Identify Disorders in Health Records using Conditional Random Fields and Metamap

AEHRC at ShARe/CLEF 2013 eHealth Evaluation Lab Task 1

G. Zuccon[1], A. Holloway[1,2], B. Koopman[1,2], A. Nguyen[1]

[1] The Australian e-Health Research Centre (CSIRO), Brisbane, Australia
[2] Queensland University of Technology, Brisbane, Australia
{guido.zuccon, alex.holloway, bevan.koopman, anthony.nguyen}@csiro.au

Abstract. The Australian e-Health Research Centre (AEHRC) recently participated in the ShARe/CLEF eHealth Evaluation Lab Task 1. The goal of this task is to individuate mentions of disorders in free-text electronic health records and map disorders to SNOMED CT concepts in the UMLS metathesaurus. This paper details our participation to this ShARe/CLEF task. Our approaches are based on using the clinical natural language processing tool Metamap and Conditional Random Fields (CRF) to individuate mentions of disorders and then to map those to SNOMED CT concepts.
Empirical results obtained on the 2013 ShARe/CLEF task highlight that our instance of Metamap (after filtering irrelevant semantic types), although achieving a high level of precision, is only able to identify a small amount of disorders (about 21% to 28%) from free-text health records. On the other hand, the addition of the CRF models allows for a much higher recall (57% to 79%) of disorders from free-text, without sensible detriment in precision. When evaluating the accuracy of the mapping of disorders to SNOMED CT concepts in the UMLS, we observe that the mapping obtained by our filtered instance of Metamap delivers state-of-the-art effectiveness if only spans individuated by our system are considered ('relaxed' accuracy).

1 Introduction

The automatic identification of clinical conditions, such as disorders, abnormalities, body sites, medications, procedures, devices, and their normalisation to a standard terminology of reference, are tasks of key importance for the analysis of free text electronic health records (e.g. discharge summaries). Solutions that tackle these tasks are fundamental to unlock clinical information trapped in the natural language of clinical narratives, which can be used to improve access, reporting, reasoning and discovery. These capabilities have been, for example, at the basis of previous research we conducted on cancer reporting [1–3], radiology reconciliation [4], and medical information retrieval [5, 6]. Other application areas include disease monitoring and pharmacological surveillance [7].

Task 1 of the ShARe/CLEF eHealth Evaluation Lab aims to provide researchers with a standard benchmark for evaluating clinical information extraction and normalisation systems [8]. The task comprises of two objectives (i.e., subtasks):

1. Identify the boundaries of mentions of disorders in discharge summaries;
2. Map each mention of disorder to a UMLS CUI (restricted to CUIs referring to SNOMED CT concepts).

Details of this task can be found in the Lab overview paper [8].

To discover mentions of disorders in the free-text of discharge summaries we implemented two solutions (runs TeamAEHRC.1 and TeamAEHRC.2) based on Metamap[9] and Conditional Random Fields [10]. Our first approach relies on Metamap (as integrated in AEHRC's Medtex medical text analysis platform [11]) to recognise mentions of disorders. The output of Metamap is filtered according to the UMLS semantic types associated with disorders as identified in the ground truth labels of the training data. Our second approach complements the output of Metamap by using Conditional Random Fields models built on training annotations, as provided by the task organisers [12]. The CRF models are built from lexical features (e.g. tokens, word shapes, etc), as well as from CUIs and semantic types as recognised by Metamap. Metamap was used to produce mappings of disorders to concepts in the UMLS metathesaurus. Disorders identified by the CRF models but not by Metamap were mapped to CUI-less concepts. Details of our approaches are given in section 2.

Empirical results obtained in the ShARe/CLEF 2013 challenge suggest that the approach implementing CRF and Metamap is more effective than using Metamap alone for identifying disorders from free-text discharge summaries. Our approach based on the combination of Metamap and CRF achieved considerably higher recall within the 'relaxed' evaluation context than when considering the 'strict' settings. This is because of a bug in the code we used for submitting our run. The bug produced a miss-alignment between the correct output of the CRF models and the text of the discharge summaries. Thus, the spans present in the submission representing this approach contain token misalignments, resulting in better effectiveness when the 'relaxed' settings are considered. Further considerations on the results of our approaches in this task are presented in section 3.2.

2 Methods

As part of our participation in this challenge, we first investigated the effectiveness of a system based on Metamap, and then the contribution that a supervised named entity recognition approach based on Condition Random Fields would have when added to the approach that relies solely on Metamap.

Metamap is a well-known software tool that uses natural language processing and knowledge-intensive approaches to identify biomedical keywords and map them to UMLS Metathesaurus concepts. We used Metamap as integrated within

AEHRC's Medtex medical text analysis platform [11]. Our instance of Medtex used the server version of Metamap 2011v2, with CUI mappings restricted to concepts belonging to the SNOMED CT terminology. Metamap was used to identify spans of text in the discharge summaries that referred to biomedical keywords. We first considered the concepts that Metamap identified for (fully or partially overlapping) mentions of disorders as identified by human expert assessors in the training discharge summaries of the ShARe/CLEF 2013 challenge. We grouped reference mentions of disorders in the training data by their semantic types as identified by Metamap. Table 1 summarises the semantic types disorders belong to; semantic types were ranked by number of occurrences.

Table 1. Statistics collected for mentions of disorders as identified in the reference standard of the ShARe/CLEF 2013 Task 1 training data. No-type refers to concepts that have been identified as disorders by the expert assessors, but that are not mapped to any concept by Metamap (and thus do not belong to any semantic type). Note that more than one semantic type can be associated with a disorder in the reference standard.

Semantic Type	Number of unique CUIs	Occurrences
Disease or Syndrome	418	1895
No-type	–	1675
Sign or Symptom	166	903
Pathologic Function	137	589
Injury or Poisoning	105	249
Congenital Abnormality	22	183
Neoplastic Process	54	107
Mental or Behavioural Dysfunction	36	106
Anatomical Abnormality	34	105
Acquired Abnormality	8	29
Finding	13	28
Mental Process	1	3
Body Substance	1	1
Cell or Molecular Dysfunction	1	1

Table 1 identifies which semantic types (as identified by Metamap) are most commonly associated with disorders (as identified by the expert assessors). To produce our first submission to this year's ShARe/CLEF 2013, we use Metamap to identify spans of biomedical keywords and their associated UMLS concepts in the discharge summaries of the test set. To retain only spans that may refer to disorders, we filter out concepts that do not belong to the top 10 semantics types identified from the training dataset (i.e., we consider only the semantic types listed in Table 1 except Mental Process, Body Substance and Cell or Molecular Dysfunction). Normalisation is achieved by considering the CUIs of the resulting concepts as provided by Metamap. The submission to the ShARe/CLEF challenge that implements this method is identified as TeamAEHRC.1.

According to Table 1, a large number of disorders identified by expert assessors in the training discharge summaries are not identified by Metamap: in fact, 1,675 disorders have no semantic type. If the training dataset is representative of the testing dataset, a similar trend is likely to be observed when testing the previous approach on unseen data. This would then result in poor recall, as Metamap would miss a large number of mentions of disorders. With the objective of improving recall of our Metamap-based approach, we complete that solution with the use of a supervised machine learning model for name entity recognition. Specifically, we chose to implement Conditional Random Fields (CRF) models to automatically identify spans of text that refer to disorders. We have used CRF models in previous work on de-identification of electronic health records, and we found that, provided enough training data is made available, CRF models are able to effectively identify targeted named entities [13].

A Conditional Random Fields classifier is a discriminative undirected probabilistic graphical model that, given a observed sequence, defines a log-linear distribution over labelled sequences. To build the CRF models we used the following lexical and semantic features:

- the word tokens;
- word shapes features (e.g. the presence of capitalised characters at the beginning of the word token or across the whole token);
- letter n-grams (n = 6);
- disjunctive features, which capture disjunctions of words and word shapes within windows of tokens;
- position, which captures the position of a word in the sentence;
- the UMLS CUIs as provided by Metamap;
- a disorder flag as provided by our first approach (i.e. TeamAEHRC.1).

Features were extracted from discharge summaries in the training and testing datasets. The CRF model was trained using discharge summaries from the ShARe/CLEF Task 1 training set only. The submission to the ShARe/CLEF challenge that implements this method is identified as TeamAEHRC.2.

3 Evaluation on the ShARe/CLEF Challenge

3.1 Evaluation Measures

To evaluate the effectiveness of approaches to identify mentions of disorders in discharge summaries, the organisers of the ShARe/CLEF Task 1 considered:

- Precision (P): TP / (TP + FP);
- Recall (R): TP / (TP + FN);
- F-measure (F): (2 * Recall * Precision) / (Recall + Precision);

where true positive (TP) indicates that a system identified a disorder in the same span as that identified by the expert assessors, false positive (FP) refers to the identification of an incorrect span, and false negative (FN) indicates that

a system failed to identify a disorder-span that was instead identified by the expert assessors. The 'strict' and 'relaxed' evaluation settings refer to the case where the automatically identified span is identical to the reference span, and that identified span overlaps with the reference span, respectively. We refer the reader to the task overview paper for more details.

Accuracy was chosen to evaluate the effectiveness of approaches to map mentions of disorders in discharge summaries to SNOMED CT concepts in the UMLS. Accuracy is defined as the ratio of correctly mapped concepts to the total number of mentions of disorders. In the 'strict' evaluation settings, the total number of mentions of disorders is computed over the reference standard identified by the expert assessors. In the 'relaxed' settings, the total number of mentions of disorders is computed over the mentions identified by the system that strictly overlap with the reference standard.

3.2 Results

Results obtained by the two submitted runs are reported in Tables 2 and 3 for the identification of disorders and their mapping to UMLS concepts, respectively.

Table 2. Precision (P), Recall (R) and F-measure (F) in the strict and relaxed evaluation settings of TeamAEHRC submissions to ShARe/CLEF Task 1a. The effectiveness of the best system at ShARe/CLEF Task 1a is reported in the bottom row.

	Strict			Relaxed		
Run	P	R	F	P	R	F
TeamAEHRC.1	0.699	0.212	0.325	0.903	0.275	0.422
TeamAEHRC.2	0.613	0.566	0.589	0.886	0.785	0.833
Best ShARe/CLEF System	0.800	0.706	0.750	0.925	0.827	0.873

Table 3. Precision (P), Recall (R) and F-measure (F) in the strict and relaxed evaluation settings of TeamAEHRC submissions to ShARe/CLEF Task 1b. The effectiveness of the best system at ShARe/CLEF Task 1b is reported in the bottom row. Note that TeamAEHRC.1 is the submission that achieved highest accuracy on Task 1b when the relaxed evaluation setting is considered.

	Strict	Relaxed
Run	Accuracy	Accuracy
TeamAEHRC.1	0.199	0.939
TeamAEHRC.2	0.313	0.552
Best ShARe/CLEF System	0.589	–

Results of TeamAEHRC.1 for task 1a (Table 2) suggest that our Metamap instance only identifies a limited number of disorders, with recall between 21.2% and 27.5%. While the relaxed evaluation setting does not sensibly affect the recall effectiveness of TeamAEHRC.1, it is observed that precision increases over the run evaluated within the strict setting. This suggests that if span overlaps between system annotation and reference standard are considered, the disorders identified by the system are highly likely to have been identified also by expert assessors. The high 'relaxed' accuracy achieved by TeamAEHRC.1 on task1b (Table 2; indeed, the highest accuracy across the challenge systems) highlights that when spans of disorders are correctly identified, Metamap is highly effective in providing a mapping consistent with those of the expert assessors.

Results of TeamAEHRC.2 submission are affected by a bug in our code that prevents the correct alignment between positions of disorder annotations as identified by our system and the token positions in the challenge's reference standard. This explains why this system achieves significantly better effectiveness across the 'relaxed' evaluation settings than the 'strict' one, both for task 1a and task 1b. The 'relaxed' settings for task 1a account for partial overlaps between system annotations and reference standards; this partially corrects the bug in our system. The analysis of Table 1 suggested that the system based on Metamap may not identify a large number of disorders: this has proven to be the case (the recall of TeamAEHRC.1 is low). The intuition for complementing the Metamap-based approach with CRF models was that these may identify patterns in the text of discharge summaries that would allow to identify mentions of disorders Metamap would not recognise. The (relaxed) recall achieved by TeamAEHRC.2 suggests that this has been indeed the case: CRF models enable to identify about 3 times more mentions of disorders than the system based solely on Metamap (TeamAEHRC.1). The highest recall provided by CRF is traded off for a loss in precision (Precision of TeamAEHRC.1: 0.903; Precision of TeamAEHRC.2: 0.886): some of the spans identified by the system are actually not mentions of disorders according to expert assessors. However, this loss is minimal (1.88%) and indeed TeamAEHRC.2 obtains almost double the F-measure than that of TeamAEHRC.1 (TeamAEHRC.1: 0.422; TeamAEHRC.2: 0.833).

4 Conclusions

In this paper we have presented the methods used in our submissions to the ShARe/CLEF 2013 eHealth Evaluation Lab Task 1. Our methods are based on an instance of Metamap and on Conditional Random Fields. Empirical results suggest that if a disorders mention is correctly identified by Metamap, then its mapping to a CUI provided by this system is highy likely to be correct. However, Metamap does only recognise a handful of mentions of disorders: many of the disorders identified by the ShARe/CLEF expert assessors are not recognised by our instance of Metamap. To increase recall, we have complemented our Metamap solution with Conditional Random Fields models. Our implementation was affected by a software bug, which prevented correct alignment of identifing

spans and tokens. The 'relaxed' evaluation settings partially addresses our span alignment issue. When 'relaxed' effectiveness is considered, the solution that mixes Metamap and Conditional Random Fields (TeamAEHRC.2) is able to identify a large number of disorders, trading off only a small amount of the precision provided by Metamap. When considering mapping of free-text to a standard reference terminology (task 1b), the mapping provided by Metamap is found to be highly accurate for the mentions that have been correctly identified by the system.

References

1. Nguyen, A., Moore, J., Zuccon, G., Lawley, M., Colquist, S.: Classification of pathology reports for cancer registry notifications. In: Health Informatics: Building a Healthcare Future Through Trusted Information-Selected Papers from the 20th Australian National Health Informatics Conference (Hic 2012). Volume 178., Ios PressInc (2012) 150
2. Zuccon, G., Nguyen, A., Bergheim, A., Wickman, S., Grayson, N.: The impact of ocr accuracy on automated cancer classification of pathology reports. In: Health Informatics: Building a Healthcare Future Through Trusted Information-Selected Papers from the 20th Australian National Health Informatics Conference (Hic 2012). Volume 178., Ios PressInc (2012) 250
3. Butt, L., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Classification of cancer-related death certificates using machine learning. Australasian Medical Journal **6**(5) (2013) 292–299
4. Zuccon, G., Wagholikar, A., Nguyen, A., Chu, K., Martin, S., Lai, K., Greenslade, J.: Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. In: AMIA Clinical Research Informatics. (2013)
5. Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., Butt, L.: Exploiting medical hierarchies for concept-based information retrieval. In: Proceedings of the Seventeenth Australasian Document Computing Symposium, ACM (2012) 111–114
6. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: Graph-based concept weighting for medical information retrieval. In: Proceedings of the Seventeenth Australasian Document Computing Symposium, ACM (2012) 80–87
7. LePendu, P., Iyer, S.V., Bauer-Mehren, A., Harpaz, R., Mortensen, J.M., Podchiyska, T., Ferris, T.A., Shah, N.H.: Pharmacovigilance using clinical notes. Clinical Pharmacology & Therapeutics (2013)
8. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Mowery, D., Leveling, J., Goeuriot, L., Kelly, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: CLEF 2013. Lecture Notes in Computer Science (LNCS), Springer (2013)
9. Aronson, A.R., Lang, F.M.: An overview of MetaMap: historical perspective and recent advances. JAMIA **17**(3) (2010) 229–236
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of ICML'01. (2001) 282–289
11. Nguyen, A., Moore, J., Lawley, M., Hansen, D., Colquist, S.: Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In: Health Informatics Conference. (2011) 117–124

12. Noémie Elhadad, W.C., O'Gorman, T., Palmer, M., Savova, G.: The share schema for the syntactic and semantic annotation of clinical texts (2013 (under review))
13. Zuccon, G., Strachan, M., Nguyen, A., Bergheim, A., Grayson, N.: Automatic de-identification of electronic health records: An australian perspective. In: 4th International Workshop on Health Document Text Mining and Information Analysis (LOUHI'13). (2013) 1–5