Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature

Noha Alnazzawi, Paul Thompson and Sophia Ananiadou
School of Computer Science, University of Manchester, UK
alnazzan@cs.man.ac.uk, {paul.thompson,
sophia.ananiadou@manchester.ac.uk}

Abstract

Narrative information in Electronic Health Records (EHRs) and literature articles contains a wealth of clinical information about treatment, diagnosis, medication and family history. This often includes detailed phenotype information for specific diseases, which in turn can help to identify risk factors and thus determine the susceptibility of different patients. Such information can help to improve healthcare applications, including Clinical Decision Support Systems (CDS). Clinical text mining (TM) tools can provide efficient automated means to extract and integrate vital information hidden within the vast volumes of available text. Development or adaptation of TM tools is reliant on the availability of annotated training corpora, although few such corpora exist for the clinical domain. In response, we have created a new annotated corpus (PhenoCHF), focussing on the identification of phenotype information for a specific clinical sub-domain, i.e., congestive heart failure (CHF). The corpus is unique in this domain, in its integration of information from both EHRs (300 discharge summaries) and literature articles (5 full-text papers). The annotation scheme, whose design was guided by a domain expert, includes both entities and relations pertinent to CHF. Two further domain experts performed the annotation, resulting in high quality annotation, with agreement rates up to 0.92 F-Score.

1 Introduction

An ever-increasing number of scientific articles is published every year. For example, in 2012, more than 500,000 articles were published in MEDLINE (U.S. National Library of Medicine , 2013). A researcher would thus need to review at least 20 articles per day in order to keep up to date with latest knowledge and evidence in the literature (Perez-Rey et al., 2012).

EHRs constitute a further rich source of information about patients' health, representing different aspects of care (Jensen et al., 2012). However, clinicians at the point of care have very limited time to review the potentially large amount of data contained within EHRs. This presents significant barriers to clinical practitioners and computational applications (Patrick et al., 2006).

TM tools can be used to extract phenotype information from EHRs and the literature and help researchers to identify the characteristics of CHF and to better understand the role of the deterioration in kidney function in the cycle of progression of CHF.

2 Related work

There are many well-known publicly available corpora of scientific biomedical literature, which are annotated for biological entities and/or their interactions (often referred to as *events*) (Roberts et al., 2009; Xia & Yetisgen-Yildiz, 2012). Examples include GENIA (Kim et al., 2008), BioInfer (Pyysalo et al., 2007) GREC (Thompson et al., 2009), PennBioIE (Kulick et al., 2004), GENETAG (Tanabe et al., 2005) and LLL'05 (Hakenberg et al., 2005). However, none of these corpora is annotated with the types of entities and relationships that are relevant to the study of phenotype information.

On the other hand, corpora of clinical text drawn from EHRs are rare, due to privacy and confidentiality concerns, but also because of the time-consuming, expensive and tedious nature of producing high quality annotations, which are reliant on the expertise of domain experts (Uzuner et al., 2011). A small number of corpora, however, have been made available, mainly in the context of shared task challenges, which aim to encourage the development of information extraction (IE) systems. These corpora vary in terms of the text type and annotation granularity. For example, the corpus presented in (Pestian et al., 2007) concerns only structured data from radiology reports, while the corpus presented in (Meystre & Haug, 2006) contains unstructured parts of EHRs, but annotated with medical problem only at the document level.

Other corpora are more similar to ours, in that that they include text-bound annotations

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 69–74,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

corresponding to entities or relations. CLEF (Clinical E-Science Framework) (Roberts et al., 2008) was one of the first such corpora to include detailed semantic annotation. It consists of a number of different types of clinical records, including clinic letters, radiology and histopathology reports, which are annotated with a variety of clinical entities, relations between them and co-reference. However, the corpus has not been made publicly available. The more recent 2013 CLEF-eHEALTH challenge (Suominen et al., 2013) corpus consists of EHRs annotated with named entities referring to disorders and acronyms/abbreviations, mapped to UMLS concept identifiers.

The Informatics for Integrating Biology at the Bedside (i2b2) NLP series of challenges have released a corpus of de-identified clinical records annotated to support a number of IE challenges with multiple levels of annotation, i.e., entities and relations (Uzuner et al., 2008; Uzuner, 2009). The 2010 challenge included the release of a corpus of discharge summaries and patient reports in which named entities and relations concerning medical problems, tests and treatments were annotated (Uzuner et al., 2011). A corpus of EHRs from Mayo Clinic has been annotated with both linguistic information (part-of–speech tags and shallow parsing results) and named entities corresponding to disorders (Ogren et al., 2008; Savova et al., 2010).

3 Description of the corpus

The discharge summaries in our PhenoCHF corpus constitute a subset of the data released for the second i2b2 shared task, known as "recognising obesity" (Uzuner, 2009). PhenoCHF corpus was created by filtering the original i2b2 corpus, such that only those summaries (a total of 300) for patients with CHF and kidney failure were retained.

The second part of PhenoCHF consists of the 5 most recent full text articles (at the time of query submission) concerning the characteristics of CHF and renal failure, retrieved from the PubMed Central Open Access database.

4 Methods and results

The design of the annotation schema was guided by an analysis of the relevant discharge summaries, in conjunction with a review of comparable domain specific schemata and guidelines, i.e., those from the CLEF and i2b2 shared tasks. The schema is based on a set of requirements developed by a cardiologist. Taking into account our chosen focus of annotating phenotype information relating to the CHF disease, the cardiologist was asked firstly to determine a set of relevant entity types that relate to CHF phenotype information and the role of the decline in kidney function in the cycle of CHF (exemplified in Table 1), secondly to locate words that modify the entity (such as polarity clues) and thirdly to identify the types of relationships that exist between these entity types in the description of phenotype information (Table 2) .

Secondly, medical terms in the records are mapped semi-automatically onto clinical concepts in UMLS, with the aid of MetaMap (Aronson, 2001).

The same annotation schema and guidelines were used for both the discharge summaries and the scientific full articles. In the latter, certain annotations were omitted, i.e., organ entities, polarity clues and relations. This decision was taken due to the differing ways in which phenotype information is expressed in discharge summaries and scientific articles. In discharge summaries, phenotype information is explicitly described in the patient's medical history, diagnoses and test results. On the other hand, scientific articles summarise results and research findings. This means that certain types of information that occur frequently in discharge summaries are extremely rare in scientific articles, such that their occurrences are too sparse to be useful in training TM systems, and hence they were not annotated.

The annotation was carried out by two medical doctors, using the Brat Rapid Annotation Tool (brat) (Stenetorp et al., 2012), a highly-configurable and flexible web-based tool for textual annotation.

Annotations in the corpus should reflect the instructions provided in the guidelines as closely as possible, in order to ensure that the annotations are of ahigh quality. A standard means of providing evidence regarding the reliability of annotations in a corpus is to calculate a statistic known as the inter-annotator agreement (IAA). IAA provides assurance that different annotators can produce the same annotations when working independently and separately. There are several different methods of calculating IAA, which can be influenced by the exact nature of the annotation task. We use the measures of precision, recall and F-measure to indicate the level of inter-annotator reliability (Hripcsak & Rothschild, 2005). In order to carry out such calculations, one set of annotations is considered as a gold standard and the total number of correct entities is the total number of entities annotated by this annotator.

Precision is the percentage of correct positive predictions annotated by the second annotator, compared to the first annotator's assumed gold standard. It is calculated as follows:

$$P = TP / TP + FP$$

Recall is the percentage of positive cases recognised by the second annotator. It is calculated as follows:

$$R = TP / TP + FN$$

F-score is the harmonic mean between precision and recall.

$$F\text{-}score =$$

$$2 * (Precision * Recall) / Precision + Recall$$

We have calculated separate IAA scores for the discharge summaries and the scientific articles. Table 3 summarises agreement rates for term annotation in the discharge summaries, showing results for both individual entity types and macro-averaged scores over all entity types. Relaxed matching criteria were employed, such that annotations added by the two annotators were considered as a match if their spans overlapped. In comparison to related efforts, the IAA rates shown in Table 3 are high. However, it should be noted that the number of targeted classes and relations in our corpus is small and focused, compared to other related corpora.

Agreement statistics for scientific articles are shown in Table 4. Agreement is somewhat lower than for discharge summaries, which this could be due to the fact that the annotators (doctors) are more used to dealing with discharge summaries in their day-to-day work, and so are more accustomed to locating information in this type of text. Scientific articles are much longer and generally include more complex language, ideas and analyses, which may require more than one reading to fully comprehend the information within them. Table 5 shows the agreement rates for relation annotation in the discharge summaries. The agreement rates for relationships are relatively high. This can partly be explained by the deep domain knowledge possessed by the annotators and partly by the fact that the relationships to be identified were relatively simple, linking only two pre-annotated entities.

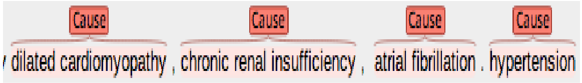
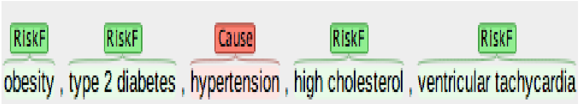
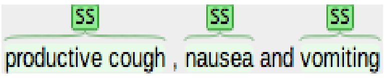
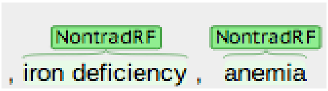
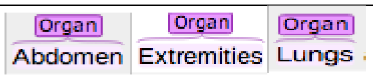
Entity Type	Description	Example
Cause	any medical problem that contributes to the occurrence of CHF	 Cause Cause Cause Cause / dilated cardiomyopathy , chronic renal insufficiency , atrial fibrillation . hypertension
Risk factors	A condition that increases the chance of a patient having the CHF disease	 RiskF RiskF Cause RiskF RiskF obesity , type 2 diabetes , hypertension , high cholesterol , ventricular tachycardia
Sign & symptom	any observable manifestation of a disease which is experienced by a patient and reported to the physician	 SS SS SS productive cough , nausea and vomiting
Non-traditional risk factor	Conditions associated with abnormalities in kidney functions that put the patient at higher risk of developing "signs & symptoms" and causes of CHF	 NontradRF NontradRF , iron deficiency , anemia
Organ	Any body part	 Organ Organ Organ Abdomen Extremities Lungs .

Table 1. Annotated phenotype entity classes

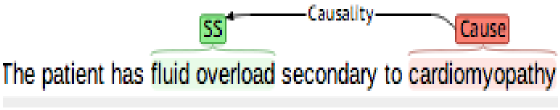
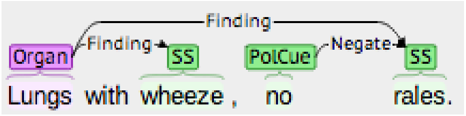
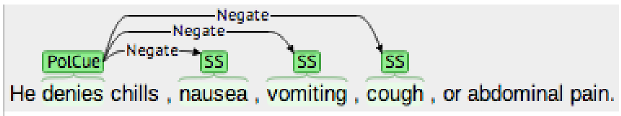
Relation Type	Description	Example
Causality	This relationship links two concepts in cases in which one concept causes the other to occur.	 The patient has fluid overload secondary to cardiomyopathy
Finding	This relationship links the organ to the manifestation or abnormal variation that is observed during the diagnosis process.	 Lungs with wheeze , no rales.
Negate	This is one-way relation to relate a negation attribute (polarity clue) to the condition it negates.	 He denies chills , nausea , vomiting , cough , or abdominal pain.

Table 2. Description of Annotated Relations

	Causality	Risk factor	Sign & Symptom	Non-traditional risk factor	Polarity clue	Organ	Macro-average
F-score	0.95	0.94	0.97	0.83	0.94	0.92	0.92

Table 3. Term annotation agreement statistics for discharge summaries

	Cause	Risk factor	Sign & Symptoms	Non-traditional risk factor	Macro-average
F-score	0.82	0.84	0.82	.77	0.81

Table 4. Overall agreement statistics for terms annotation in scientific articles

	Causality	Finding	Negate	Macro-average
F-score	0.86	0.94	0.95	0.91

Table 5. Relation annotation and agreement statistics for discharge summaries

5 Conclusion

This paper has described the creation of a new annotated corpus to facilitate the customisation of TM tools for the clinical domain. The corpus[1] consists of 300 discharge summaries and 5 full-text articles from the literature, annotated for CHF phenotype information, including causes, risk factors, sign & symptoms and non-traditional risk factors. Discharge summaries have also been annotated with relationships holding between pairs of annotated entities. A total 7236 of entities and 1181 relationships have been annotated. Extracting phenotype information can have a major impact on our deeper understanding of disease ethology, treatment and prevention (Xu et al., 2013). Currently we are working on confirming the utility of the annotated corpus in training and customising TM tools, i.e., adapting different sequence tagging algorithms (such as Conditional Random Fields (CRF) and Hidden Markov Model (HMM)) to extract comprehensive clinical information from both discharge summaries and scientific articles.

[1] Guidelines and stand-off annotation are publicly available at https://code.google.com/p/phenochf-corpus/source/browse/trunk

References

MEDLINE citation counts by year of publication.

Aronson, A.R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. Proceedings of the AMIA Symposium, American Medical Informatics Association.

Hakenberg, J., Plake, C., Leser, U., Kirsch, H. and Rebholz-Schuhmann, D. (2005). *LLL'05 challenge: Genic interaction extraction-identification of language patterns based on alignment and finite state automata*. Proceedings of the 4th Learning Language in Logic workshop (LLL05).

Hripcsak, G. and Rothschild, A.S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3): 296-298.

Jensen, P.B., Jensen, L.J. and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6): 395-405.

Kim, J.-D., Ohta, T. and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Ungar, L., Winters, S. and White, P. (2004). *Integrated annotation for biomedical information extraction*. Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).

Meystre, S. and Haug, P.J. (2006). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics*, 39(6): 589-599.

Ogren, P.V., Savova, G.K. and Chute, C.G. (2008). *Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition*. LREC.

Patrick, J., Wang, Y. and Budd, P. (2006). *Automatic Mapping Clinical Notes to Medical Terminologies*. Australasian Language Technology Workshop.

Perez-Rey, D., Jimenez-Castellanos, A., Garcia-Remesal, M., Crespo, J. and Maojo, V. (2012). CDAPubMed: a browser extension to retrieve EHR-based biomedical literature. *BMC Medical Informatics and Decision Making*, 12(1): 29.

Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K.B. and Duch, W. (2007). *A shared task involving multi-label classification of clinical free text*. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J. and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1): 50.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Setzer, A. and Roberts, I. (2008). *Semantic annotation of clinical text: The CLEF corpus*. Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I. and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5): 950-966.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C. and Chute, C.G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5): 507-513.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J.i. (2012). *BRAT: a web-based tool for NLP-assisted text annotation*. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L. and Jones, G.J. (2013). Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Springer**:** 212-231.

Tanabe, L., Xie, N., Thom, L.H., Matten, W. and Wilbur, W.J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1): S3.

Thompson, P., Iqbal, S., McNaught, J. and Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1): 349.

Uzuner, Ö., Goldstein, I., Luo, Y. and Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1): 14-24.

Uzuner, Ö. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4): 561-570.

Uzuner, Ö., South, B.R., Shen, S. and DuVall, S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5): 552-556.

Xia, F. and Yetisgen-Yildiz, M. (2012). *Clinical corpus annotation: challenges and strategies*. Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Xu, R., Li, L. and Wang, Q. (2013). Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics*, 29(17): 2186-2194.