Evaluation of Medical Concept Annotation Systems on Clinical Records

Hamed Hassanzadeh Anthony Nguyen Bevan Koopman

The Australian e-Health Research Centre, CSIRO, Brisbane, QLD, Australia

{hamed.hassanzadeh, anthony.nguyen, bevan.koopman}@csiro.au

Abstract

Large volumes of electronic health records, including free-text documents, are extensively generated within various sectors of healthcare. Medical concept annotation systems are designed to enrich these documents with key concepts in the domain using reference terminologies. Although there is a wide range of annotation systems, there is a lack of comparative analysis that enables thorough understanding of the effectiveness of both the concept extraction and concept recognition components of these systems, especially within the clinical domain. This paper analyses and evaluates four annotation systems (i.e., MetaMap, NCBO annotator, Ontoserver, and QuickUMLS) for the task of extracting medical concepts from clinical free-text documents. Empirical findings have shown that each annotator exhibits various levels of strengths in terms of overall precision or recall. The concept recognition component of each system, however, was found to be highly sensitive to the quality of the text spans output by the concept extraction component of the annotation system. The effects of these components on each other are quantified in such way as to provide evidence for an informed choice of an annotation system as well as avenues for future research.

1 Introduction

With the advent of electronic health records, large volumes of mostly free-text clinical documents — discharge summaries, radiology reports, pathology reports, and patients progress notes — are now present in the health ecosystem. While these documents contain much valuable information, it can only be exploited if effective computational methods of dealing with clinical free-text are devised. The goal here is to automatically extract clinical concepts from unstructured clinical documents, thus providing a structured representation that enables fast and effective access and analysis.

To facilitate the extraction of clinical concepts from free-text, many automatic systems (known as medical concept annotators) have been developed. These systems analyse natural language and annotate specific spans of text to concepts defined in some external medical terminology/thesaurus. This workflow can be considered as a two-step process of extracting candidate spans of concepts within a given document (known as "concept extraction") and then assigning appropriate concept identifiers to each candidate span based on the defined concepts in the domain ontologies (known as "concept recognition"). Such systems are widely used in a variety of e-health settings and are critical for activities such as clinical information analysis and reporting (Zuccon et al., 2013), derivation of phenotypic descriptions (Groza et al., 2013b; Collier et al., 2014) and medical information retrieval (Zuccon et al., 2012; Koopman, 2014).

Although there are a wide range of available annotation systems, there is a lack of comparative analysis that provides enough evidence for an informed decision in choosing the most suitable system. Many of these system are developed for a specific domain (e.g., medical journal article abstracts) and may not be suited to dealing with clinical text. Deployment of these systems can often only be done in a black-box fashion: without an underlying understanding of the individual components of a system and its effectiveness.

This paper aims to analyse and evaluate four annotation systems on the task of extracting medical concepts from clinical free-text documents. Specifically, we investigate the following research

Hamed Hassanzadeh, Anthony Nguyen and Bevan Koopman. 2016. Evaluation of Medical Concept Annotation Systems on Clinical Records. In *Proceedings of Australasian Language Technology Association Workshop*, pages 16−25.

questions:

1. How well do common medical concept annotation systems perform on clinical free-text?
2. What is the impact of the core components of an annotation system (i.e., concept extraction and concept recognition) on their overall performance?

The analysis of the performances of the annotation systems show that different components of the annotation systems exhibit different levels of strengths in terms of overall precision or recall. When evaluating the performance of the individual concept extraction and concept recognition components of the systems, it was found that the concept recognition performance was highly dependent on a high performing concept extraction component. This leads to a set of insights over annotation systems from both application and development perspectives.

2 Related Work

Due to the advances in electronic health records and the availability of large volumes of clinical text documents, significant interest has been directed towards automating their processing and analyses. Several workshops and shared tasks have been designed in recent years to attract researchers to the domain and challenge different ideas and methodologies for such tasks. The ShARe/CLEF eHealth shared task in 2013 is one of them that focuses on the application of Natural Language Processing (NLP), Machine Learning (ML), and Information Retrieval (IR) for leveraging health care data[1]. Task 1 in the CLEF ShARed Task focuses on the concept recognition problem, more specifically, on identifying disorder concepts from clinical documents. It comprises two subtasks: *(i) Task 1a* a concept extraction task that evaluates the systems according to their ability to extract correct spans of text for disorder concepts; and *(ii) Task 1b* a concept recognition task that is about assigning the correct class of concept (i.e., a Concept Unique Identifier or CUI) to each text span using the Unified Medical Language System (UMLS) terminology (Suominen et al., 2013; Keith and others, 1998). Note that, only a subset of UMLS concepts were used for this annotation task (i.e., only those UMLS concepts that were associated to particular disorder-related concepts in the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT ontology)) A data set was provided to the participants in order to develop and test their automatic systems (more information about the data set is described in Section 3.3). A similar task was offered in the following year in SemEval 2014 Task 7 (Pradhan et al., 2014), which applied the same data set as the ShARe/CLEF task as a follow-up on the concept recognition task. In this paper, we also apply the ShARe/CLEF data set as it provides manually annotated concepts that can be used to evaluate concept annotation systems. However, different to the systems that were specifically designed for the task and tailored to the data set, we investigate the performance of off-the-shelf annotation systems for annotating this data set with medical concepts.

Mirhosseini et al. (Mirhosseini et al., 2014) also applied a subset of the same data set (i.e., the train set of ShARe/CLEF data) to compare medical annotation systems (e.g. MetaMap (Aronson, 2001; Aronson and Lang, 2010), Ontoserver (McBride et al., 2012)), and a number of standard IR techniques for the concept recognition component of the shared task (i.e. Task 1b). They considered the concept recognition task as an Information Retrieval technique and used queries with the spans of text associated with the concepts in the gold standard. The responses of annotation systems were then evaluated using standard IR evaluation measures, such as Reciprocal Ranker and Success@K. They converted all the UMLS concepts IDs in the ShARe/CLEF data to one or more corresponding SNOMED CT IDs and performed the evaluation on this new version of the data. In this paper, we investigate an extended number of annotation systems and use the original dataset and evaluation metrics for evaluating the end-to-end effectiveness of the annotation systems (as opposed to only the concept recognition component of the systems).

Funk et al. (Funk et al., 2014) compared three annotation systems (i.e., MetaMap, NCBO Annotator (Jonquet et al., 2009a), and ConceptMapper (Tanenblatt et al., 2010)) by focusing on tuning their configurable parameters according to particular ontologies on full-text articles in the biomedical domain. They evaluated the systems un-

[1] `https://sites.google.com/site/shareclefehealth/home`

der different settings according to eight ontologies. They found that the systems did not achieve the best performance with their default parameters and changes in these parameters had a significant effect on effectiveness. The ConceptMapper system was found to be the best performing system across the majority of the ontologies. Different to our study, is their use of mainly genetic-related ontologies as opposed to the clinical SNOMED CT ontology, and their use of published articles (which are written in more formal language) compared to narrative clinical documents (which are often in the form of unstructured, ungrammatical, and often fragmented free-text).

Groza et al. (Groza et al., 2013c) compared four open medical concept recognition systems (i.e., cTAKES (Savova et al., 2010), NCBO Annotator (Jonquet et al., 2009a), BeCAS (Nunes et al., 2013) and MetaMap) with their default settings. These comparisons are performed over one semi-gold and one silver standard data sets comprising of clinical trials and published abstracts. Their silver and semi-gold standard corpora were (semi-)automatically generated using different combinations of the output of their studied annotation systems. Like Groza et al. (Groza et al., 2013c), we study a range of medical concept recognition systems but with a focus on clinical records and associated gold standard that has been curated by domain experts.

3 Methodology

3.1 Annotation Systems

Automatic annotation systems commonly comprise of two distinct components: *(i)* ***Concept Extraction***, and *(ii)* ***Concept Recognition***. The concept extraction component of the systems is responsible for the extraction of candidate text spans from the input document that potentially refer to medical concepts, such as, disorders as in the ShARe/CLEF data. The concept recognition component then aims to assign a domain concept (using one or more base terminologies) that is semantically related to the candidate span of text.

In this paper, we evaluated medical concept annotation systems from both the concept extraction and concept recognition perspectives. The investigated systems in this study include two of the most popular medical concept annotators (i.e., MetaMap and NCBO annotators) and two of the more recent systems (i.e., QuickUMLS and Ontoserver). Brief descriptions of the systems are provided in the following:

MetaMap is an annotation program that is developed by the National Library of Medicine (NLM) to annotate a given text with appropriate concepts (i.e., UMLS Metathesaurus). MetaMap has a range of configurable parameters and options to tune its different NLP and retrieval components and its output (Aronson, 2001; Aronson and Lang, 2010). The MetaMap service usually requires considerable time in order to process the input text and annotate concepts (Shah et al., 2009; Soldaini and Goharian, 2016).

NCBO Annotator is an annotation service that covers a wide range of ontologies (i.e., more than 500 ontologies) (Jonquet et al., 2009a). Its workflow consists of a syntactic concept extraction step that employs concept names and synonyms and a semantic expansion step that tries to enrich the extracted concepts with the semantic features from ontologies. NCBO provides a set of configurable options that can be customised according to different settings and applications (Jonquet et al., 2009b).

QuickUMLS is a concept recognition approach that employs an approximate dictionary matching technique (Soldaini and Goharian, 2016). Given a text, it tries to find highly similar concepts (using the concept's string) to the given text. Instead of calculating similarities between all the concepts in the dictionary and the given text, it applies CPMerge to reduce computation costs (Okazaki and Tsujii, 2010). CPMerge is an algorithm for approximate dictionary matching. It finds a subset of concepts that have a number of features in common with the given input.

Ontoserver is a terminology server that provides an Information Retrieval solution to medical concept annotation [2]. It employs SNOMED CT as the base terminology but also supports the Australian Medicines Terminology (AMT) and Logical Observation Identifiers Names and Codes (LOINC). It exploits a purposely-tuned retrieval function and linguistic capabilities such as spell checking, restrictions and inferences on the source ontology (McBride et al., 2012). Unlike the above systems, Ontoserver currently only supports the concept recognition phase of an annotation system. As a result, Ontoserver is currently unable to use as input the whole document and perform the

[2] http://ontoserver.csiro.au:8080/

concept extraction to generate suitable text spans for concept recognition.

Table 1 shows an overview of main components of the above-mentioned annotation systems. It can be observed that the annotation systems support the UMLS terminology to annotate input documents, with the exception that Ontoserver is based on the SNOMED CT ontology. MetaMap and NCBO annotators are mainly designed to annotate biomedical literature while Ontoserver is targeted towards searching for specific clinical terminology and QuickUMLS is a generic annotator. All of the annotation systems provide APIs to access and deploy their respective medical concept annotation systems.

3.2 Concept Extraction

Concept extraction refers to the identification of appropriate spans of text that can represent a domain concept. Most annotation systems have built-in concept extraction modules. However, to control for the concept extraction component of these systems, three different concept extraction approaches, one manual and two computational approaches, were investigated to generate candidate text spans to evaluate the concept recognition component of the annotation systems.

3.2.1 Gold Standard

In order to assess the systems concept recognition performance, the exact gold standard spans of text were submitted to the systems. The gold standard text spans were generated by human experts, and hence, they can be used as a benchmark to assess the effectiveness of automatic concept extraction approaches.

3.2.2 Noun Phrase Parser

From a lexical perspective, the disorder-related terminologies are mainly in the form of subjects or objects of sentences rather than predicates or actions (e.g., the post-verb component in the following sentence: “The patient was admitted with headache and dysarthria.”). It is considered that the noun phrases of sentences in clinical documents are the dominant sources of medical concepts, especially for disorder concepts. Hence, a parser is employed to extract noun phrases from documents and form the input for concept annotation systems. One issue associated with this approach is that the clinical documents are commonly ungrammatical. As a result, an English noun phrase parser algorithm used as a black-box will face issues around the parsing of improper sentences, and hence, likely to produce noisy noun phrase text spans.

3.2.3 CRF Concept Extractor

A Conditional Random Field classifier (CRF) (Lafferty et al., 2001) can be used to automatically extract the boundary of candidate text chunks. CRF is a probabilistic undirected graphical model that has shown promising results in sequence labelling and text classification problems, especially in medical domain (Hassanzadeh et al., 2014; Kholghi et al., 2016; Groza et al., 2013a; de Bruijn et al., 2011; Hassanzadeh and Keyvanpour, 2013). The CRF model was trained over the training set of the ShARe/CLEF task corpus using the following features: words and their lemmas, Part of Speech (POS) tags, orthographic information (e.g., flagging if words contain initial capital letter, numerics, punctuations, etc.), character n-grams (i.e, 2 to 4-grams), and sequential features by including previous and next words (and their POS tags) in the feature vector of a given word and flagging if the word is the first/last word of a sentence. All-punctuation tokens (such as “||||” used as a separator) and determiner tokens (including numerical values) are removed in a preprocessing step. Although punctuations and determiners are not considered as independent tokens, they still participate in the feature vector of their adjacent words (i.e., a word that has such tokens in its preceding or following keeps this information in its feature vector).

3.3 Data

The ShARe/CLEF corpus was employed to evaluate the performance of the annotation systems (Suominen et al., 2013). This corpus contains de-identified clinical reports of diverse types, such as discharge summaries, electrocardiogram reports, and echocardiogram and radiology reports. In each document, those spans of text that correspond to disorder concepts were manually annotated by experts. These annotations were based on the UMLS Concept Unique Identifiers (CUIs) (Keith and others, 1998). Disorder concepts were considered to be concepts that are subcategories of the Disorder semantic group in the SNOMED CT ontology. Each span of text, which can refer to non-adjacent tokens in the documents, is annotated with a single CUI. Spans of text in the

Table 1: Annotation Services Specifications.

	Supported Terminology	Domain	Software Infrastructure
Metamap	UMLS	Biomedical literature	Prolog
NCBO	UMLS/NCBO	Biomedical literature	Java
Ontoserver	SNOMED CT/AMT/LOINC*	Clinical terminology use within health sector	Java
QuickUMLS	UMLS*	Generic	Python & C++

* Can be extended to employ other terminologies.

Table 2: The ShARe CLEF disorder concept recognition corpus statistics.

	Train Set	Test Set
No. Documents	199	99
All disorder	5,874	5,351
CUI-less disorder	1,661(28%)	1,750 (33%)
Non-CUI-less disorder	4,213 (72%)	3,601 (67%)
Disjoint disorder	660 (11%)	439 (8%)
Non-disorder tokens	59,835	56,610

corpus where annotators annotated them as disorders but no UMLS concept have been found for them were annotated with a "CUI-less" label. The ShARe CELF corpus comprises separate train and test sets that consist of 199 and 99 clinical documents, respectively. Detailed statistics of this corpus are shown in Table 2. *Disjoint* concepts refer to concepts where their spans cover discontinuous tokens. Recognising such concepts is more challenging than the regular concepts as the recogniser should be able to foresee possible tokens that can be assigned to a concept as a whole.

3.4 Evaluation Measures

The annotation systems were evaluated based on standard Information Extraction measures, namely, Precision, Recall, and F1-Score:

Precision (P): TP / (TP + FP);

Recall (R): TP / (TP + FN);

F1-Score (F1): (2 * Recall * Precision) / (Recall + Precision); i.e, Harmonic mean of Precision and Recall.

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 identified by the expert assessors.

For the evaluation of the concept extraction component, The "exact span" and "overlapping span" evaluation settings refer to the case where the automatically identified span is identical to the gold standard span boundaries, and that the identified span overlaps with the gold standard span boundaries, respectively.

3.5 Experimental Setup

The ShARe/CLEF data set only contains disorder concepts. Hence, the annotation systems were guided to look for disorder concepts only. Due to the annotation guideline of ShARe/CLEF data set (Suominen et al., 2013), a concept is in the disorder semantic group if it belonged to one of the following UMLS semantic types: 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, and Signs and Symptoms. Occurrences of "*CUI-less*" spans and concepts in the gold standard were removed from the data set as we cannot expect the annotation systems to find appropriate concepts for disorder text spans if appropriate concepts cannot be found by a human expert.

Table 3 shows the settings of the annotation systems. These parameters can be used to reproduce the results that are reported in this paper. It can be observed that MetaMap, NCBO, and QuickUMLS systems were restricted to the above-mentioned UMLS semantic types. Since Ontoserver does not provide options for such restriction, we filter the output of this system to only those semantic types in a post-processing step. In addition, Ontoserver's annotations are based on SNOMED CT concept IDs. Since the annotations in the data set are UMLS concept IDs, the resulting SNOMED CT IDs were mapped to UMLS concept IDs using NLM's Metathesaurus mapping table [3].

The Stanford CoreNLP toolkit was applied to extract noun phrases from the clinical docu-

[3] Version 2015AB: https://www.ncbi.nlm.nih.gov/books/NBK9685

Table 3: Annotation System Settings.

	System Parameters
Metamap	-J acab,comd,anab,cgab,dsyn,emod,inpo,mobd,neop,patf,sosy, -R SNOMEDCT_US, -q
NCBO	include=prefLabel,cui, ontologies=SNOMEDCT, exclude_numbers=true, longest_only=true, semantic_types=T020,T049,T190,T019,T047,T050,T037,T048,T191,T046,T184
Ontoserver	findConceptsByTermPrefixes, versionedId=http://snomed.info/sct/32506021000036107/version/20160731
QuickUMLS	threshold=0.7, window=5, similarity_name=jaccard, accepted_semtypes='T020','T049','T190','T019','T047','T050','T037','T048','T191','T046','T184'

ments (Manning et al., 2014). In this approach, the resulting parse tree generated from each document was processed to extract the noun phrases (NPs) from the associated subtrees of clauses of sentences.

The MALLET implementation of CRF was used in this paper to train a concept recogniser model (McCallum, 2002). The text spans of disorder concepts from the ShARe/CLEF training data set was used to train the CRF model. The data was converted into BIO format (Begin/Inside/Outside of spans) in order to have an appropriate formulation of concepts with multiple tokens.

4 Results

Table 4 presents the performance of the annotation systems. The first column of results shows system results when the whole document was used as input. The results here would reflect the end-to-end annotation system for both their built-in concept extraction and concept recognition components of the system. MetaMap achieved the highest results with 0.5948 F1-score followed by QuickUMLS and then NCBO. Despite NCBO having the lowest F1-score of the three systems, its precision was considerably higher than MetaMap and QuickUMLS. Ontoserver currently only supports the annotation of short phrases and does not have a built-in concept extraction module to support annotations at a document level.

To further investigate the effectiveness of an annotation system's concept recognition component, the input to the annotation systems were controlled by providing each system the same spans of text. The second column of results in Table 4 shows the results when spans from the gold standard dataset were used as input into the annotators. The remaining columns show the performance of the annotation systems when input spans were generated by the noun phrase parser and the CRF model were used as input.

As expected, system performance on the gold standard chunks achieved the highest results compared to other concept extraction techniques. This simulated the upper bounds of these annotation systems as the human expert generated spans of text were used as input to the systems. The best performing concept recognition system in this setting was Ontoserver with 0.7426 F1-score. QuickUMLS and MetaMap achieved comparable results of 0.7409 and 0.7321 F1-score respectively. Noteworthy was Ontoserver's ability to achieve a very high precision of 0.9058, while QuickUMLS achieved the best recall (i.e., 6893).

For the input spans generated by the noun phrase parser and the CRF model, a similar pattern could be observed in the performance of the systems: MetaMap and QuickUMLS achieved higher F1-scores while NCBO and Ontoserver showed similar performance. Again, Ontoserver achieved the highest precision, particularly when applied to the span generated by the noun phrase parser (precision = 0.6305).

Concept extraction techniques generated candidate spans of text to input into the concept recognition component of the systems. The results suggest that the concept extraction technique greatly impacted the performance of the concept recognition component. To further investigate this impact, Table 5 shows the evaluation of the two concept extraction approaches against the gold standard text spans. For some application, it may be sufficient to identify overlapping rather than exact spans. Therefore, two evaluation scenarios (i.e., *Exact* and *Overlapping*, as described in Section 3.4) were employed to report the results. The results show that the concept extraction approaches studied follow a naive methodology and were far from optimal. The results for both noun phrase generation approaches, however, show that if the text span evaluation criteria were relaxed to overlapping spans then a significant improvement can be achieved in both precision and recall results.

Table 4: Concept recognition results. For whole documents as input, MetaMap and QuickUMLS achieved higher overall F1 scores, while NCBO showed higher precision. Over the various noun phrases, systems showed much superior results on the gold standard input spans with Ontoserver, in general, achieving the highest precision and QuickUMLS achieving the highest recall.

	Document input			**Span input (via concept extraction)**								
	(built-in concept extractor)			**Gold standard**			**Noun phrase parser**			**CRF**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
MetaMap	0.5650	0.6278	0.5948	0.8076	0.6695	0.7321	0.5027	0.4903	0.4964	0.3702	0.0401	0.0723
NCBO	0.6364	0.3742	0.4712	0.7679	0.3758	0.5047	0.5789	0.2982	0.3936	0.3767	0.0226	0.0426
Ontoserver	-	-	-	0.9058	0.6292	0.7426	0.6305	0.3335	0.4363	0.3322	0.0267	0.0495
QuickUMLS	0.5140	0.6197	0.5619	0.8008	0.6893	0.7409	0.4622	0.5075	0.4838	0.3518	0.0406	0.0729

Table 5: Performance of concept extraction approaches in identifying the gold standard text spans. "Exact Spans" and "Overlapping Spans" refer to the case where the automatically identified span is identical to the gold standard span boundary, and that the identified span overlaps with the gold standard span boundaries, respectively. Results show that concept extraction performance is very poor but significant improvements can be achieved when the text span evaluation criteria was relaxed to overlapping spans.

	Precision	**Recall**	**F1**
		Exact Spans	
Noun phrase parser	0.0686	0.4986	0.1206
CRF	0.0517	0.0443	0.0477
		Overlapping Spans	
Noun phrase parser	0.1262	0.9334	0.2224
CRF	0.1884	0.1608	0.1735

To assist in the analysis of the concept extraction and concept recognition components of the systems, Table 6 was included to show the number of input text spans and the number of concept annotations output by each of the annotation systems. It can be observed that the noun phrase parser generates a large number of candidate spans (i.e., all noun phrases in a document), which leads to higher recall in both exact and overlapping text span scenarios (0.4986 and 0.9334, respectively) but low precision (0.0686 and 0.1262). On the other hand, the CRF model generated fewer candidates and achieved poorer results, especially in the exact text span scenario.

5 Discussion

Annotation systems perform two primary steps: concept extraction and concept recognition. While most previous evaluations considered the end-to-end process (Jonquet et al., 2009a; Aronson and Lang, 2010; Groza et al., 2013c; Nunes et al., 2013; Mirhosseini et al., 2014), this papers attempts to consider the impact of these two components separately. The findings are that the concept extraction component significantly impacts the concept recognition phase. One reason for this was that the various concept extraction methods (noun phrase parser, CRF and the built-in methods within each annotator) all produced widely varying spans of text. There was a large difference in the performance between using the gold standard span, which represent an upper bound, and the spans produced by concept extraction methods. The built-in concept extraction methods all performed better than the naive noun phrase parsing and CRF methodology. Therefore we, conjecture that the noun phrase parser and CRF start to show promise when the text span evaluation criteria was relaxed to overlapping spans. Despite this, there was less variation in different concept recognition methods for the same spans of text. The lesson here is that efforts to improve annotation systems are best directed toward improving concept extraction.

The concept recognition results show that some methods were optimal in terms of precision (e.g., Ontoserver), while others were optimal in terms of recall (e.g., QuickUMLS). There are different use cases for concept annotation systems — some precision focused (e.g, accurate coding of diagnoses according to medical classification systems for reimbursement purposes where incorrect codes could lead to substantial penalties (Pestian et al., 2007)) and some recall focused (e.g. searching pa-

Table 6: Number of output annotations by the systems over the test set. NCBO's built-in concept extractor found far less concepts compared to MetaMap and QuickUMLS. In addition, Noun phrase parser generated a large number of candidate input spans while the CRF model generated fewer candidates.

	Built-in concept extractor	**Gold Standard**	**Noun phrase parser**	**CRF**
# Input spans	-	3,610	26,113	3,074
MetaMap	4,599	3,456	4,036	445
NCBO	2,246	1,874	1,963	231
Ontoserver	-	2,499	1,900	289
QuickUMLS	4,331	3,103	3,944	415

tient records for rare diseases where clinicians are concerned with trying to get as high recall as possible, and will tolerate lower precision results). To facilitate these different use cases it would be advantageous to configure the annotation system to optimise for either precision or recall. This may involve adapting the system to use different concept extraction or concept recognition methods. In general, it would be advantageous, both from a system design and system evaluation perspective, to decouple the concept extraction and concept recognition component of such systems.

5.1 Future Work

The medical concept annotation systems studied were observed to comprise of concept extraction and concept recognition components with different levels of strengths (e.g., NCBO's concept extraction module showed less success than its concept recognition module – resulting in low recall but considerable precision). Investigating the effectiveness of the integration of these components across annotation systems should see gains in the overall performances. For example, using the QuickUMLS concept extractor (as it resulted in the best recall) as inputs to the Ontoserver concept recogniser (as it showed the highest precision). Furthermore, an ensemble of these systems working together may also show promising results (Kang et al., 2012). For example, a voting system can be designed to enrich the final annotations with the best outcomes of different systems.

A thorough investigation into the effectiveness and efficiency of annotation systems including evaluations of systems for recognising concepts beyond disorders is also warranted. Comparison of other dimensions, such as execution time, robustness in terms of domain (e.g., radiology, pathology, emergency) and type of input clinical document (e.g., discharge letter vs progress notes), and larger datasets (e.g., i2b2 (Uzuner et al., 2011) or CADEC (Karimi et al., 2015) corpora), and more detailed comparison of concept extraction and recognition components (e.g., effect of overlapping spans on concept recognition) will all be the subject of ongoing work.

6 Conclusion

This paper investigated and evaluated four annotation systems (i.e., MetaMap, NCBO, Ontoserver, and QuickUMLS). The focus was on evaluating and assessing the performances of annotation systems on annotating clinical free-text documents. Concept extraction and concept recognition, which are two main components of a concept annotation system, were independently evaluated in order to provide an in-depth comparison of their performances. The experimental results showed that each annotator exhibited varied performance and that the text spans output by the concept extraction component of an annotation system significantly impacts on the performance of the concept recognition and overall end-to-end performance of the system.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Nigel Collier, Anika Oellrich, and Tudor Groza. 2014. Concept selection for phenotypes and disease-related annota-tions using support vector machines. In *Proc. PhenoDay and Bio-Ontologies at ISMB 2014*.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC bioinformatics*, 15(1):1.

Tudor Groza, Hamed Hassanzadeh, and Jane Hunter. 2013a. Recognizing scientific artifacts in biomedical literature. *Biomedical informatics insights*, 6:15.

Tudor Groza, Jane Hunter, and Andreas Zankl. 2013b. Mining skeletal phenotype descriptions from scientific literature. *PloS one*, 8(2).

Tudor Groza, Anika Oellrich, and Nigel Collier. 2013c. Using silver and semi-gold standard corpora to compare open named entity recognisers. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 481–485. IEEE.

Hamed Hassanzadeh and Mohammadreza Keyvanpour. 2013. A two-phase hybrid of semi-supervised and active learning approach for sequence labeling. *Intelligent Data Analysis*, 17(2):251–270.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170.

Clement Jonquet, Nigam Shah, and Mark Musen. 2009a. The open biomedical annotator. In *AMIA summit on translational bioinformatics*, pages 56–60.

Clement Jonquet, Nigam Shah, Cherie Youn, Chris Callendar, Margaret-Anne Storey, and M Musen. 2009b. Ncbo annotator: semantic annotation of biomedical data. In *International Semantic Web Conference, Poster and Demo session*, volume 110.

Ning Kang, Zubair Afzal, Bharat Singh, Erik M Van Mulligen, and Jan A Kors. 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics*, 45(3):423–428.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

E Keith et al. 1998. The unified medical language system: Toward a collaborative approach for solving terminological problems. *JAMIA*, 5:12–16.

Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2016. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, 23(2):289–296.

Bevan Koopman. 2014. *Semantic Search as Inference: Applications in Health Informatics*. Ph.D. thesis, Queensland University of Technology, Brisbane, Australia.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Simon McBride, Michael Lawley, Hugo Leroux, and Simon Gibson. 2012. Using australian medicines terminology (amt) and snomed ct-au to better support clinical research. In *Studies in Health Technology and Informatics*, pages 144–149.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Shahin Mirhosseini, Guido Zuccon, Bevan Koopman, Anthony Nguyen, and Michael Lawley. 2014. Medical free-text to concept mapping as an information retrieval problem. In *Proceedings of the 2014 Australasian Document Computing Symposium*, page 93. ACM.

Tiago Nunes, David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Becas: biomedical concept recognition services and visualization. *Bioinformatics*, 29(15):1915–1916.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 851–859. Association for Computational Linguistics.

John P Pestian, Christopher Brew, Paweł Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 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.

Nigam H Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie P Chiang, and Mark A Musen. 2009. Comparison of concept recognizers for building the open biomedical annotator. *BMC bioinformatics*, 10(Suppl 9):S14.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Michael A Tanenblatt, Anni Coden, and Igor L Sominsky. 2010. The conceptmapper approach to named entity recognition. In *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC?10)*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Guido Zuccon, Bevan Koopman, Anthony Nguyen, Deanne Vickers, and Luke Butt. 2012. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, Dunedin, New Zealand, December.

Guido Zuccon, Amol S Wagholikar, Anthony N Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. 2013. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. *AMIA Summits on Translational Science Proceedings*, 2013:300.