Medical Free-Text to Concept Mapping as an Information Retrieval Problem

Shahin Mirhosseini[1], Guido Zuccon[1], Bevan Koopman[2,1], Anthony Nguyen[2], Michael Lawley[2]

[1]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia
[2]Australian e-Health Research Centre, CSIRO, Brisbane, Australia

shahin.mirhosseini@connect.qut.edu.au, g.zuccon@qut.edu.au, {bevan.koopman, anthony.nguyen, michael.lawley}@csiro.au

ABSTRACT

Concept mapping involves determining relevant concepts from a free-text input, where concepts are defined in an external reference ontology. This is an important process that underpins many applications for clinical information reporting, derivation of phenotypic descriptions, and a number of state-of-the-art medical information retrieval methods. Concept mapping can be cast into an information retrieval (IR) problem: free-text mentions are treated as queries and concepts from a reference ontology as the documents to be indexed and retrieved. This paper presents an empirical investigation applying general-purpose IR techniques for concept mapping in the medical domain. A dataset used for evaluating medical information extraction is adapted to measure the effectiveness of the considered IR approaches. Standard IR approaches used here are contrasted with the effectiveness of two established benchmark methods specifically developed for medical concept mapping. The empirical findings show that the IR approaches are comparable with one benchmark method but well below the best benchmark.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Experimentation

Keywords

Concept Mapping, Information Extraction

1. INTRODUCTION

Recognising mentions of medical concepts in free-text is an important task for medical Natural Language Processing (NLP) and is often critical to enable further activities such as clinical information analysis and reporting [12], derivation of phenotypic descriptions [5, 4] and medical information retrieval [11, 6]. The automatic recognition of medical concepts can be divided into two sub-tasks: concept extraction and concept mapping. *Concept extraction* refers to the identification of text spans that refer to entities of interest, such as medical problems, disorders, abnormalities, etc. *Concept mapping* refers to the process of identifying the relevant concept (or concepts) referred to by the text span, where the concept is indicated by its identifier within an ontology or terminology resource, e.g., the UMLS and SNOMED CT. Often concept extraction and concept mapping are combined into an overall *concept recognition* process.

Considerable research effort has been directed in developing methods and systems for automatic *medical* concept recognition from free-text. The current state-of-the-art systems consist of specialised clinical NLP pipelines that combine linguistic, statistical, and rule-based techniques. For example, the Metamap system [1] is a widely used tool developed by the National Library of Medicine for medical concept recognition using the UMLS metathesaurus. Metamap uses linguistic and statistical methods, combined into a pipeline that includes sentence and boundary detection, tokenisation, part-of-speech tagging, abbreviation and acronym identification and expansion, dictionary lookup, shallow parsing, and word sense disambiguation. These pipelines are generally capable of performing both concept extraction and concept mapping.

In this paper, we focus on the problem of concept mapping: given a (often short) free-text fragment that refers to a mention of a concept of interest, find the corresponding concept identifier (CUI) from a reference ontology. Thus, we assume that a system that performs concept extraction has been already applied and we consider only the second step of the concept recognition problem. We translate the problem of concept mapping into an information retrieval problem, where free-text mentions of concepts are treated as queries and concepts from a reference ontology are treated as the documents to be indexed and retrieved.

The aim of this paper is to evaluate the effectiveness of standard retrieval models applied to the concept mapping problem, and contrast this to the effectiveness obtained by more established approaches that combine linguistic, statistical and rule-based methods devised specifically for this task. Previous research has compared the effectiveness of dedicated clinical NLP pipelines for concept mapping from a classification perspective [10, 4]. This evaluation instead is

(c) 2014 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
ADCS'14, November 27–28 2014, Melbourne, VIC, Australia.
Copyright 2014 ACM 978-1-4503-3000-8/14/11 ...$15.00
http://dx.doi.org/10.1145/2682862.2682880.

conducted to determine whether standard IR systems provide comparable effectiveness in concept mapping to more expensive (both in terms of computational costs and in terms of development costs), specialised clinical NLP pipelines. This initial investigation can be used as a first step towards the development of concept mapping approaches based on IR models that only rely on shallow linguistic techniques (e.g., stemming and stop-word removal) and word-count statistics. A by-product of this investigation is the evaluation of the approaches using a rank-based perspective, and thus rank-based measures, rather than the traditional classification approach (and thus measuring true/false positives and true/false negatives) used in previous evaluations [10, 4].

2. IR APPROACH TO CONCEPT MAPPING

Figure 1(a) shows an example of sentence containing a free-text mention of a medical concept, along with the expert annotation for the mention that is considered for relevance assessment. The corresponding concept, as it is recorded in the SNOMED CT ontology, is reported in Figure 1(b).

In this paper, a free-text mention of a medical concept is treated as the query, while the concepts in the reference ontology are treated as documents (thus, concept-documents) that are searched to find the relevant concept (or concepts) given the query text. Concepts in the reference ontology are composed of a concept description and a concept unique identifier or CUI. Often concepts in resources such as SNOMED CT include synonyms or alternative concept descriptions, their semantic type (e.g., finding, disease or syndrome, procedure), and their relations with other concepts (relationship types include for example is-a, finding site, causative agent, etc.). These could be used to improve the document representations and thus be exploited for retrieval.

Formally, a free-text mention of a medical concept is modelled as a sequence Q of one or more terms $t_1, \ldots, t_n$. A concept in a medical ontology is modelled as a document C, which is composed of a sequence of one or more terms $t_1, \ldots t_m$. Alternatively, the document C may be split in different fields, each containing sequences of terms (in this case a document is modelled as a set of fields, each field containing one or more sequences $t_1, \ldots, t_m$). These fields would correspond to concept description, synonyms and alternative descriptions, semantic types, relations, etc.

The use of sequences and fields allows for the use of proximity retrieval models, like MRF [8], and field-based retrieval approaches, like BM25F [9]. In the initial investigation presented in this paper however, we treat concepts as being formed by bag-of-words rather than sequences and thus we reduce the representation of a document C to the set of terms $\{t_1, \ldots, t_m\}$. Similarly, we leave the study of field-based retrieval approaches to future work, and treat the fields *FullySpecifiedName* and *Descriptions(Synonyms)* as forming the concept-documents (identified by the *conceptId* field). Other fields that are not related to the direct description of the concept may have been included in the concept-document representation, e.g., the field *DefiningRelationships* in SNOMED CT, which includes the fully specified names of the concepts that are connected to the current concept in the ontology. The influence of these fields on retrieval effectiveness is left to future work; note, however, that state-of-the-art clinical NLP pipelines like NCBO Annotator[1] do take into account this information.

[1] `http://bioportal.bioontology.org/annotator`, last visited 29/09/2014.

25064002

```
... the patient had headaches and was home ...
```

(a) Example of a free-text sentence and a concept mention, with the corresponding SNOMED CT concept annotation.

CONCEPT ID:	25064002
FULLY SPECIFIED NAME:	Headache (finding)
SYNONYMS:	HA - Headache Headache Cephalalgia Head pain Pain in head Cephalodynia Cephalgia
PREFERRED TERMS:	GB English : Headache US English : Headache

(b) The information encoded in SNOMED CT for concept 25064002 (Headache).

Figure 1: Example of free-text mention of a medical concept with the corresponding concept information from the SNOMED CT medical ontology.

The problem of medical concept mapping can then now be cast in a retrieval problem: given a query Q, retrieve the concept-documents $C_1 \ldots, C_k$ from the index, obtained by processing the reference ontology, such that the retrieved documents are relevant to the provided query. Concept mappings manually assigned to free-text mentions by clinical coders or expert annotators can be used as relevance assessments to evaluate the quality of the mapping methods.

General purpose information retrieval methods can be tested on the concept mapping problem. These methods represent a low cost alternative to standalone, complex systems purposely developed for the task of concept extraction in the medical domain. In this paper we consider standard IR baseline methods: TF-IDF, BM25, Jelinek-Mercer language model (JMLM) and Dirichlet Language model (DLM).

3. EXPERIMENTAL SETUP

3.1 Data and Test Collection

Queries and relevance assessments were adapted from the ShARe/CLEF 2013 eHealth Evaluation Lab dataset [10] aimed at evaluating concept recognition systems, where concepts represented problems and disorders mentioned in free-text clinical reports from a U.S. intensive care unit. The training dataset comprised 200 clinical reports, with free-text concept mentions annotated with the corresponding UMLS concept identifiers restricted to the SNOMED CT terminology. The test dataset (containing 100 reports) could not be used for our study because the corresponding annotations were not publicly released by the CLEF organisers.

For the evaluation presented in this paper, we considered only the spans of free-text that had been annotated with concept identifiers; these formed the queries used by the retrieval system to generate mappings (duplicate queries were removed). Some free-text spans were annotated as corresponding to concepts with no associated unique concepts identifiers (called CUI-less) — these were removed from the evaluation queries. This left 1,669 unique queries, along

with the corresponding UMLS concept identifiers restricted to the SNOMED CT terminology. Finally, UMLS concept identifiers were translated to SNOMED CT concept identifiers using the relevant translation tables. Each UMLS CUI corresponded to one or more SNOMED CT concepts. These were treated as the relevant documents that a concept mapping system should retrieve. Queries and relevance assessments are made available at `anonymised`. Our evaluation was similar to the original CLEF 2013 Task 1b evaluation, but it differed because we matched SNOMED CT rather than UMLS concepts and we considered a rank-based evaluation rather than a true/false positive and true/false negative evaluation.

As described in Section 2, concepts from the SNOMED CT medical terminology [3] release 20140731, were considered as the unit of retrieval, i.e., the documents that were indexed and retrieved by the tested systems.

3.2 Evaluation Measures

Generally a user of a concept mapping system is interested in obtaining one CUI (concept) for each of the text spans entered, even if more than one CUI is applicable to that text span. To this aim we evaluated the rankings of candidate concept mappings produced by the systems using reciprocal rank (RR), that is, the reciprocal rank of the first relevant retrieved document (concept). We also report success@k, which measures whether a relevant document has been retrieved up to a cut-off k ($k = 1, 5, 10$).

3.3 IR Systems and NLP Benchmarks

The implementations of TF-IDF, BM25, JMLM and DLM provided in the Apache Lucene 4 software package [2] were used for the empirical evaluation reported in this paper. For each method, we only recorded the top 100 results.

The retrieval effectiveness of the standard information retrieval baselines was compared with the effectiveness of two specialised concept mapping systems: Metamap and Ontoserver.

Metamap [1] outputs UMLS concepts using a combination of linguistic and statistical methods within a staged clinical NLP pipeline. The same process of converting SNOMED CT concepts to UMLS concepts described in Section 3.1 was utilised to convert Metamap's output into a ranking of SNOMED CT concepts.[2]

Ontoserver [7] is a concept mapping tool developed at the Australian e-Health Research Centre and natively returns SNOMED CT concepts given a free-tex query. Ontoserver exploits a purposely-tuned retrieval function and linguistic capabilities such as spell checking, restrictions and inferences on the source ontology. In this paper we used version 2.3.0 of Ontoserver, which is publicly available at `http://ontoserver.csiro.au:8080/`.

3.4 Parameter Tuning

The investigated IR methods (with the exception of TF-IDF) have a number of parameters that require tuning. We perform two explorations of the parameter space.

Firstly, we performed a linear search (grid-search for BM25) of the parameter space to find the parameter settings that provided the overall best performance on the query set. The studied parameter values were BM25: $b \in [0, 1]$ with steps of 0.1, $k_1 \in [0, 2]$ with step of 0.1; JMLM: $\lambda \in [0, 1]$ with steps of 0.05; DLM: $\mu \in [100, 3000]$ with steps of 100. This exploration allowed us to find the best effectiveness that would be achieved by a perfectly tuned system (oracle effectiveness).

Secondly, we set parameter values according to a 10-fold cross validation experiment. This cross validation was repeated 100 times for each retrieval method to allow for random different fold splittings. This exploration allowed for a more realistic tuning of the system that assumes 90% of the data is available for training, while the remaining 10% is withheld for evaluation.

Both benchmark systems (Ontoserver and Metamap) were treated as black-box systems with no access to any degree of tuning. Each system returned less than 100 documents per query.

[2]We used the configuration:
`metamap -K -I -b -R SNOMEDCT_US`.

System	RR	S@1	S@5	S@10
Metamap	0.2723	0.1857	0.3901	0.5285
Ontoserver	0.6166	0.5219	0.7376	0.7879
TF-IDF	0.3823*	0.2888*	0.4883*	0.5674*
BM25	0.3802*	0.2888*	0.4859*	0.5620*
(cross-eval)	0.3800*	0.2887*	0.4858*	0.5620*
JMLM	0.3581*	0.2690*	0.4596*	0.5488*
(cross-eval)	0.3562*	0.2672*	0.4582*	0.5480*
DLM	0.2761	0.1750	0.3853	0.4961*
(cross-eval)	0.2761	0.1750	0.3853	0.4961*

Table 1: Retrieval results on the concept mapping task using benchmark systems and standard IR techniques. All differences between IR techniques and Ontoserver are statistically significant with $p < 2.2 * 10^{-16}$ (paired t-test); statistical significant difference between IR techniques and Metamap are marked with $*$ ($p < 0.01$).

4. RESULTS AND DISCUSSION

Table 1 reports the retrieval effectiveness for both standard IR baselines and benchmark systems. Ontoserver significantly outperformed the other methods, providing a relevant concept-document in the first rank position (S@1) for more than 50% of the queries — about double the success rate of the other techniques. While the IR baselines retrieved a relevant concept within the first ten rank positions (S@10) for about 50% of the queries, Ontoserver retrieved a relevant concept in the first ten rank positions for about 80% of the cases. Notably, S@10 for IR baselines was actually comparable to S@1 for Ontoserver. Interestingly, while the increase in success between rank 5 and 10 for IR baselines was significant, the same increase was of minor effect for Ontoserver. While IR techniques were significantly less effective than Ontoserver, they were comparable with the other baseline of Metamap.

The size of the ranking lists produced by Ontoserver and IR techniques was similar (Ontoserver retrieved on average 82 concepts/query and IR techniques 80 concepts/query), while Metamap returned, on average, only 4.7 concepts/query.

Retrieval effectiveness was significantly reduced by queries for which *no relevant* concept was returned at all: 210 queries for Ontoserver, 450 for IR techniques (TFIDF: 426, BM25: 442, JMLM: 456, DLM: 471) and 749 for Metamap.

There were a number of queries for which systems did not retrieve *any result at all* (i.e., empty result list). Table 2 provides the details of the number of queries with no docu-

	Metamap	Ontoserver	IR
Metamap	191	9/193	41/211
Ontoserver	-	11	9/63
IR	-	-	61

Table 2: The first diagonal of the table reports the number of queries with no retrieved result for each of the systems; the remaining cells report the size of the intersection and of the union of the sets of queries with no retrieved result for each pair of systems.

System	RR	S@1	S@5	S@10
Metamap	0.3015	0.2032	0.4354	0.5941
Ontoserver	0.6315	0.5323	0.7576	0.8111
TF-IDF	0.3959*	0.2967*	0.5069*	0.5920
BM25	0.3925*	0.2953*	0.5048*	0.5852
JMLM	0.3691*	0.2747*	0.4766	0.5714
DLM	0.2914	0.1848	0.4059	0.5227*

Table 3: Retrieval results on the concept mapping task using benchmark systems and standard IR techniques and excluding queries where no result is returned by at least one approach. All differences between IR techniques and benchmark systems are statistically significant with $p < 2.2 * 10^{-16}$ (paired t-test); statistical significant difference between IR techniques and Metamap are marked with $*$ ($p < 0.01$).

ment returned by each approach, along with the size of the intersection and union of the sets of queries with no result returned when systems were pairwise compared. Overall, there were 212 queries for which at least one system did not return a result and 43 queries for which no system returned any results. This highlights that although all systems suffer from not retrieving results for certain queries — more so for the IR approaches and Metamap; thus these approaches are characterised by poor matching (recall). However, IR approaches did retrieve concepts for a minority of queries for which Ontoserver retrieved no results.

Table 3 reports the retrieval effectiveness of the methods on the queries for which all systems returned at least one result (1,457 queries): while the effectiveness was naturally higher than that reported in Table 1 (because queries with 0 effectiveness are removed), the results exhibit the same trends observed in the previous analysis. Results of the cross-validation experiments are omitted because their value was similar to the oracle tuning, as it was the case in Table 1. These results highlight that not only IR approaches suffer from poor matching when compared to Ontoserver, but they also exhibit poor ranking choices (precision).

5. FUTURE WORK AND CONCLUSION

In this paper we have investigated the effectiveness of general-purpose, baseline IR approaches on the task of (medical) concept mapping, i.e., the labelling of a free-text extract with a concept identifier from a reference ontology. The concept mapping problem was cast into a retrieval problem and the effectiveness of the IR methods was compared with the results obtained by two complex, comprehensive and dedicated clinical NLP pipelines. As a by-product, the mapping problem was evaluated from a ranked-based standpoint rather than the traditional classification standpoint used in previous work [10].

The empirical results suggested that, although the IR methods are comparable with one of the benchmark methods (Metamap), state-of-the-art custom benchmark methods (Ontoserver) are still far more effective than the standard IR approaches. In addition, we found that probabilistic language modelling approaches are actually worse than the heuristic methods (TF-IDF and BM25). Other specific IR models, such as the translational language models, might be better suited to this task because they may also consider reformulations of the free-text terms that match relevant concepts.

Acknowledgements. Shahin Mirhosseini was supported by a CSIRO student scholarship; experiments were executed on equipment purchased with the support of the QUT SEF Large Equipment Grant #94.

6. REFERENCES

[1] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] A. Białecki, R. Muir, and G. Ingersoll. Apache lucene 4. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 17–24, 2012.

[3] L. Bos et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Stud Health Technol Inform*, 121:279–290, 2006.

[4] N. Collier, A. Oellrich, and T. Groza. Concept selection for phenotypes and disease-related annota-tions using support vector machines. In *Proc. PhenoDay and Bio-Ontologies at ISMB 2014*, 2014.

[5] T. Groza, J. Hunter, and A. Zankl. Mining skeletal phenotype descriptions from scientific literature. *PloS one*, 8(2), 2013.

[6] B. Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, Queensland University of Technology, Brisbane, Australia, 2014.

[7] S. McBride, M. Lawley, H. Leroux, and S. Gibson. Using australian medicines terminology (AMT) and SNOMED CT-AU to better support clinical research. *Studies in health technology and informatics*, 178:144–149, 2012.

[8] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[9] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[10] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer, 2013.

[11] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *ADCS'12*, pages 111–114, 2012.

[12] G. Zuccon, A. S. Wagholikar, A. N. Nguyen, L. Butt, K. Chu, S. Martin, and J. Greenslade. 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, 2013.