Medical Concept Resolution

Nitish Aggarwal°, Ken Barker*, and Chris Welty[a]

*IBM Watson Research, Yorktown Heights, NY, USA
kjbarker@us.ibm.com
°Insight-Centre, National University of Ireland Galway, Ireland
nitish.aggarwal@insight-centre.org
[a]Google Inc.
chris.welty@gmail.com

Abstract. In this paper, we present a problem that we refer to as Medical Concept Resolution for finding concept identifiers in a large knowledge base, given medical terms mentioned in a text. We define the problem with its unique features and novel algorithms to address it. We compare performance to MetaMap and find distinct and complementary behavior.

1 Introduction

Linking phrases in text to concepts in a knowledge base (KB) such as the Unified Medical Language System (UMLS) [1] is especially difficult when it consists of multiple, merged source taxonomies. Although standard open domain entity linking (EL) [1, 2] deals with finding concepts or entities in a KB that match a given phrase (mention) in a text, it assumes that there is only one correct match at a time for a given mention. However, in domain specific EL, there is often more than one correct match for a given mention depending upon the context. Moreover, open domain entity linking finds the entries in a KB that match the mention text exactly, so there is no need to "discover" the candidate entries with partial matches.

In this paper, we present Medical Concept Resolution (MCR), a system for finding a concept (more precisely, a concept's Concept Unique Identifier (CUI)) in UMLS, for medical terms mentioned in text. We describe three unique challenges in mapping text spans to CUIs in UMLS, but we note that the properties likely apply to any large concept repository, especially those with concepts from multiple different sources.

Discovery In a large concept repository such as UMLS, the hierarchy of concepts and the labels supplied for them can be arbitrary. Frequently, there is no concept whose label matches a span of text exactly. Given the variability of medical language, the span detection problem for medical terms is significantly more difficult than typical entity linking tasks. For these reasons it can be very difficult to determine whether a concept exists in the repository. For example, there is no UMLS concept with a label that matches the text span "distended jugular

[1] UMLS: http://www.nlm.nih.gov/research/umls/

vein". Relaxing the term order (e.g., "jugular vein distended") and substituting synonyms (e.g., "engorged jugular vein"), we still find no UMLS concepts with matching labels. There are, however, concepts in UMLS for "jugular vein distention" and "jugular venous distension".

Multiplicity In concept repositories that combine multiple sources, there are often multiple entries for the same domain concept. So even after a concept is discovered, there may be other appropriate concepts. In UMLS there are many CUIs with the label "pain". Therefore, MCR has to deal with concept disambiguation similar to open domain entity linking.

Granularity Even when there are close superficial matches between concept labels and text spans, more specific concepts often exist that capture more of the semantics for the span, given a larger context. For instance, the CUI for "jugular vein distention" may match a text span perfectly, but the more specific CUI for "jugular venous distension with inspiration" may be more appropriate when considering a larger context.

2 Approach

Our approach for mapping text spans ("mentions") to UMLS CUIs consists of two main steps: candidate over-generation and candidate reranking. For obtaining the mentions from text, we use a CRF-based method for extracting medical terms [3].

2.1 Candidate Over-generation

The intuition behind over-generation is that there may be a mismatch between the mentions in text and the variant labels of target CUIs. Over-generation finds all CUIs having any variant containing any of the tokens in the mention text. The resulting candidates include many irrelevant CUIs, but also relevant CUIs that are more general than the mention and CUIs that are more specific. For example, candidates for the string "pupil miosis" include the CUIs for pupil, miosis, school pupil, congenital miosis, pupillary miosis of the right eye, etc. Candidate over-generation may produce an enormous number of candidates. For efficiency, only those candidates that are most similar to the mention are considered in the subsequent reranking step. The most similar candidates are determined by inverse document frequency (IDF) rank weighted similarity of their labels to the mention text. The n tokens in the original mention are ranked according to their IDF in a medical corpus. The ranks are converted to weights $w_i = r_i/n$, where r_i is the IDF rank of the i^{th} token. The least frequent (highest IDF) token has rank n, the most frequent token, rank 1. For example, in the phrase "pupillary miosis" of the right eye, the weighted word vector would be: [pupillary:0.83, miosis:1.0, of:0.33, the:0.17, right:0.5, eye:0.67]. The weights are used in calculating weighted cosine similarity between the mention tokens and each candidate CUI variant. All candidates (up to 100) having variants with similarity to the mention text above a threshold are kept for the reranking step. Converting IDF values to rank weights normalizes and smoothes the IDF values.

2.2 Candidate Reranking

The candidate CUIs are reranked by measuring the similarity between mention context and candidate context. The mention context is a relatively large window of text (averaging 6-7 sentences) surrounding the mention. Both the mention context and the candidate context are treated as bags of words in computing the cosine similarity. We considered different context window sizes. The results reported below use the full sentence containing the mention span (Sm). On the candidate CUI side, we implemented following three context generators:

Gloss-Based Medical Concept Resolution (gbmcr) Two contributing sources in UMLS are MeSH and NCI, which together contribute definitions for only roughly 3% of the concepts. Nevertheless, for 86% of the medical concept mentions in our experiments, at least one of the filtered candidate concepts had at least one MeSH or NCI definition. In gbmcr, candidates are ranked according to the cosine similarity between the words in the mention span (Sm) and the words from the MeSH definition of the candidate, if one exists, or the words from the NCI definition.

Neighbor-Based Medical Concept Resolution (nbmcr) In addition to taxonomic relations, UMLS contains semantic relations. In nbmcr, we consider "neighbor concepts" related to the candidate concept via a subset of "clinically relevant" relations based on the semantic type of the candidate CUI. For example, for candidates of type Disease or Syndrome, we consider related symptoms, treatments, risk factors, etc. Candidates are ranked according to the similarity between the words in Sm and the words in variants of the neighbor CUIs semantically related to the candidate.

Variants-Based Medical Concept Resolution The bag of words of all of the variant labels in UMLS of the candidate CUI make up the vbmcr candidate context.

3 Evaluation

In order to evaluate our MCR methods, we compare their performance to MetaMap (mmap)[2] on a dataset that contains 1,570 medical term spans extracted from 100 short descriptions (averaging roughly 8 sentences, 100 words) of patient scenarios. The MCR algorithms can produce a ranked list of as many CUIs as there are filtered candidates (arbitrarily capped at 100). We included the top three ranked concepts for each factor in the evaluation. Five human judges were randomly assigned roughly 560 factor-CUI mappings each. The assignments did not overlap, but judges also rated 41 mappings in common. The average pairwise kappa for judge agreement was 0.6. For each CUI found for a medical term, judges gave a score of: 0 (inappropriate for the factor), 1 (appropriate for the factor) and 2 (appropriate for the factor and better than concepts scoring 1 for the same factor). The purpose of distinguishing two grades of appropriate is to verify one of the original motivations for MCR: that even when a CUI appropriate to the exact span exists, there are often more specific CUIs that are more appropriate

[2] MetaMap: http://metamap.nlm.nih.gov/

considering more context. For example, consider the sentence “She has pain in the epigastric region and sometimes on the right side of her abdomen”, here, the CUI for “pain” is appropriate, but UMLS also has CUIs for “abdominal pain” and “right sided abdominal pain”.

	Best-is-correct			Appropriate-is-correct		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
mmap	0.614	0.608	0.611	**0.794**	**0.787**	**0.790**
gbmcr	0.450	0.427	0.438	0.628	0.596	0.611
nbmcr	0.567	0.538	0.552	0.624	0.592	0.608
vbmcr	**0.721**	**0.685**	**0.703**	0.762	0.723	0.742

Table 1. Performance

3.1 Results and Discussion

We calculate the precision, recall and F1-measure by considering two settings: *best* (CUIs scoring 2 are correct; CUIs scoring 1 are correct only if there are no CUIs scoring 2); and *appropriate* (CUIs scoring 2 or 1 are correct). We can compare the performance of different methods in a strict environment and a more relaxed one. Table 1 shows that our method vbmcr outperforms all other approaches in strict environment (Best-is-correct). Particularly, vbmcr achieved more than 13% improvement over state of the art method of medical entity linking i.e. MetaMap. The other two MCR methods (nbmcr and gbmcr) did not perform as well. MetaMap achieved the highest scores in the relaxed setting (Appropriate-is-correct). This shows that MCR is able to find more specific concepts and takes into consideration more context; MetaMap performs well when there is a close match between the mention text and CUI variants, and no more specific CUIs exist.

4 Conclusion

We introduced the notion of Medical Concept Resolution (MCR), a knowledge base lookup task in which terms expressed in medical text are identified in a knowledge base. We argued that MCR is more difficult than standard entity linking problems because medical terms themselves are far more composable and contextual, and determining the correct span of text to search for in a knowledge base is more complex. We introduced three aspects of this complexity: discovery, multiplicity, and granularity. Further, we presented a set of new algorithms for performing MCR and showed that our methods outperformed state-of-the-art methods.

References

1. N. Aggarwal and P. Buitelaar. Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*, 2014.
2. H. Ji, R. Grishman, and H. T. Dang. Overview of the tac 2010 knowledge base population track. In *TAC 2010*.
3. C. Wang and J. Fan. Medical relation extraction with manifold models. In *ACL 2014*.