Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-446

Thesaurus-Based Hierarchical Semantic Grouping of Medical Terms in Information Extraction

Yassine LASSOUED[1] and Léa DELERIS
IBM Research, Ireland

Abstract. In this paper we describe a semantic approach for grouping medical terms into a hierarchy of concepts based on the UMLS meta-thesaurus. The context of this work is Medical Recap, a Web system that automatically extracts risk information from PubMed abstracts, and then aggregates this knowledge into dependence graphs or Bayesian networks.

Keywords. Information extraction, Semantic grouping of medical terms, UMLS

1. Introduction

Considering the speed at which medical publications and documents are growing and being made available, manual processing of these resources with respect to extracting information is no longer an option. Instead, medical research is turning towards automatic information extraction tools to extract structured information from natural text. Medical Recap [5] is one such medical information extraction system, developed by IBM Research. The tool extracts structured relationships amongst diseases and risk factors from PubMed abstracts, and then aggregates those into dependence graphs and Bayesian networks.

While the output of information extraction tasks is, as intended, structured (e.g., tabular, graph triples, etc.), the extracted information (field values, graph nodes, etc.) may still be free text values. As such, it may contain multiple representations of similar or related concepts. This makes it difficult to further aggregate the extracted knowledge as is. For instance, in Medical Recap, the output of the information retrieval task is a list of dependence relationships or probability statements, each linking two variables. Variables may be diseases or risk factors. But in all cases they are mere terms extracted from natural text. Several extracted terms, such as "*breast cancer*", "*breast carcinoma*", and "*ER+*", or "*BMI*" and "*body mass index during adulthood*" need to be grouped under common concepts, e.g., Breast Cancer and Body Mass Index. This is essential in order to aggregate the results into a dependence graph or a Bayesian network.

In this paper, we describe a thesaurus-based approach we developed for Medical Recap, aimed to semantically group similar and related terms under a hierarchy of concepts. We rely on the Unified Medical Language System (UMLS) meta-thesaurus as a domain knowledge thesaurus.

[1] Corresponding Author: YLassoue@ie.ibm.com

2. Background and Related Work

Medical Recap is a Web system that automatically extracts medical risk information from text (PubMed abstracts) in the form of dependence or probability relationships such as "***high young adult BMI*** *was* ***associated with*** *decreased* ***premenopausal ER+ cancer***", or "*the OR for* ***breast cancer*** *was* ***1.085*** *(95 % CI: 1.015 to 1.160) for* ***low-level (< 21g/d) alcohol drinkers***".

The extracted information can be viewed as a set of links between variables. In the first example above, Medical Recap may identify "*high young adult BMI*" and "*premenopausal ER+ cancer*" as variables. In the second case, variables are "*breast cancer*" and "*low-level (< 21 g/d) alcohol drinkers*". In order for Medical Recap to be able to aggregate these links into a dependence graph or a Bayesian network, it needs to group the extracted variables semantically into concepts. Figure 1 shows a concrete example of dependence data extracted by Medical Recap.

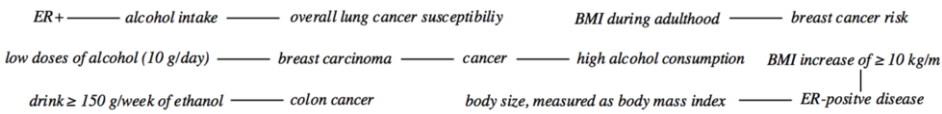


Figure 1. Variable dependence relationships between variable extracted by Medical

The most common way to approach this problem is clustering. Typically clustering relies on the provision of a semantic similarity or relatedness measure between terms or texts to cluster. It then applies a suitable clustering algorithm to group similar terms under clusters. There exists a panoply of similarity measures in the literature. Distributional similarity methods [6, 9, 12] define similarities between words according to their distributions or co-occurrences in large text corpora. Knowledge based semantic methods, such as [10, 11], provide similarities between concepts of a taxonomy or thesaurus (e.g., WordNet), which may then be generalised to text fragments. More recent approaches [1, 2] are based on learned word embeddings as abstract vector representations of words. In the medical domain, knowledge based semantic similarity approaches rely on the Unified Medical Language System (UMLS) meta-thesaurus, MeSH, or the SNOMED-CT ontology, see for instance [2, 7, 8].

In this paper, we explore an alternative to similarity-based clustering suitable for terms, i.e., short text fragments. The objective is to provide an accurate grouping of terms into hierarchical concepts using a domain thesaurus (UMLS). This has the advantage of not depending on the quality of a similarity measure or clustering algorithm. In addition, it allows users to apply term groupings at custom conceptualisation (abstraction) levels depending on their needs.

3. Approach

Our approach relies on the availability of a domain knowledge thesaurus. It uses UMLS to identify concept mentions in the extracted variables. It then selects the most relevant concept mentions, and organises them hierarchically according to their meanings (using both UMLS and the term structures). For example, a possible output for the variables shown in the above example is shown in Figure 2. We describe the sequence of steps that leads to this result in the following sub-sections.

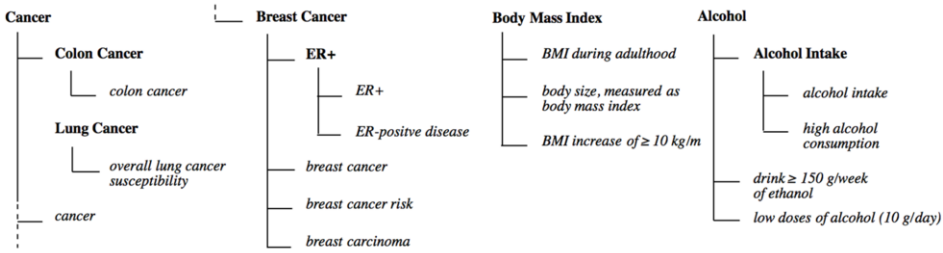


Figure 2. Term Hierarchical Semantic Grouping into UML Concepts

3.1. Identify Concept Mentions

Our input consists of a set of variables, each being a term (c.f. Figure 1). The objective of this phase is to identify the concepts mentioned in each term. We use a hierarchical concept matching approach in which each input term is represented as a hierarchy of sub-terms. We designed two methods for obtaining such sub-term hierarchies:

1. Constituency Parse Tree: We parse the input term using a natural language processing constituency parser, and only keep the relevant nodes of the parse tree, e.g., noun phrases, adjective phrases, etc.
2. N-Gram Lattice: We build the lattice of all the n-grams of the input term, and then remove the n-grams that are syntactically invalid or irrelevant (e.g., starting or ending with a preposition, etc.).

Next, we traverse the so-obtained hierarchy starting from the root (full term) and try to match each node (sub-term) to UMLS concepts. While matching, we only consider UMLS concepts from a predefined set of relevant semantic types (e.g., disease, behaviour, substance, etc.). Figure 2 shows an example of the result of this process. Concept mentions are scored to reflect the confidence that the concept matches the sub-term in question. Only matches with scores above a certain threshold are considered in order to reduce noise. Sub-terms that do not have concept matches are skipped.

body size, measured as body mass index:
- *body size*: C0005901 (**Body Size**)
  - *body*: C0152338 (**BCd**), C0242821 (**Bodies, Human**), C0460148 (**Body**)...
  - *size*: C0221872 (**Sizers**), C0456389 (**025-026 SIZES**), C0522511 (**Has size**)...
- *measured*: C0444706 (**Measured**), C0449768 (**Measured to**), C3641261 (**Not Measured**)...
- *body mass index*: C0005893 (**BMI**), C1305855 (**BMI**), C2240399 (**Body Mass Index**)
  - *body mass*: C0518010 (**Body mass**)
    - *body*: C0152338 (**BCd**), C0242821 (**Bodies, Human**), C0460148 (**Body**)...
    - *mass*: C0577559 (**A mass**), C1306372 (***Mass**), C1414542 (**FBN**), C1546709 (**Mass**)...
  - *index*: C0021200 (**Indexing**), C0600653 (**Indexes**), C0918012 (**Index**)...

Figure 3. Example of a Hierarchical Concept Mapping for the Input Term "*body size, measured as body mass index*"

3.2. Fuse Similar Concepts

From the preceding step, we obtain a set of hierarchical concept mappings. Each node in such a hierarchy consists of a sub-term and a set of scored UMLS concept matches. The objective of this phase is to identify synonyms across all the hierarchies and fuse their concept matches into one common virtual concept. For this purpose, we compare all the nodes across all the hierarchies and group those that share at least one matching

concept. This is performed in a transitive (iterative) way, i.e., if nodes *a* and *b* share a UMLS concept *X*, and if nodes *b* and *c* share another UMLS concept *Y*, then all three nodes (sub-terms) *a*, *b*, and *c* are considered as synonymous. Figure 4, shows an example with two concept mapping hierarchies.

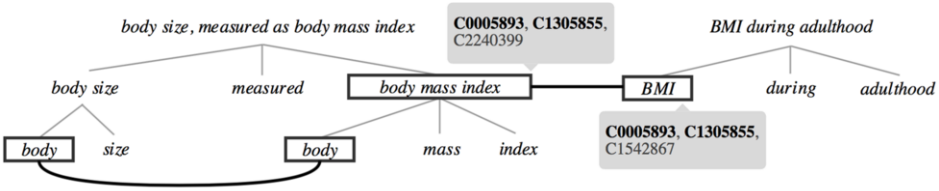
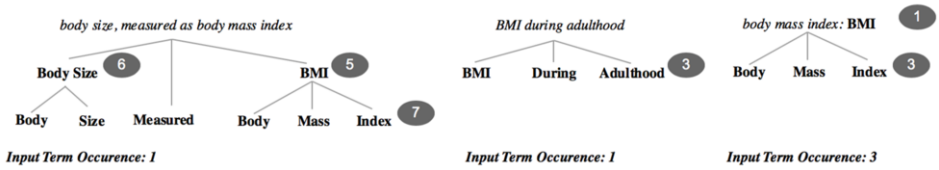


Figure 4. Grouping sub-term nodes sharing at least one concept. Both sub-terms "body mass index" and "BMI" match UMLS concepts C005893 and C1305855. They are considered as synonymous, and they will be represented by the union of their matching concepts: C005893, C1305855, C2240399, and C1542867.

Now, synonyms across the concept mapping hierarchies are identified. For each group of synonymous sub-terms, the sets of their matching UMLS concepts are fused to create one ***virtual concept***. The score of the best UMLS concept match in each node is considered as the virtual concept's score and is preserved for the next step.

3.3. Re-Score Concepts According to Relevance

At this stage, we have a set of hierarchical mapping nodes. Each mapping node consists of a sub-term, a matching virtual concept, and a score. This phase aims to filter out the least relevant nodes carried throughout the previous steps, e.g., "*mass*", "*index*", "*measured*", etc. The objective is to identify the (virtual) concepts that will be included in the final concept hierarchy (c.f., Figure 2). The idea is to boost those concepts that most frequently match longer spans in the input terms, in other words, the nodes that frequently appear in the top levels of the concept mapping hierarchies.



	Initial Score	Average Depth	New Score = Old Score / Average Depth
BMI	1	(1*5 + 1*3 + 3*1) / (1+1+3) = **2.2**	0.45
Body Size	1	6 / 1 = **6**	0.17
Index	1	(1*7 + 3*3) / (1+3) = **4**	0.25

Figure 5. Re-scoring concepts based on their average depths in the n-gram hierarchies: Concept depths are illustrated with the circled numbers. For a term with n tokens, and a sub-term with k tokens, the depth of the sub-term in the term's n-gram hierarchy is n-k+1. Term frequency is taken into consideration.

We start by computing the average depth for each virtual concept in the entire concept mapping hierarchies. Then, we re-score each concept by dividing the initial match score by the concept's average hierarchical depth. In such a way, and as illustrated in Figure 5, "*body mass index*" gets boosted, while "*body size*" or "*index*" receive lower scores. After rescoring, we select in each concept mapping hierarchy the node that has the highest score. For example, the input term "*body size, measured as body mass index*" will be represented by the node "BMI" as this has the highest score.

3.4. Create Hierarchy

At this point, for each term in our initial input, we have an associated preferred concept. The hierarchical links between concept mappings are considered as lexical narrower/broader links between the corresponding concepts. For instance "breast cancer" is considered as narrower (more specific) than "cancer". This provides a basic set of hierarchical relationships obtained lexically from the term hierarchies. We enrich this set using semantic hierarchical (narrower/broader) relationships extracted from UMLS. For each pair of preferred virtual concepts from two input terms, we extract the hierarchical relationships between their respective UMLS concepts. A virtual concept *X* is considered as broader than another, *Y*, if it contains a UMLS concept that is broader (in the transitive sense) than a UMLS concept in *Y*. Using the semantic relationships between the preferred concepts, we now build the hierarchy illustrated in Figure 2.

4. Conclusion

The approach described above for hierarchical semantic grouping of medical terms has been implemented and tested as part of the Medical Recap system. The result is an interactive tool that groups medical terms under a hierarchy of concepts, which then may be further edited by the user. The tool supports flexible term grouping, as it allows the user to select the concept levels at which the final term grouping should be applied.

References

[1] C. Banea, D. Chen, R. Mihalcea, C. Cardie, and J. Wiebe, SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity, in *Proceedings of the 8th International Workshop on Semantic Evaluation SemEval* (2014).

[2] V.N. Garla and C. Brandt, Semantic Similarity in the Biomedical Domain: An Evaluation across Knowledge Sources, *BMC Bioinformatics*, 13(261) (2012).

[3] T. Kenter and M. de Rijke, Short Text Similarity with Word Embeddings, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management CIKM '15* (2015) 1411–1420.

[4] D. Kiela, F. Hill, and S. Clark, Specializing Word Embeddings for Similarity or Relatedness, in *Proceedings of the Empirical Methods in Natural Language Processing Conference EMNLP* (2015).

[5] Y. Lassoued, C. Jochim, S. Deparis, B. Sacaleanu, and L.A. Deleris, Medical Risk Modeling Made Easy, in *Proceedings of the 15th World Congress on Health and Biomedical Informatics MEDINFO* (2015).

[6] D. Lin, An Information-Theoretic Definition of Similarity, in *Proceedings of the 15th International Conference on Machine Learning ICML'98* (1998), 296–304.

[7] B.T. McInnes, T. Pedersen, and S.V. Pakhomov, UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity, in the *Proceedings of the Annual Symposium of the American Medical Informatics Association* (2009).

[8] S. Pakhomov, B.T. McInnes, T. Adam, Y. Liu, T. Pedersen, and G.B. Melton, Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study, in *Proceedings of the Annual Symposium of the American Medical Informatics Association* (2010).

[9] F. Pereira, N. Tishby, and L. Lee, Distributional Clustering of English Words, in *Proceedings of ACL-93* (1993), 183–190.

[10] P. Resnik, Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in *Proceedings of IJCAI-95* (1995), 448–453.

[11] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, Text relatedness based on a word thesaurus, *Journal of Artificial Intelligence Research* **37** (2010), 1–40.

[12] P.D. Turney and P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research* **37** (2010), 141–188.