Evaluating a dictionary of human phenotype terms focusing on rare diseases

Simon Kocbek[1,2], **Toyofumi Fujiwara**[3], **Jin-Dong Kim**[3], **Toshihisa Takagi**[4],
Tudor Groza[1,5]
[1]Kinghorn Center for Clinical Genomics, Garvan Institute of Medical Research, Australia
[2]Dept of Computing and Information Systems, The University of Melbourne, Australia
[3]Database Center for Life Science, ROIS, Tokyo, Japan
[4]Dept Biological Sciences, Gard School of Science, The University of Tokyo, Japan
[5]St Vincent's Clinical School, Faculty of Medicine, UNSW, Australia
skocbek@gmail.com, fujiwara@dbcls.rois.ac.jp,
jdkim@dbcls.rois.ac.jp, tt@bs.s.u-tokyo.ac.jp,
t.groza@garvan.org.au

Abstract

Annotating medical text such as clinical notes with human phenotype descriptors is an important task that can, for example, assist in building patient profiles. To automatically annotate text one usually needs a dictionary of predefined terms. However, do to the variety of human expressiveness, current state-of-the art phenotype concept recognizers and automatic annotators struggle with specific domain issues and challenges. In this paper we present results of annotating gold standard corpus with a dictionary containing lexical variants for the Human Phenotype Ontology terms. The main purpose of the dictionary is to improve the recall of phenotype concept recognition systems. We compare the method with four other approaches and present results.

1 Introduction

Human phenotype descriptions are the composite of one's observable characteristics/traits (e.g., renal hypoplasia, enlarged kidneys, etc.). These descriptions are important for our understanding of genetics and enable the computation and analysis of a varied range of issues related to the genetic and developmental bases of correlated characters (Mabee et al., 2007).

Concept Recognition (CR) is the identification of entities of interest in free text and their resolution to ontological terms with the aim of structuring knowledge from unstructured data. Linking from the literature to ontologies such as the Human Phenotype Ontology (HPO) has gained a substantial interest from the text mining community (e.g., Uzuner et al., 2012; Morgan et al., 2008). Although phenotype CR is similar to other tasks such as gene and protein name normalization, it has its specific domain issues and challenges (Groza et al., 2015). In contrast to gene and protein names, phenotype concepts are characterized by a wide lexical variability. As a result, simple methods like exact matching or standard lexical similarity usually lead to poor results. Additional challenges in performing CR on phenotypes include the use of abbreviations (e.g., *defects in L4-S1*) or of metaphorical expressions (e.g., *hitchhiker thumb*).

A fairly challenging task of phenotype concept recognizers is detecting lexical variants of tokens due to high variety of human expressiveness. For example, detecting similar words with classical similarity metrics such as the Levenshtein distance might group words with different meaning like *zygo-*

Proceedings of the 5th International Workshop on Computational Terminology,
pages 104–109, Osaka, Japan, December 12 2016.

matic (a cheek bone) and *zygomaticus* (cheek muscle) into one lexical cluster (even when using a high similarity threshold). On the other hand, less similar words with same meaning like, for example, irregular nouns (e.g., *phalanx* vs *phalanges,* or *femur* vs *femora*) might be grouped into different clusters.

Therefore, this paper presents results of experiments designed to evaluate a dictionary that tries to address the lexical variability of phenotype terms. Extending dictionaries with new terms has improved performance of, for example, gene phenotype recognizers (Funk et al., 2016). To help improve the performance (focusing on recall) of automatic phenotype CR process, we previously generated a dictionary of lexical variants for all HPO tokens (Kocbek and Groza, 2016), and here we present results of using this dictionary to annotate a gold corpus capturing text spans from 228 abstracts. The latter were manually annotated with Human Phenotype Ontology (HPO) concepts and harmonized by three curators (Groza et al., 2015).

We expect that adding lexical variants will improve the recall of the annotation process, however, we also try to measure the effect of parameter tuning on the precision of the system.

2 Methods

We used the dictionary of lexical variant clusters for all concepts and their synonyms in the HPO. Each HPO term and synonym was then extended with combinations of all words in the corresponding clusters. We automatically annotated the gold standard corpus and compared results of five different approaches.

2.1 The Human Phenotype Ontology and the gold standard corpus

The HPO (Köhler et al., 2014) is often used for the annotation of human phenotypes and offers a tool for large-scale computational analysis of the human phenotype, focusing on rare diseases. The HPO has been used in applications such as linking human diseases to animal models (Washington et al., 2009), describing rare disorders (Firth et al., 2009), or inferring novel drug indications (Gottlieb et al., 2011).

Most terms in the HPO contain descriptions of clinical abnormalities and additional sub-ontologies are provided to describe inheritance patterns, onset/clinical course and modifiers of abnormalities. Each term has a name and can have other synonyms (e.g., "Triangular head shape" is a synonym for "Trigonocephaly"). Each name and synonym may consist of several tokens (e.g., the term "synostosis of some carpal and tarsal bones" has 7 tokens).

Terms in HPO usually follow the Entity-Quality formalism where they combine anatomical entities with qualities (Mungall et al., 2007) For instance, the term "wide anterior fontanelle" describes an anatomical entity "anterior fontanelle" with the quality "wide". Entities can usually be grounded in ontologies such as the Foundational Model of Anatomy (Rosse and Mejino, 2003), while qualities usually belong to the Phenotype and Trait Ontology (Gkoutos et al., 2009). We have previously shown that rich lexical variability comes from the quality part of phenotype terms – due to their widespread usage in common English (Kocbek and Groza, 2016)

The manually annotated HPO gold standard corpus used in this study (Groza et al., 2015) comprises 1,933 annotations in 228 abstracts with an average length of 2,42 tokens per annotation. The gold standard was harmonized by three curators. The corpus covers 460 unique HPO concepts that include abnormalities of nervous system, neoplasms, abnormalities of the integument, and abnormalities of the skeletal system.

2.2 Dictionary construction

We used the HPO released in July 2016 to generate two dictionaries, i.e., collections of labels and their corresponding identifiers. In the first dictionary (Dict1), we extracted labels and their synonyms for each HPO term. This resulted in 25,603 unique dictionary entries that were used as a baseline in our annotation experiments described in Section 2.3. Each label and synonym belonging to the same HPO term were linked to the same corresponding HPO identifier adorned with a postfix. For example, the label "Sclerosis of 5th toe phalanx" and its synonym "Increased bone density in pinky toe bone" for the HPO term with identifier HP:0100929 would have identifiers HP:0100929_0 and HP:0100929_1 respectively.

For the second dictionary (Dict2), we developed a simple tokenizer that broke each name and synonym into series of lower case tokens. The following characters were removed: . / ( ) ' > < : ; and the space and backslash characters were then used as delimiters. We ignored numbers and short tokens (i.e., shorter than 3 characters). Then the NLM Lexical Variant Generator (LVG), 2016 release (The Lexical Systems Group, 2016) was used to create lexical variants for all HPO tokens. This way we created 29,602 variants grouped into 6,480 clusters with average size of 4.57 tokens per cluster. All combinations of token variants were then used to create the collection of lexical variants of the original term. Again, the identifiers were adorned with postfix. Figure 1 illustrates the process of creating Dict2. Please note that we generated lexical variants only for the 460 HPO terms annotated in the gold standard.

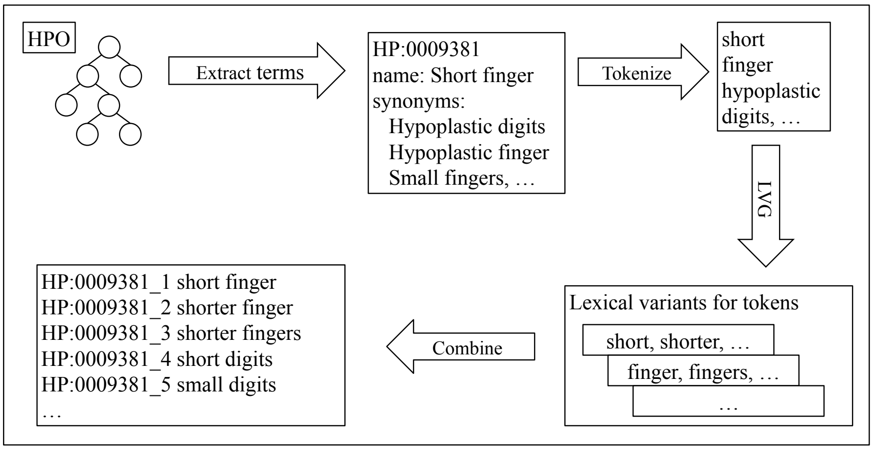

Figure 1: An example of lexical variants for the HPO term "HP:0009381 Short finger"

2.3 Annotation

To test the effectiveness of the created dictionary, we annotated the HPO gold standard corpus (Groza et al., 2015). For annotation with exact matching, we developed a simple annotator that compared lower case text against all terms in the developed dictionary. Common English stop word were ignored. Overlapping annotations with the same HPO identifier were considered as one annotation. The terms that were found in the text were annotated with text spans. For the similarity matching annotation we used the Jaccard coefficient, which is one of the commonly used metrics to measure similarity of two strings and is defined as the size of the intersection divided by the size of the union of the sample sets. For general performance of Jaccard coefficient in comparison with other approaches, readers are referred to (Cohen et al., 2003). Specifically we used PubDictionaries[1], a public service for text annotation using a dictionary (where the latter represent a collection of labels and their corresponding identifiers). A label is a natural language term that refers to the object identified by the corresponding identifier. PubDictionaries provides a REST service for text annotation using dictionaries which are plug-able, and it implements Jaccard coefficient for string similarity computation. The input to the REST services is the text, the type of the annotation, i.e., exact matching or similarity matching and the threshold coefficient in the case of the latter.

We have chosen to use PubDictionaries for our experiments not only because it eases our experiments with its pluggable dictionary system, but also by keeping the dictionaries in the public service, the experiments will remain replicable by any one. As PubDictionaries is an open source project, the experiments should be replicable. The following three similarity thresholds were used for annotation through PubDictionaries: 0.75, 0.85 and 0.95.

2.4 Evaluation

We defined true positive (TP), false positive (FP) and false negative (FN) concept annotations as follows. TPs were the annotations with the same HPO identifier found in both the dictionary and the gold standard corpus and an overlapping text span. For example, if in the following text: "A syndrome

[1] Available on: www.pubdictionaries.org

of brachydactyly (absence of some middle or distal phalanges)" the terms "brachydactyly" and "syndrome of brachydactyly" are both mapped to the same ID, they will represent the same annotation, since they overlap. On the other hand, when terms with the same identifier are found on different positions in the text, they represent different annotations. FP annotations were those detected with automatic annotator, but were not included in the gold standard corpus, while FN annotations represent annotations found in the gold standard and not detected with the automatic annotator.

Based on these three values, we evaluate the model with and report the Precision, Recall and F-Score values. Precision of positive class is the ratio of correctly annotated positive values to the number of all instances annotated as positive, this is also known as Positive Predictive Value. Recall of positive class is computed as the number of correctly annotated instances from the positive class divided by the number of all instances from the positive class; this is also known as sensitivity. F-Score is the weighted harmonic mean of Precision and Recall.

3 Results

Figure 2 summarizes results for five different approaches. One can notice that extending HPO terms with lexical variants in *Dict 2 E* reaches the highest F-Score (0.45), while it shares the highest Recall with the *Dict1 0.75* approach (0.45). The highest precision was achieved with *Dict1 0.95* (0.47).

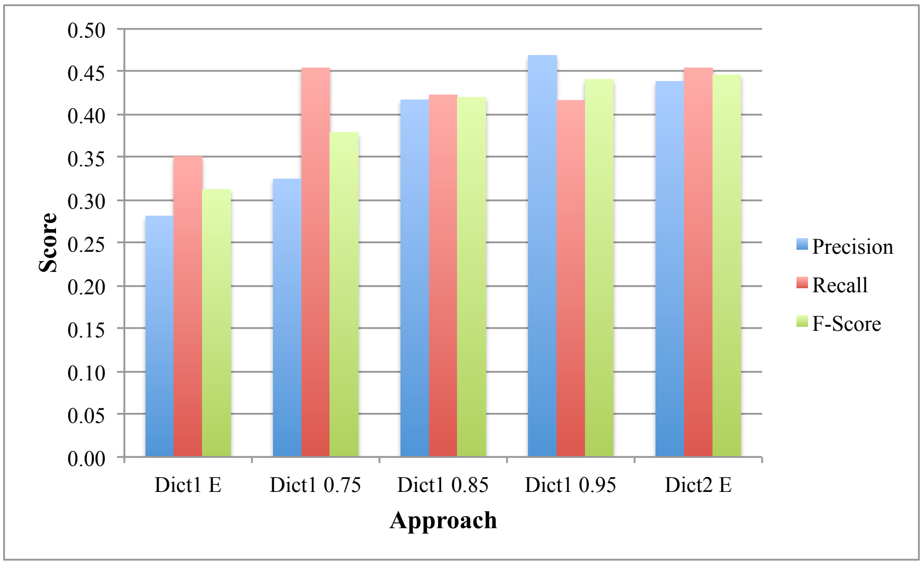

Figure 2: Precision, Recall and F-Score values for 5 different approaches: Dictionary 1 with Exact (E) matching and similarity matching (thresholds 0.75, 0.85 and 0.95), and Dictionary 2 with Exact matching.

Table 1 presents results of five different annotation combinations that were/were-not detected with different approaches. The term "syndrome of brachydactyly" is an example where a similarity metric with a low threshold value preforms better than our approach with added lexical variants. Terms "autosomal dominant trait" and "synostosis of some carpal and tarsal bones" are examples where extending the dictionary with lexical variants works well, while the term "malformed pinna" was not detected with any approach.

Table 1: Some examples of annotations that were (Yes) or were not (No) detected with different approaches.

Annotation	Dict1 E	Dict1 0.75	Dict1 0.85	Dict1 0.95	Dict2 E
syndrome of brachydactyly	No	Yes	No	No	No
brachydactyly	Yes	Yes	Yes	Yes	Yes
autosomal dominant trait	No	Yes	Yes	Yes	Yes
synostosis of some carpal and tarsal bones	No	No	No	No	Yes
malformed pinna	No	No	No	No	No

4 Conclusion

We presented and evaluated a dictionary of human phenotype terms and their lexical variants. Using a gold standard HPO corpus we measured Precision, Recall and F-Score, and compared five different approaches. The results showed that extending HPO terms with their lexical variants significantly improves the Recall and F-Score values compared to the original dictionary with no lexical variants. However, the method did not achieve the highest Precision of the system. Depending on the task and application, one might consider using our dictionary when Recall plays a more important role than Precision. Please note that we also used a relaxed method for defining true positives as described in Section 2.4. In case of strict exact matching, the results would be affected.

In the current version of the dictionary, we extended only a small subset of all HPO terms. In the future we plan to extend also other terms, however, with the current approach this would result in a large number of irrelevant and incorrect terms (such as, for example, "low blooded pressure" for the original term "low blood pressure"). Therefore, we are planning to address this issue before generating the full dictionary. In addition, we are planning to consider other reference corpus in the evaluation step. The current version of the dictionary is publicly available through PubDictionaries (HP_Garvan).

Reference

William W. Cohen, Stephen E. Fienberg, Pradeep D. Ravikumar, and Stephen E. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. *Proceedings of IJCAI-03 Workshop on Information Integration on the Web*:73–78.

Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533.

Christopher S. Funk, K. Bretonnel Cohen, Lawrence E. Hunter and Karin M. Verspoor. 2016. Gene Ontology synonym generation rules lead to increased performance in biomedical concept recognition. *Journal of Biomedical Semantics*, 7:52.

Georgios V. Gkoutos, Chris Mungall, Sandra Ďolken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Ǩohler, and Peter N. Robinson. 2009. Entity/quality-based logical definitions for the human skeletal phenome using PATO. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pages 7069–7072.

Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(496):496.

Tudor Groza, S. Kohler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M. Couto, Gareth Baynam, Andreas Zankl, Peter N. Robinson, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M. Couto, Gareth Baynam, Andreas Zankl, and Peter N. Robinson. 2015. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database*, 2015(0):bav005-bav005.

Simon Kocbek and Tudor Groza. 2016. Building a dictionary of lexical variants for human phenotype descriptors. In *BioNLP Workshop*, pages 186–190. ACL.

Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C M Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. Fitzpatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, et al. 2014. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1).

Paula M. Mabee, Michael Ashburner, Quentin Cronk, Georgios V. Gkoutos, Melissa Haendel, Erik Segerdell, Chris Mungall, and Monte Westerfield. 2007. Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology and Evolution*, 22(7):345–350.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William W Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, et al. 2008. Overview of BioCreative II gene normalization. *Genome biology*, 9 Suppl 2(SUPPL. 2):S3.

Chris Mungall, Georgios Gkoutos, Nicole Washington, and Suzanna Lewis. 2007. Representing phenotypes in OWL. In *CEUR Workshop Proceedings*, volume 258.

Cornelius Rosse and José L V Mejino. 2003. A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500.

NLM The Lexical Systems Group. 2016. Lexical Tools, 2016, https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2016/web/index.html, accessed June 2016

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2012. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–6.

Nicole L. Washington, Melissa A. Haendel, Christopher J. Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biology*, 7(11).