Building a dictionary of lexical variants for human phenotype descriptors

Simon Kocbek[1,2]
skocbek@gmail.com

Tudor Groza[1]
t.groza@garvan.org.au

[1]Kinghorn Center for Clinical Genomics, Garvan Institute of Medical Research,
Darlinghurst, Sydney, NSW 2011, Australia
[2]Department of Computing and Information Systems, University of Melbourne,
Melbourne, VIC 3000, Australia

Abstract

Detecting phenotype descriptors in text and linking them to ontology concepts is a challenging task. Current state-of-the art concept recognizers struggle with several issues due the variety of human expressiveness. Here we present initial results of creating a dictionary of lexical variants for the Human Phenotype Ontology. This work is a smaller but important part of a larger project with a goal to improve recall in phenotype concept recognizers.

1 Introduction

Phenotype descriptions (i.e., the composite of one's observable characteristics/traits) are important for our understanding of genetics. These descriptions 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). Scientific literature contains large amounts of phenotype descriptions, usually reported as free-text entries.

Concept Recognition (CR) is the identification of entities of interest in free text and their resolution to ontological concepts with the aim of leveraging structured 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*).

Consequently, phenotype CR is an ongoing research area with a demand for improvement. For example, systems such as OBO Annotator (Taboada et al., 2014), NCBO Annotator (Jonquet et al., 2009) and Bio-Lark (Groza et al., 2015) have been evaluated with maximum precision, recall and F-score values of 0.65, 0.49 and 0.56 respectively (Groza et al., 2015).

Here we present initial results of experiments designed to address the lexical variability of phenotype terms. We generate a dictionary of lexical variants for all HPO tokens. When completed, such a dictionary will help improve, in particular, the low recall of phenotype CR systems.

Generating lexical variants for HPO tokens is a fairly challenging task. For example, grouping similar words with classical similarity metrics such as the Levenshtein distance (even when using a high threshold) might group words with different meaning like *zygomatic* (a cheek bone) and *zygomaticus* (cheek muscle) into one lexical cluster. On the other hand, less similar words with same meaning like irregular nouns (e.g. *phalanx*, *phalanges,* or *femur*, *femora*) might be grouped into different clusters. Here, we experiment with the NLM Lexical Variant Generator (LVG) (The Lexical Systems Group, 2016) to generate lexical variants.

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 186–190,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

2 Methods

To generate the dictionary of lexical variants, we extracted all concept names and their synonyms from the HPO. The text was then tokenized and a cluster of lexical variants was created for each token. Tokens with overlapping lexical variants were merged into one cluster. We manually analyzed the clusters for their quality and coverage, and performed a preliminary automatic evaluation. In addition, we identified those parts of phenotype terms that display the largest lexical variability. For the latter, we used the following two additional ontologies: Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003), and the Phenotype and Trait Ontology (PATO) (Gkoutos et al., 2009). Details of data and methods used are described in the following sections.

2.1 The Human Phenotype Ontology

The HPO's primary goal is to offer a tool that allows large-scale computational analysis of the human phenotype (Köhler et al., 2014). The HPO is often used for the annotation of human phenotypes and has repeatedly been adopted in biomedical applications aiming to understand connection between phenotype and genomic variations. Some examples of using the HPO are 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.

Below is an example of part of a term in the OBO format:

```
id: HP:0000260
name: Wide anterior fontanel
def: "Enlargement of the anterior fontanelle with respect to age-
dependent norms." [HPO:curators]
synonym: "Large anterior fontanel" EXACT []
…
xref: UMLS:C1866134 "Wide anterior fontanel"
is_a: HP:0000236 ! Abnormality of the anterior fontanelle
property_value: HP:0040005 "Enlargement of the `anterior fonta-
nelle` (FMA:75439) with respect to age-dependent norms."
xsd:string {xref="HPO:curators"}
```

Terms in HPO usually follow the Entity-Quality formalism where they combine anatomical entities with qualities (Mungall et al., 2007) For instance, in the above example, *anterior fontanelle* describes an anatomical entity with the quality *wide*. Entities can usually be grounded in ontologies such as the FMA, while qualities usually belong to the PATO. It is assumed that rich lexical variability comes from the quality part of phenotype terms – due to their wide spread usage in common English.

For this study, we used the OBO versions of the HPO Apr 2016 and the PATO Nov 2015 ontologies, and the FMA OWL version 3.2.1.

2.2 Pre-processing text in ontologies

We extracted labels and synonyms for all HPO, PATO and FMA terms. The OWL API (Horridge and Bechhofer, 2011) was used for parsing.

After a manual inspection of a random subset of names and synonyms, 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 (< 3 characters). The final set contained 8,098 HPO; 1,959 PATO and 8,502 FMA tokens.

2.3 Generating clusters of lexical variants

We use the NLM Lexical Variant Generator (LVG), 2016 release (The Lexical Systems Group, 2016) to create lexical variants for the HPO tokens. LVG is a suite of utilities that can generate, transform, and filter lexical variants from the given input. Its intention is to create robust indexes and to transform user queries into retrievable entries from those indexes. Although LVG focuses on biomedical terms, it is not specialized for phenotype domain.

There are more than 60 functions (flow components) in LVG and each function has a set of parameters. In this work, the following two functions were used with the LVG Java API:

- Generating inflectional variants (IVs), which include the singular and plurals for nouns, the various tenses of verbs, the positive, superlative and the comparative of adjectives and adverbs.
- Generating derivational variants (DVs), which are terms that are related to the original term but do not necessarily, share the same meaning. Often, the derivational variant changes syntactic category from the original term. Only DVs with the same prefix as the original token (i.e., first two characters) were considered.

Both IVs and DVs can be generated with two methods: a) using an internal dictionary, and b) using a set of predefined rules. When generating lexical variants, we experimented with the following three configurations (Cs):

- C1: Generating IVs using the dictionary.
- C2: Generating IVs DVs using the dictionary.
- C3: Generating IVs and DVs using the dictionary first, and using the set of rules for those tokens that did not have any variants in the dictionary.

Generating lexical variants for HPO tokens can be described with the following algorithm:

```
Generate lexical variants for HPO:

N: number of HPO terms
$H_i$: single HPO term
S: set of names and synonyms for an HPO term
T: set of unique tokens
M: number of tokens
$T_j$: single token
$ID_j$: set of dictionary based IVs for $T_j$
$DD_j$: set of dictionary based DVs for $T_j$
$IR_j$: set of rule based IVs for $T_j$
$DR_j$: set of rule based DVs for $T_j$
V: sets of lexical variants
$V_k$: single set of lexical variants, where $0 < k < 4$
C: sets of clusters
$C_k$: single set of clusters, where $0 < k < 4$

For i = 1 to N do:
  Extract name/synonyms for $H_i$ and save them into S
  Tokenize S and save unique tokens into T
Initialize $C_1$, $C_2$ and $C_3$
For j = 1 to M do:
  Initialize $V_1$, $V_2$ and $V_3$
  Generate dictionary based inflectional variants for $T_j$ and save them into $ID_j$
  If $ID_j$ is empty then do:
      Generate rule based inflectional variants for $T_j$ and save them into $IR_j$
  Generate dictionary based derivational variants for $T_j$ and save them into $DD_j$
  If $DD_j$ is empty then do:
      Generate rule based derivational variants for $T_j$ and save them into $DR_j$
  $V_1 = ID_j$
  $V_2 = ID_j + DD_j$
  $V_3 = ID_j + DD_j + IR_j + DR_j$
  For k = 1 to 3 do:
      If a cluster in $C_k$ has a variant from $V_k$ then do:
          Put variants from $V_k$ into the existing cluster
      Else do:
          Create a cluster from $V_k$ in $C_k$
```

2.4 Inspecting/evaluating lexical clusters

For each configuration we calculated the coverage of extended tokens (i.e., the number of tokens for which at least one variant was found), and manually inspected lexical variants for 10 randomly selected tokens. In addition, we inspected clusters for the following two specific tokens of interest that are known to be problematic in phenotype CR: *phalanx* and *shortening*. The former is an irregular noun that changes to *phalanges* in plural form, while the latter represents a participle that is usually not correctly normalized for our need. For example, we would expect *short* and *shortening* in the same cluster (*short finger* vs. *shortening of the finger*). We also inspected variants for *zygomatic* and *zygomaticus* that should not be in the same cluster.

In addition to the manual inspection, we also performed a preliminary automatic evaluation of the clusters. The HPO has been integrated into Unified Medicine Language System (UMLS) Metathesaurus (Humphreys et al., 1998) since the 2015AB version (Dhombres et al., 2015). This potentially gives new synonyms for the HPO terms. The synonyms can contain lexical variants of the HPO term tokens. For example, *acute promyelocytic* ***leukemia***, does not contain any synonyms in the HPO. However, UMLS contains the synonyms *acute promyelocytic* ***leukaemia.*** Similarly, *ascending* ***aortic*** *aneurysm* has no HPO synonyms, while we can find *aneurysm of ascending* ***aorta*** in UMLS. Therefore, we developed an algorithm for counting those HPO terms that increased the coverage of tokens in UMLS synonyms for these terms (e.g., the above two terms would be counted).

As mentioned in section 2.1, it is assumed that most tokens with rich lexical variability are associated with the quality part of HPO terms. To test this assumption, we finally examined coverage of the HPO tokens in the FMA and the PATO. We then analyzed lexical cluster sizes for these tokens. In case the assumption is true, we expect the cluster sizes of PATO tokens (i.e. quality) larger than tokens found in FMA (i.e. entity).

3 Results and discussion

Table 1 summarizes the number of variants, the number of clusters, the average number of variants in each cluster and the number of tokens with no variants (NV) for different configurations.

Table 1: Results summary for each configuration

	#Variants	#Clusters	Average	#NV
C1	13,471	6,355	2.12	877
C2	18,080	5,620	3.22	877
C3	29,602	6,480	4.57	0

The same tokens with no variants were found using only the dictionary in C1 and C2, which implies that these tokens are not covered with LVG's dictionary. After the manual examination of generated clusters we can identify some examples of tokens without generated variants as follows: *spelling errors* (e.g., accesory, dermititis), *latin words* (e.g., ambiguus), *chemical compounds* (e.g., 23-diphosphoglycerate), *abbreviations* (eg., gnrh, pirc), *roman numbers* (e.g., xii, xiii), and *ordinal numbers* (e.g., 1st, 2nd). Using the rule-based approach in C3 generated variants for these tokens.

Examining the clusters showed that C1 generated several disjoint clusters that should be merged. Some examples are tokens like *abdomen* and *abdominal*, *abnormal* and *abnormality*, *external* and *externally*, and *yellow* and *yellowish*. As for the tokens of our particular interest, *phalanx* contained the following variants in the same cluster: *phalange*, *phalanges*, *phalanx*, and *phalanxes*; while *shortening* was clustered with the following variants: *shorten*, *shortened*, *shortening*, and *shortenings* and was missing words like *short*, *shorter* and *shortest*. Variants for *zygomatic* and *zygomaticus* were in separated clusters in all three approaches.

According to Table 1, the C2 approach generated more variants distributed into less clusters when compared to C1. Manual examination revealed that several disjoint clusters from previous paragraph merged into larger clusters (*abdomen* and *abdominal*, *abnormal* and *abnormality*, and *external* and *externally).* The *phalanx* cluster gained a new variant *phalangeal*, which was previously in a different cluster. There was no change in the *shortening* cluster.

Clusters in C3 extended tokens with no variants in LVG's dictionary with rule generated terms. However, variants for tokens like spelling errors or ordinal numbers were incorrect. For example, *accesory* would be extended with variants like *accesoryed* and *accesoryer*. In addition, participles were not in correct clusters (e.g., shortening). Unfortunately, terms like *brachymesomelia* or *trichromacy* were also extended with wrong variants. This implies that rules defined in LVG might not be appropriate for phenotype terms and we must define our own rules. This investigation is left for future work.

Testing with UMLS, we found that 6,580 (62%) of the HPO terms contained UMLS synonyms. 16% of these terms increased the coverage of synonym tokens with new lexical variants, which indicates that the generated dictionary does include quality variants. We plan to investigate the results in depth in the future.

When testing the coverage of HPO tokens in the PATO and the FMA, we found that 10% and 26% of the HPO tokens can also be found in the PATO and the FMA respectively. Figure 1 shows ratios for different lexical cluster sizes of the overlapping tokens created with the C2 approach (minimum/maximum size of 1 and 11 respectively). One can notice that the PATO tokens tend to form larger clusters, which indicates that these tokens have more lexical variants compared to the FMA tokens. This confirms the assumption from Section 2.1, that the quality part of phenotype term offers more lexical variability than the entity part.

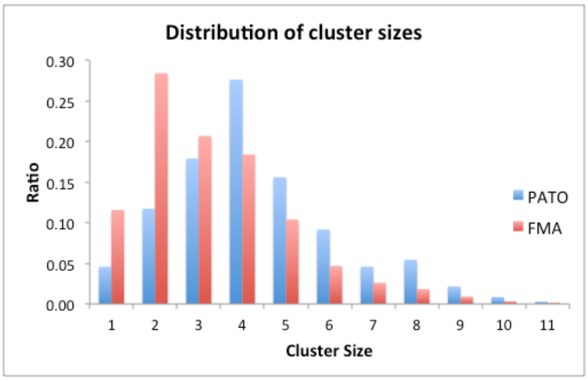

Figure 1: Distribution of lexical cluster sizes for those HPO tokens that were also found in the PATO/FMA.

4 Conclusion

In this paper we presented initial results for creating a dictionary of lexical variants of all tokens in the Human Phenotype Ontology. This task is a part of bigger project with aim to improve phenotype concept recognition. Using the NLM Lexical Variant Generator, we experimented with three configurations where different combinations of inflectional and derivational variants were used to extend original HPO token space. We examined the clusters and performed a preliminary automatic evaluation of these clusters. We also identified parts of phenotype terms that are likely to express more lexical variability.

In the future, we are planning to perform a detailed analysis of the generated clusters and improve the automatic evaluation. As seen in the results section, there are some phenotype tokens that are not covered in external dictionaries such as LVG. We will try to identify patterns of these tokens and see how we can extend them with lexical variants. In addition, we will improve the quality of generated clusters with removing incorrect variants (e.g., results of spelling errors), or tokens that are actually not phenotypes.

Focus of our future work will be the quality part of phenotype terms, since we showed that quality tokens display larger lexical variability than entity tokens. In addition, we have not managed to automatically generate clusters for all participles.

Reference

Ferdinand Dhombres, Rainer Winnenburg, James T. Case, and Olivier Bodenreider. 2015. Extending the coverage of phenotypes in SNOMED CT through post-coordination. In *Studies in Health Technology and Informatics*, volume 216, pages 795–799.

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.

Georgios V. Gkoutos, Chris Mungall, Sandra D̈olken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian K̈ohler, 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.

Matthew Horridge and Sean Bechhofer. 2011. The OWL API: A Java API for OWL ontologies. *Semantic Web*, 2(1):11–21.

Betsy L. Humphreys, Donald a. B. Lindberg, Harold M. Schoolman, and G. Octo Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11.

Clement Jonquet, Nigam H Shah, H Cherie, Mark a Musen, Chris Callendar, and Margaret-Anne Storey. 2009. NCBO Annotator : Semantic Annotation of Biomedical Data. *Iswc*:2–3.

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.

M. Taboada, H. Rodriguez, D. Martinez, M. Pardo, and M. J. Sobrido. 2014. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014(0):bau045–bau045.

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).