Semantic similarity-driven decision support in the skeletal dysplasia domain

Razan Paul[1], Tudor Groza[1], Andreas Zankl[2,3], and Jane Hunter[1]

[1] School of ITEE, The University of Queensland, Australia
razan.paul@uq.edu.au, tudor.groza@uq.edu.au, jane@itee.uq.edu.au
[2] Bone Dysplasia Research Group
UQ Centre for Clinical Research (UQCCR)
The University of Queensland, Australia
[3] Genetic Health Queensland,
Royal Brisbane and Women's Hospital, Herston, Australia
a.zankl@uq.edu.au

Abstract. Biomedical ontologies have become a mainstream topic in medical research. They represent important sources of evolved knowledge that may be automatically integrated in decision support methods. Grounding clinical and radiographic findings in concepts defined by a biomedical ontology, e.g., the Human Phenotype Ontology, enables us to compute semantic similarity between them. In this paper, we focus on using such similarity measures to predict disorders on undiagnosed patient cases in the bone dysplasia domain. Different methods for computing the semantic similarity have been implemented. All methods have been evaluated based on their support in achieving a higher prediction accuracy. The outcome of this research enables us to understand the feasibility of developing decision support methods based on ontology-driven semantic similarity in the skeletal dysplasia domain.

1 Introduction

Similarity plays a central role in medical knowledge management. Like most scientific knowledge, medical knowledge is also inferred from comparing different concepts (such as phenotypes, populations, and species) and analyzing their similarities and differences. However, medical science is unlike other sciences in that its knowledge can seldom be reduced to a mathematical form. Thus, medical scientists usually record their knowledge in free form text, or lately in biomedical ontologies. New concepts that emerge in the domain are firstly compared and judged based on their degree of similarity to existing concepts before being integrated into the overall domain knowledge.

Biomedical ontologies are knowledge bases that have emerged and evolved over time following this process. Most of them are used not only to model and capture specific domain knowledge, but also to annotate, and hence enrich, diverse resources like patient cases or scientific publications. The adoption of biomedical ontologies for annotation purposes provides a means for comparing

medical concepts on aspects that would otherwise be incomparable. For example, the annotation of a set of disorders (directly or via patient cases) using the same ontology enables us to compare them, by looking at the underpinning annotation concepts. The actual comparison is subject to a semantic similarity measure, i.e., a function that takes two or more ontology concepts and returns a numerical value that reflects the degree of similarity between these concepts.

Over the course of the last decade, there has been significant research performed on semantic similarities over biomedical ontologies. One key remark that needs to be taken into account is that meaningful similarity measures are dependent on the domain knowledge, as only by using the explicit semantics of the domain one can compare concepts in an appropriate manner. In this paper we report on our experiences with using semantic similarity over domain knowledge and annotated patient cases for disorder prediction in the skeletal dysplasia domain.

Skeletal dysplasias are a group of heterogeneous genetic disorders affecting skeletal development. There are currently over 450 recognised bone dysplasias, structured into 40 groups. Patients suffering from such disorders have complex medical issues, ranging from bowed arms and legs to neurological complications. Since most dysplasias are very rare ($<$ 1:10,000 births), data on clinical presentation, natural history and best management practices is very sparse. A different perspective on data sparseness is introduced also by the small number of clinical and radiographic phenotypes typically exhibited by patients from the vast range of possible characteristics globally associated with these disorders.

Decision support methods can usually assist clinicians and researchers both in the research, as well as in the decision making process, in general, in any domain. However, building efficient or meaningful decision support methods in a domain affected by data sparseness, such as bone dysplasias, is a very challenging task. On the other hand, semantic similarity measures can facilitate the objective interpretation of clinical and radiographic findings by using knowledge captured in biomedical ontologies or annotated patient cases to provide decision support. In this paper we aim to bridge the two worlds, by investigating different approaches for determining the semantic similarity between sets of phenotypes encoded as ontological concepts and its application to disorder prediction.

The context of our work is provided by the SKELETOME project that develops a community-driven knowledge curation platform for the bone dysplasia domain [1]. The underlying foundation of the platform is a two-phase knowledge engineering cycle which enables: (1) semantic annotation of patient cases – connecting domain knowledge to real-world cases; and (2) collaborative diagnosis, collaborative knowledge curation and evolution – evolving the domain knowledge, based on real-world cases. The semantic annotation process relies on clinical and radiographic findings grounded in the Human Phenotype Ontology (HPO) [2] – an emerging de facto standard for capturing, representing and annotating phenotypic features encountered in rare disorders. At the same time, the domain knowledge is modeled via the Bone Dysplasia Ontology (BDO) [3], which

at a conceptual level associates bone dysplasias and phenotypes represented by HPO terms.

These two sources of knowledge, i.e., domain knowledge from BDO and raw knowledge from annotated patient cases, together with the structure of HPO, which underpins the formalization of phenotypes, enable us to investigate the use of several semantic similarity measures in order to achieve disorder prediction. More concretely, this paper: (i) analyzes which semantic similarity performs better on each of the two types of data, and (ii) performs an extensive empirical evaluation of the application of these semantic similarities for disorder prediction, using a real-world dataset.

The remainder of the paper is structured as follows: Section 5 discusses existing related work, Section 2 provides a comprehensive background on the knowledge and data sources used within our experiments, while Section 3 details our methodology. Before concluding in Section 6 we present an extensive evaluation and discuss the experimental results in Section 4.

2 Background

This section provides a brief overview of the background of our work. It introduces the Human Phenotype Ontology (HPO) and discusses some of its characteristics (Section 2.1), then describes the two knowledge sources used in our experiments, i.e., the Bone Dysplasia Ontology and the largest bone dysplasia patient dataset (Section 2.2) and finally, presents briefly some of the most commonly used similarity measures (Section 2.3).

2.1 Human Phenotype Ontology

The Human Phenotype Ontology [4] is a controlled vocabulary that captures and represents clinical and radiographic findings (or phenotypes in general), in principle, in hereditary diseases listed in Online Mendelian Inheritance in Man (OMIM) database [5]. The ontology consists of around 9,000 concepts describing modes of inheritance, onset and clinical disease courses and phenotypic abnormalities. This last category represents around 95% of the ontology and is the main subject of our study. Phenotypic abnormalities are structured in a hierarchical manner (via class–subclass relationships) from generic (e.g., `HP_0000929` – *Abnormality of the skull*) to specific abnormalities (e.g., `HP_0000256` – *Macrocephaly*).

One aspect that needs to considered when using the structure of HPO is the multiple inheritance. All children of a particular class share some information (which is logical in a typical ontology), however, the type of this shared information (i.e., not the specific information) can be different. More concretely, abnormalities may share their anatomical localization or they may share the

[4] `http://www.human-phenotype-ontology.org/`
[5] `http://www.omim.org/`

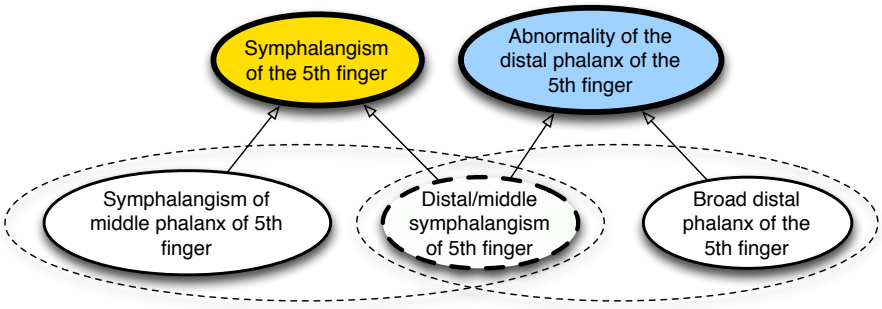

Fig. 1. An example of multiple inheritance in HPO (arrows denote class–subclass relations).

intrinsic type of abnormality. Fig. 1 depicts an example of such multiple inheritance. `HP_0009244` (*Distal/middle symphalangism of 5th finger*) is a sibling of `HP_0009178` (*Symphalangism of middle phalanx of 5th finger*) – they represent the same type of abnormality, i.e., *Symphalangism*, and hence are both children of `HP_0004218` (*Symphalangism of the 5th finger*), but also a sibling of `HP_0009240` (*Broad distal phalanx of the 5th finger*) – they share the anatomical localization of the abnormality, and hence are both children of `HP_0004225` (*Abnormality of the distal phalanx of the 5th finger*). This remark is important because it influences the computation of the most specific common ancestor for two concepts, a central element of most semantic similarity measures.

2.2 Bone dysplasia knowledge sources

As mentioned in Section 1, in the context of the SKELETOME project, we have two major knowledge sources: the Bone Dysplasia Ontology (BDO) [6] and a set of semantically annotated patient cases. The clinical and radiographic findings that characterize both are underpinned by the Human Phenotype Ontology.

BDO has been developed to model and capture essential (and mature) knowledge in the skeletal dysplasia domain. As depicted in Fig. 2, it associates bone dysplasias to gene mutations and phenotypic characteristics, which are then further specialised via concepts defined by external ontologies, such as HPO. In [3] we provide a comprehensive overview of the design process of BDO. With respect to the work described in this paper, there is one remark that is worth being noted. BDO describes associations (via class axioms) between more than 250 disorders (out of the 450 in total) and around 2,000 findings (represented by HPO concepts). These associations have been created from the clinical synopses of the corresponding disorders in OMIM and represent, in principle, the current state of *conceptual* understanding of their clinical manifestations. As a result, the phenotypic findings listed there are a mixture of more generic (e.g., *abnormal*

[6] `http://purl.org/skeletome/bonedysplasia`

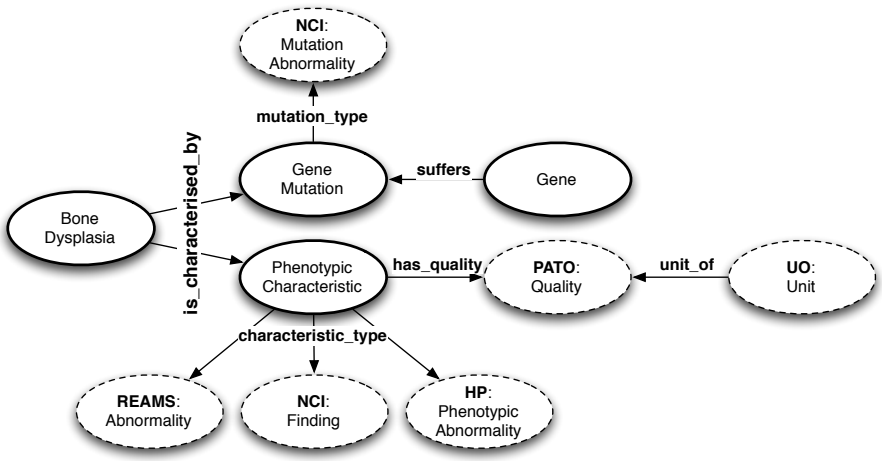

Fig. 2. A snapshot of the Bone Dysplasia Ontology from [1].

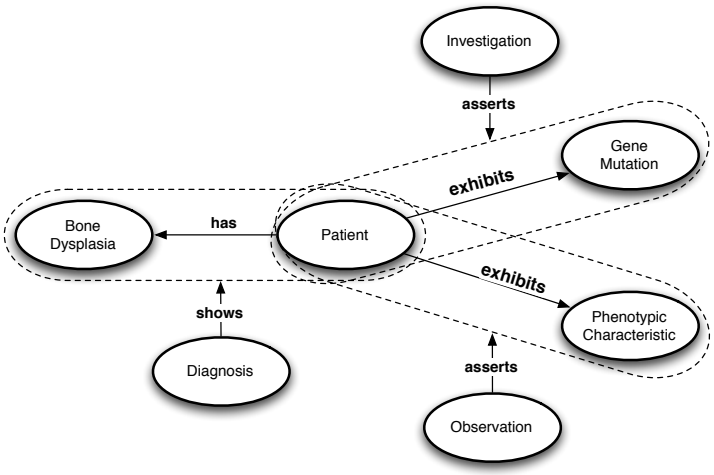

Fig. 3. A snapshot of the Patient Ontology from [1].

femoral neck) and fairly specific (e.g., *short, broad femoral neck*) terms. This reflects the balance achieved by capturing both the clinical interpretation of sets of patient cases (for the more common disorders), as well as singular or particular patient cases (for those that are extremely rare).

In addition to the domain knowledge, SKELETOME focuses also on capturing instance data, i.e., annotated patient cases. The actual modeling is done via the Patient Ontology (depicted in Fig. 3), which associates patients to clinical and radiographic findings, gene mutations and bone dysplasias. The main

source of patient data is the registry of the European Skeletal Dysplasia Network (ESDN) [7], which is a pan-European research and diagnostic network aimed to provide community driven help and diagnostic expertise for rare bone disorders. Our current dataset comprises a total of 1,200 semantically annotated closed ESDN cases. Each patient case has been modeled using the Patient Ontology and captures HPO concepts denoting clinical and radiographic findings and BDO dysplasias denoting the final diagnosis. In contrast to the knowledge in BDO, the level of specificity present in the clinical descriptions is, as expected, fairly high, i.e., the general tendency is to find more specific findings rather than more generic ones.

2.3 Semantic similarity

As mentioned earlier, there has been a great amount of research done on semantic similarities. Here, we intend only to introduce some basic concepts and to provide a brief overview of the measures used within our experiments. A detailed survey on semantic similarity on biomedical ontologies can be found in [4].

There are two main types of semantic similarities: (1) node-based similarities and (2) edge-based similarities. The former uses nodes and their properties as information source, whereas the latter focuses on edges and their types.

Node based approaches rely on the notion of Information Content (IC) to quantify the informativeness of a concept. IC values are usually calculated by associating probabilities to each concept in ontology by computing the negative likelihood of its frequency in large text corpora. The basic intuition behind the use of the negative likelihood in the IC calculation is that the more probable the presence of a concept in a corpus is, the less information it conveys. IC is expressed in Eq. 1, with $p(c)$ being the probability of occurrence of c in a specific corpus. In our case $p(c)$ represents the probability of occurrence of an HPO concept in the context of a bone dysplasia, either from the domain knowledge, or from the raw patient cases.

The foundational node based similarity measures are Resnik [5], Lin [6] and Jiang and Conrath [7]. Resnik was the first to leverage IC for computing semantic similarity and expressed semantic similarity between two terms as the IC of their most informative common ancestor (MICA – Eq. 2). The intuition is that similarity depends on the amount of information two concepts, c_1 and c_2, share. This, however, does not consider how distant the terms are in their information content and from a hierarchical perspective. Consequently, Lin (Eq. 3) and Jiang and Conrath (Eq. 4) have proposed variations of Resnik's similarity to take into account these aspects.

$$IC(c) = -logp(c) \tag{1}$$

$$sim_{Res}(c_1, c_2) = IC(c_{MICA}) \tag{2}$$

[7] `http://www.esdn.org`

$$sim_{Lin}(c_1, c_2) = \frac{2 * IC(c_{MICA})}{IC(c_1) + IC(c_2)} \tag{3}$$

$$sim_{JC}(c_1, c_2) = 1 - IC(c_1) + IC(c_2) - 2 * IC(c_{MICA}) \tag{4}$$

Edge-based approaches take into account the paths existing between the concepts in the ontology. Subject to the domain and ontology, such paths could be considered by following *is-a* relationships (the most common approach) or other types of relationships defined by the ontology. Examples of such similarity measures include: (i) Wu & Palmer [8] (Eq. 5), where LCS is the least common subsumer of c_1 and c_2 and N_1 is the length of the path from c_1 to root, N_2 the length of the path from c_2 to root and N_3 the length of the path from LCS to root; or (ii) Leacock-Chodorow [9] (Eq. 6), where D is the overall depth of the ontology. A more recent measure has been described in [10] and considers, among other aspects, the number of changes in direction of the shortest path between two concepts (i.e., how many times on the shortest path the traversing direction changes from child to parent and vice-versa).

$$sim_{W\&P}(c_1, c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \tag{5}$$

$$sim_{L\&C}(c_1, c_2) = -log \frac{len(c_1, c_2)}{2 * D} \tag{6}$$

A third category of similarity measures could be considered for the hybrid approaches, i.e., combining node and edge-based similarities (e.g., [11] or [12]). Our work aims to integrate both information content and structural relationships in order to gain as much as possible from the semantics provided by HPO. As described in the following section, we also propose a series of such hybrid measures tailored on specific requirements emerged from our knowledge sources.

3 Methodology

The goal of our work is to predict disorders given an annotated patient case description. More concretely, given a background knowledge base (i.e., BDO or the annotated patient dataset) and a set of HPO concepts (representing clinical and radiographic findings of a new patient case), we aim to predict the most plausible bone dysplasias, ranked according to their probability. This is a typical multi-class classification problem, however, due to data sparseness that characterises the skeletal dysplasia domain, typical Machine Learning algorithms achieved a very low accuracy [8]. Our intuition is that by using semantic similarity measures on patient findings (i.e., HPO concepts) we are able to leverage and use intrinsic associations between phenotypes that cannot, otherwise, be acquired by typical

[8] A series of classification experiments we have performed revealed a maximal accuracy of around 35% for Naive Bayes, in a setting in which we have considered only six disorders, i.e., those that had more than 20 cases.

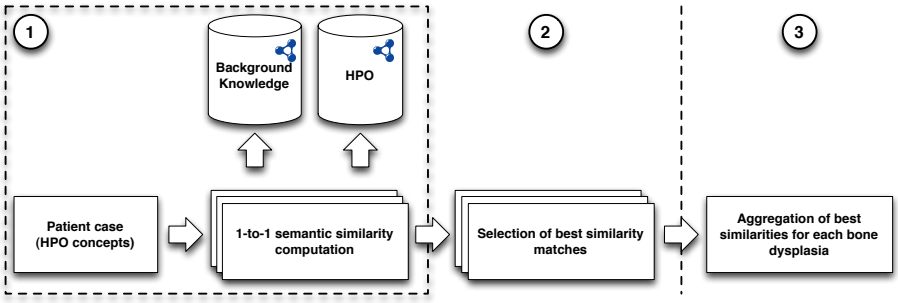

Fig. 4. Block diagram of the prediction methodology.

Machine Learning methods (due to their term-based matching process). As an example, if the background knowledge base lists `HP_0000256` (*Macrocephaly*) as a phenotype of `Achondroplasia` and a new patient exhibits `HP_0004439` (*Craniofacial dysostosis*) we want to use the semantic similarity between `HP_0000256` and `HP_0004439` to also associate the later to `Achondroplasia` with a certain probability [9]. The semantic similarity between the two concepts could be inferred via their most common ancestor `HP_0000929` (*Abnormality of the skull*). Such an association is not possible when employing typical Machine Learning methods since each term would be considered individually and only in the context provided by the background knowledge base.

Fig. 4 depicts the overall methodology. In the first step, we compute the semantic similarity between all HPO concepts representing clinical and radiographic findings of the given patient case and all phenotypes associated with bone disorders in the background knowledge base (please note that we do not make any assumptions about the background knowledge base). If we consider $\{S_1, S_2, \ldots, S_n\}$ to be patient findings and $\{P_1, P_2, \ldots, P_n\}$ phenotypes of bone dysplasia D, the best similarity match between S_i and D is given by:

$$BestMatch(S_i, D) = \arg\max_{j=1}^{n}\{sim(S_i, P_j)\} \tag{7}$$

The semantic similarity in Eq. 7 can be any of the classical similarities mentioned in Section 2 or, for example, one of the measure we introduce later in this section. The evaluation described in Section 4 has been performed on multiple such similarity measures. Once the best matches have been computed, we calculate the final probability by aggregating them:

$$P(S_1, S_2, \ldots, S_n|D) = \frac{1}{n} * \sum_{i=1}^{n} BestMatch(S_i, D) \tag{8}$$

[9] As a remark, there is no direct relationship in HPO between the concepts `HP_0000256` and `HP_0004439`. A relationship exists only via the parent of `HP_0000256` (i.e., `HP_0000240` – *Abnormality of skull size*), which is a sibling of `HP_0004439`

As mentioned previously, a good semantic similarity measure needs to take into account the specific aspects of the target domain. Below we have summarized a series of requirements for the similarity measure that have emerged from the bone dysplasia domain and the structure of HPO:

- Given two HPO concepts and their LCA (lowest common ancestor), we consider the concept closer to the LCA to be more similar to the LCA than the concept located at a bigger distance. E.g., `HP_0004439` (*Craniofacial dysostosis*) will be considered more similar to `HP_0000929` (*Abnormality of the skull*) than `HP_0000256` (*Macrocephaly*), because it is a direct descendent of `HP_0000929`;
- The information content of an LCA is dependent on its specificity (i.e., its location in the overall hierarchy). More concretely we consider the more specific LCA to be more informative. E.g., `HP_0004439` (*Craniofacial dysostosis*) (as an LCA) should be considered more informative than `HP_0000929` (*Abnormality of the skull*), which is in this case, its direct parent.
- A smoothing parameter may be required to deal with missing LCA information content. As described in [12], one of the main issues of IC is that its values are derived by analyzing large corpora (in our case a given background knowledge base), which may not even contain certain concepts. This is also the case with LCAs computed on certain pairs of findings, aspect dependent on the background knowledge base. Unfortunately, neither the intrinsic information content defined in [13], nor the extended information content defined in [12] can be employed in our domain, because we need the IC of a concept to be defined in the context of a given disorder (see below) and not only based on its children or surrounding concepts in the ontology. In other terms, we cannot use only the local IC definition provided by HPO without the scope provided by and associated disorder.

In addition to these requirements, we need to define the Information Content (IC) of a finding in the context of a disorder. Independently of the background knowledge base used for experiments, we have considered $IC(C_P)$ (i.e., the IC of the concept C grounding phenotype P) to be:

$$IC(P) = -log\frac{N_{DP}}{N_D} \tag{9}$$

where N_{DP} represents the number of disorders associated with P and N_D is the total number of disorders.

In the following we define a series of hybrid semantic similarities that take into account the above listed requirements.

HSS1 quantifies the semantic similarity between concepts according to the information content of the LCA and the position of LCA in regards to the concepts. HSS1 neglects the specificity of LCA.

$$HSS1(C_1, C_2) = \frac{Any - Node - Based - Similarity}{DIST(C_1, LCA) + DIST(C_2, LCA)} \tag{10}$$

where $DIST(C, C) = 0$ and $DIST(C_1, C_2) = len(SPath(C_1, C_2))$ (SPath = shortest path).

For example, HSS1 used with Resnik (sim_{Res}) would be:

$$HSS1(C_1, C_2) = \frac{IC(LCA)}{DIST(C_1, LCA) + DIST(C_2, LCA)} \tag{11}$$

HSS2 introduces the specificity of LCA, however, it neglects the missing LCA information content. HSS2 is defined below.

$$HSS2(C_1, C_2) = \frac{L}{D} * HSS1 \tag{12}$$

where L is the length of the path from the root to LCA and D is the depth of the ontology.

HSS3 and HSS4. In order to fulfil the last requirement, we have experimented with two additional measures (HSS3 and HSS4 defined below), that introduce different smoothing parameters: HSS3 uses a constant K, where $K = 1/N_D$ (N_D = total number of disorders), while HSS4 considers a joint information content of the two concepts.

$$HSS3(C_1, C_2) = \frac{\frac{L}{D} * (IC(LCA) + K)}{DIST(C_1, LCA) + DIST(C_2, LCA)} \tag{13}$$

$$HSS4(C_1, C_2) = \frac{\frac{L}{D} * (IC(LCA) + \frac{IC(C_1) * IC(C_2)}{IC(C_1) + IC(C_2)})}{DIST(C_1, LCA) + DIST(C_2, LCA)} \tag{14}$$

4 Experimental results

Taking into account the context provided by the SKELETOME project, i.e., a platform used by clinicians, we have tested the disorder prediction on a subset of the patient dataset described in Section 2. We performed three different experiments, described in the following:

- Firstly, we used a part of the patient dataset as knowledge source,
- Secondly, we used the Bone Dysplasia Ontology as knowledge source,
- Thirdly, we compared the semantic similarity-based prediction against a term matching-based prediction (i.e., an approach that uses only the frequency of the patient findings in the context of each disorder).

Each experiment tested different semantic similarity measures (applied in a HPO concept to concept setting). To assess the efficiency provided by the semantic similarity, we have calculated the overall accuracy of the disorder prediction. Node-based similarities have used the information content calculated on the background knowledge used in the experiment (i.e., IC on BDO or on patient cases), while the hybrid similarities have used both this information content and the structure of HPO.

Table 1. Experimental results of disorder prediction using patient cases as background knowledge.

Similarity	A@1 (%)	A@2 (%)	A@3 (%)	A@4 (%)	A@5 (%)
Resnik	10.96	21.92	32.88	35.62	41.10
Lin	6.85	12.33	17.81	28.77	34.25
J&C	2.74	9.59	12.33	13.70	20.55
HSS1	31.51	46.58	54.79	64.38	71.23
HSS2	32.87	49.32	56.16	64.38	69.86
HSS3	**39.73**	**52.05**	**61.64**	**69.86**	**75.34**
HSS4	39.73	52.05	60.27	68.49	73.97

In Section 2 we have discussed some of the foundational differences between the two knowledge sources with respect to the phenotypes' specificity. Another aspect that needs to be mentioned is that, since the raw knowledge we are using emerges from real patient cases, it will contain clinical and radiographic features that are directly related to the disorder, but also phenotypes that are not necessarily relevant. This is a normal phenomenon, because clinicians record all their findings before considering a diagnosis. For example, a clinical summary may contain findings such as, *bowed legs*, *macrocephaly* and *cleft palate*, which are relevant for the final *Achondroplasia* diagnosis, but it may also contain *fractured femur* and *decreased calcium level*, which are not relevant in the context of the final diagnosis. The set of unrelated findings are termed as noise.

Noise is the one of the most important contributing factors to the prediction accuracy, and it is inverse proportional to it. Hence, the prediction accuracy depends on the noise introduced both by the background knowledge, as well as the test data. As we are considering both the domain knowledge (via BDO) and patient cases as background knowledge bases in two different assessments, we will be able to judge which of the two types of knowledge contains more noise. This is realized by testing both on the set test dataset and comparing the resulted accuracy.

In all experiments detailed below we compute the prediction accuracy as the overall percentage of correctly predicted disorders at a given recall cut-off point (i.e., by taking into account only the top K predictions, for different values of K, where K is the recall cut-off point). Hence, a success represents correctly predicted disorder (the exact same, and not a sub or super class of it), while a miss represents an incorrectly predicted disorder. If N is the total number of test cases and L is the number of corrected predicted disorders, then Accuracy $A = L/N$. This is expressed in percentages in Tables 1, 2 and 3.

4.1 Experiment 1: Patient data as knowledge base

This first experiment considers patient cases as background knowledge. As discussed in Section 2, we collected a dataset of 1,200 patient cases from ESDN and annotated them with HPO terms. In order to provide an accurate view over the

Table 2. Experimental results of disorder prediction using BDO as background knowledge.

Similarity	A@1 (%)	A@2 (%)	A@3 (%)	A@4 (%)	A@5 (%)
Resnik	2.74	4.10	4.10	6.84	8.21
Lin	1.37	2.74	2.74	4.10	4.10
J&C	0	0	0	0	0
HSS1	16.43	21.91	32.87	43.84	**47.95**
HSS2	10.96	16.43	17.80	24.66	27.40
HSS3	10.96	16.44	19.18	23.29	27.40
HSS4	10.96	17.80	19.18	21.92	28.77

prediction, the experiment has been performed as a 5-fold cross validation with an 80-20 split (80% knowledge base, 20% test data). Table 1 lists the resulted average accuracy at five different recall cut-off points.

Overall, HSS3 has performed the best in this experiment, more or less on par with HSS4, and has confirmed that it is important for all three requirements listed in Section 3 to be fulfilled. Moreover, this experiment shows the improvement brought by a hybrid method over traditional information content based approaches. HSS1 outperforms the IC-based similarities because it considers the distance to the LCA and not only the IC of the LCA – i.e., the closer the two terms are to the LCA (and implicitly between them) the more similar they are. At the same time, HSS3 outperformed HSS1 because it smooths the missing information content, while at the same time introducing the specificity (L/D) – which is characteristic to the background knowledge. Finally, the similarity between HSS3 and HSS4 (that can also be observed in experiment 2) shows that the parameter $K = 1/N_D$ is a good approximation of the joint information content of the two concepts.

4.2 Experiment 2: BDO as knowledge base

The second experiment evaluated the disorder prediction with BDO as background knowledge. We have performed the same rounds of experiments as in the first case, i.e., we tested the prediction accuracy for the exact same 5 test folds resulted from experiment 1 and computed the final average accuracy for each semantic similarity. Results are listed in Table 2.

As in the case of the first experiment, all hybrid similarities outperformed the classical information content approaches. This time, however, HSS1 has achieved the best result, proving that the HPO concepts captured by BDO are more generic, as we have expected. The specificity factor L/D in HSS2, HSS3 and HSS4 takes low values because L is generally smaller (i.e., terms are located higher in the hierarchy and hence more generic) which leads to smaller values for these measures. This is also the reason why, the same similarities have performed worse when BDO was considered background knowledge, as opposed to using patient cases as background knowledge. The specificity of the ancestor improves

Table 3. Experimental results on term matching vs. semantic similarity.

Method	A@1 (%)	A@2 (%)	A@3 (%)	A@4 (%)	A@5 (%)
Patient cases as background knowledge					
Term matching	26.02	38.36	50.68	56.16	61.64
Semantic similarity	39.73	52.05	61.64	69.86	**75.34**
BDO as background knowledge					
Term matching	8.21	15.06	21.91	26.02	27.4
Semantic similarity	16.43	21.91	32.87	43.84	**47.95**

the accuracy on patient cases but it decreases it on domain knowledge. Finally, a different reason for the lower accuracy is the multiple inheritance used in HPO, which leads to additional missing information content for LCAs.

4.3 Experiment 3: Term matching vs. semantic similarity

Finally, in order to gain insight in the importance of using semantic similarity measures in disorder prediction, we have compared the results of the best performing similarity for each background knowledge against prediction calculated on term-based matching. Firstly, the results listed in Table 3 show that using semantic similarity is generally a good strategy as the overall accuracy is improved when compared to term-based matching, independently of the background knowledge. Interestingly, in this comparison, the specificity factor that heavily influences the accuracy based on the background knowledge has proved to be beneficial in the context of BDO, when compared against term matching. Secondly, returning to the comparison based on background knowledge, we can conclude that the domain knowledge introduces more noise than patient cases, which seems to contradict our initial belief (since clinicians will list in a case all observed findings, including those that may turn out to be irrelevant for the final diagnosis). In reality, in this case we are dealing with a different kind of noise, as the domain knowledge has the tendency to dilute the discriminatory findings when aggregating the information resulted from analyzing groups of patients. We intend to deal with this issue by including knowledge on differential diagnosis in the Bone Dysplasia Ontology.

5 Related Work

The research presented in [14] is the most relevant related work in the context of this paper. Kohler et al. have developed a semantic similarity search application named Phenomizer, which takes as input a set of HPO terms and returns a ranked list of diseases from OMIM, to their semantic similarity values. Phenomizer uses the Resnik semantic similarity and arithmetic mean as aggregation strategy (similar to our approach). According to the experiments discussed in

the paper, their solution outperforms term-based matching approaches that do no consider any relationships between terms. Our research follows closely the work done in Phenomizer, however, we use real patient data to test the disorder prediction (as opposed to the synthetically generated data in their case), and we try to tailor the semantic similarity to map onto the requirements emerging from the domain. Furthermore, we test several semantic similarities in order to get a better understanding of the most appropriate combination that serves our prediction goal. Finally, we evaluate the prediction accuracy using two types of background knowledge – domain and raw knowledge, as opposed to only domain knowledge in their case.

Additional related work includes [15], where the authors use a threshold of lowest semantic similarity value to find best-matching term pairs with the goal of predicting molecular functions of genes in Gene Ontology (GO) [16] annotations. Similar to our work, the authors tailor the semantic similarity measures according to fit the structure of GO and their application requirements. Lei et al. [17] assess protein similarity within GO to predict the subnuclear location. They compared the prediction accuracy of several similarity measures, including classical ones such as Resnik, and term-based matching to find insignificant differences between them. The authors also evaluate several aggregation strategies for the similarity values (e.g., sum, average, multiplication) and have found that the sum of the term-based matching method produces the best predictive outcome. Subnuclear location of a gene is associated with specific GO terms in most of the cases. As a result, using the hierarchical structure of the ontology via semantic similarity methods may not bring significant improvements.

In [18], the authors use ontological annotations and a proposed semantic similarity measure to find a correlation between protein sequence similarity and semantic similarity across GO. Similarly, Washington et al. [19] investigate the ontological annotation of disease phenotypes and the application of semantic similarities to discover new genotype-phenotype relationships within and across species. Finally, Ferreira et al. [20] use semantic similarity measures to classify chemical compounds and have showed that employing such techniques improves the chemical compound classification mechanisms. To achieve this, they employed measures tailored on the semantics of the Chemical Entities of Biological Interest Ontology (ChEBI).

6 Conclusion

In this paper we have reported on our experiences in using semantic similarity measures for disorder prediction in the skeletal dysplasia domain. The SKELETOME project provides two types of knowledge sources: (1) domain knowledge, modeled by and captured in the Bone Dysplasia Ontology and (2) raw knowledge emerging from patient cases. In both cases the clinical and radiographic findings are grounded in Human Phenotype Ontology concepts. The data sparseness that characterises this domain required us to consider alternative approaches in performing disorder prediction. Hence, we took advantage of the semantics pro-

vided by HPO and experimented with different semantic similarity measures, using both types of knowledge sources.

The experimental results have led to the conclusion that applying only information theoretic approaches in computing semantic similarity over the Human Phenotype Ontology, in our domain, does not provide the optimum result. Instead, we need to take into account particular requirements that emerge from the data characteristic to the bone dysplasia domain, i.e., a combined path between findings and their common ancestor, the specificity of this common ancestor and a smoothing parameter for the cases when the information content of the common ancestor is missing. Another conclusion of our experiments has been the need for differential diagnosis information in the domain knowledge in order to increase the weight of the discriminatory findings. Finally, we have shown that using semantic similarities improved the prediction accuracy when compared to term-based (frequency) matching prediction.

Acknowledgments

The work presented in this paper is supported by the Australian Research Council (ARC) under the Discovery Early Career Researcher Award (DECRA) – DE120100508 and the Linkage grant SKELETOME – LP100100156.

References

1. Groza, T., Zankl, A., Li, Y.F., Hunter, J.: Using Semantic Web Technologies to Build a Community-driven Knowledge Curation Platform for the Skeletal Dysplasia Domain. In: Proc. of ISWC 2011, Bonn, Germany (2011) 81–96
2. Robinson, P.N., Kohler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S.: The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. The American Journal of Human Genetics **83**(5) (2008) 610–615
3. Groza, T., Hunter, J., Zankl, A.: The Bone Dysplasia Ontology: integrating genotype and phenotype information in the skeletal dysplasia domain. BMC Bioinformatics **13**(50) (2012)
4. Pesquita, C., Faria, D., Falcao, A., Lord, P., Couto, F.: Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology **5**(7) (2009)
5. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the 14th IJCAI. (1995) 448–453
6. Lin, D.: An information-theoretic definition of similarity. In: Proc. of the 15th ICML. (1998) 296–304
7. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the 10th Conf. on Research on Comp. Linguistics, Taiwan (1997)
8. Wu, Z., Palmer, M.: Verb semantics and lexicon selection. In: Proc. of the 32nd ACL. (1994) 133–138
9. Chodorow, M., Leacock, C.: Combining local context and WordNet similarity for word sense identification. Fellbaum (1997) 265–283
10. Schickel-Zuber, V., Faltings, B.: OSS: A Semantic Similarity Function based on Hierarchical Ontologies. In: Proc. of the 20th IJCAI. (2007) 551–556

11. Li, Y., Bandar, Z., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. ITEE Transactions on Knowledge and Data Engineering **15**(4) (2003) 871–882
12. Pirro, G., Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Proc. of the 9th ISWC. (2010)
13. Seco, N., Veale, T., Hayes, J.: An Intrinsic Information Content measure for Semantic Similarity in WordNet. In: Proc. of ECAI 2004. (2004) 1089–1090
14. Kohler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dolken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. The American Journal of Human Genetics **85**(4) (2009) 457–464
15. Tao, Y., Sam, L., Li, J., Friedman, C., Lussier, Y.A.: Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics **23**(13) (2007) i529–i538
16. Berardini, T.Z., et al.: The Gene Ontology in 2010: extensions and refinements. Nucleic acids research **38** (2010) D331–D335
17. Lei, Z., Dai, Y.: Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC bioinformatics **7**(1) (2006) 491
18. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19**(10) (2003) 1275–1283
19. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology **7**(11) (2009)
20. Ferreira, J.D., Couto, F.M.: Semantic similarity for automatic classification of chemical compounds. PLoS computational biology **6**(9) (2010)