CharaParser+EQ: Performance Evaluation Without Gold Standard

Hong Cui[1], Wasila Dahdul[2], Alexander T. Dececchi[2], Nizar Ibrahim[3], Paula Mabee[2], James P. Balhoff[4], Hariharan Gopalakrishnan[1]

1. School of Information Resources and Library Science, University of Arizona, Tucson, AZ 85719, {hongcui, hariharang}@email.arizona.edu
2. Department of Biology, University of South Dakota, Vermillion, SD 57069, {wasila.dahdul, alex.dececchi, pmabee}@usd.edu
3. Department of Organismal Biology and Anatomy, Chicago, IL, 60637, nibrahim@uchicago.edu
4. National Evolutionary Synthesis Center, Durham, NC 27705, balhoff@nescent.org

ABSTRACT

To make phenotypic characters of organisms widely useful for computerized biology research, biocurators manually convert character descriptions to a structured format, for example the Entity-Quality (EQ) format. The manual approach is time consuming and affected by inter-curator variations. In this paper we report a software application, CharaParser+EQ, to our knowledge the first software that produces EQ statements from textual character descriptions. We report a recent experiment that evaluates the performance of the software against three experienced biocurators. While the software is still far from being able to compete with biocurators on this highly intellectual task, the results show (1) CharaParser+EQ's performance (precision and recall) is greatly improved compared to a previous version, (2) the completeness of the ontologies used in the process has significant impact both on the software's EQ generation performance and on the agreement among curators, and (3) unlimited access to external knowledge (published papers, books) by curators has no significant impact on inter-curator agreements. A detailed error analysis that compares machine and curator generated EQs is included.

Keywords

Phenotype character curation, EQ statements, Natural Language Processing, curation inconsistency, ontology search

INTRODUCTION

Phenotypic descriptions of organisms are used in almost all areas of biological research, yet the vast majority of such information is locked in a human readable form not directly usable for computerized analyses. Various projects (e.g., Blake et al., 2009; Bradford et al., 2011; Bowes et al., 2008; Howe et al. 2011) employ biocurators, mostly often postdoctoral experts with extensive specialist knowledge, to manually convert published phenotypic descriptions into a computable format. One such format gaining popularity is the Entity-Quality (EQ) format (for more information see BACKGROUND). The Phenoscape project (Mabee et al., 2012) has transformed vertebrate phenotypes from over one hundred comparative studies into Entity-Quality (EQ) statements using an EQ editor called Phenex (Balhoff et al., 2010). These EQ statements have been used in machine reasoning to generate biologically interesting and meaningful inferences (e.g,, Dececchi et al., in press).

While expert-curated EQs are of high quality, the manual approach is not scalable to converting all character descriptions into the EQ formalism. To automate the process of EQ generation, we have started to develop a software application, called CharaParser+EQ.

In 2012, the first version of CharaParser+EQ, introduced as part of the Phenoscape Curation System (PCS), participated in the BioCreative 2012 Workshop Track III: interactive text mining task (Arighi et al., 2013). After further development on the EQ module of CharaParser+EQ another CharaParser+EQ evaluation experiment was conducted in summer 2013 with the Phenoscape project.

In this paper, we report the second evaluation experiment in detail; in addition, the 2013 version of CharaParser+EQ is compared with the 2012 version. The rest of the paper is organized as follows. BACKGROUND section uses examples to explain the EQ format. Next, the data and experimental design are described, followed by CharaParser+EQ algorithms and performance evaluation metrics. We then report a set of key results organized using subheadings in the RESULTS section, which is followed by a DISCUSSION. After reviewing related works, we conclude the paper with directions for future work.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

BACKGROUND

An EQ statement is an ontologized character statement consisting of an Entity (E), a Quality (Q), and an optional Related Entity (RE). For example,

Example 1:
Character: fusion of distal-carpal-1 + 2 : absent
E: *UBERON:distal carpal bone 1*
Q: *PATO:separated from*
RE: *UBERON:distal carpal bone 2.*
The character says “distal carpal bone 1 and 2 are not fused”. The resulting EQ statement is semantically equivalent to the original character description. Here, UBERON (Mungall et al., 2012) and PATO (PATO, 2015) are the ontologies used to formalize (“ontologize”) original phrases (e.g., *distal carpal 1*) with terms (also called terms or classes) in relevant ontologies (e.g., *UBERON:distal carpal bone 1*). Note, any of the E, Q, RE could be post-coordinated (or post-composed) when a desired term can not be found in ontologies.

Example 2:
Character: lateral pelvic glands: absent in males
E: *UBERON:gland and (part_of some (BSPO:lateral region and (part_of some UBERON:pelvis and (part_of some UBERON:male organism))))*
Q: *PATO:absent.*

In this example, BSPO (Dahdul et al., 2014) is another ontology used in Phenex and a RE is not needed. Because a pre-coordinated term for *lateral pelvic gland* is not in the ontologies provided, E needs to be post-composed with component terms, such as *UBERON:gland*, *BSPO:lateral region* etc.

As these examples show, EQ curation is knowledge-intensive, requiring thorough expertise in anatomy, ontologies, in addition to knowledge representation. This work is often performed by post-docs with training in biological ontologies and knowledge representation; however, due to the complexity nature of the task, substantial inter-curator variation exists in the EQ curation process and results (Arighi et al., 2013; Camon et al., 2005).

DATA AND EXPERIMENT DESIGN

The 2013 CharaParser+EQ experiment was designed to evaluate (1) the performance of CharaParser+EQ as compared to curators, (2) the effects of the completeness of ontologies on curator and machine curation results, and (3) the effect of domain knowledge on curators’ EQ generation results. A gold-standard was not used in this evaluation because 1) such a standard does not exist and 2) because substantial inter-curator variation has been widely reported before (Arighi, et al., 2013; Camon et al., 2005; Wiegers et al, 2009). Instead, all three curators working on the Phenoscape project at the time of the experiment participated in the evaluation. Each curator has longer than one year curation experience on the Phenoscape project. CharaParser+EQ’s performance is compared to each of the curators.

A prospective power test was conducted and the result suggests that the sample size of the characters need to be greater than 150 characters. In the end, a total of 203 characters were selected for the experiment. These include 29 characters randomly selected from each of the 7 phylogenetic publications that were selected by a senior bioscientist. The paper selection aimed to maximize the taxonomic breadth of the characters (but within curators’ expertise), including both extinct and extant taxa and characters from different anatomical systems (e.g., skeletal, muscular, nervous systems). The selection of the characters was blind to CharaParser+EQ developers and the curators.

The three curators manually curated the same set of characters twice: the first time (a.k.a. the Naive Round) without access to external knowledge (published papers, specialist books), and the second time (a.k.a. the Knowledge Round) with access to external sources. In other words, in the Naïve Round, the EQs created were based solely on the character descriptions and curators’ specialist knowledge, while in the Knowledge Round, the curators were allowed to access external resources such as the full text publication, textbooks, or the Web; they were not allowed to communicate with the other curators, however. Our hypothesis was that access to external knowledge would make the EQs across all three curators more similar and would make the machine EQs more different from the human curated EQs. In total, 6 sets of EQs were created by the three curators in the two rounds of curation.

Curators and CharaParser+EQ developers used the same set of curation guidelines[1] developed by the Phenoscape project. Software developers used the guidelines to develop the heuristic rules used in the EQ generation module. Curators studied the guidelines as a group before curation. Curator curation process was blind to CharaParser+EQ developers during the experiment.

Both curators and CharaParser+EQ were given the same initial set of ontologies: UBERON (version phenoscape-ext /2013-03-15), BSPO (2013-05-17 release), and PATO (2013-06-03 release). During the two rounds of curation, the curators were allowed to independently add any needed (pre-coordinated) terms to the ontologies, as this is part of their normal practice. At the end of the two rounds of curation, six sets of augmented ontologies were generated. Besides the initial set and the six sets of augmented ontologies, an eighth set (a.k.a. the final set), was generated afterwards by adding the de-duplicated set of all term additions from three curators into the initial set of ontologies. CharaParser+EQ was run eight times, each with a different set of ontologies. EQs generated by the software

[1] http://phenoscape.org/wiki/Guide_to_Character_Annotation date Feb 2013

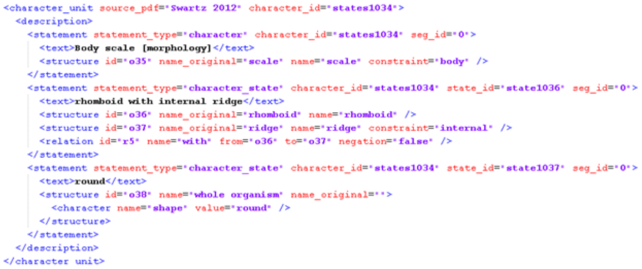
were compared to those created by all individual curators in two runs. Table 1 lists the 18 comparisons.

ID	Curator EQs	Machine EQs and Ontologies used
1	naive 1	CP+EQ using the initial set
2	naive 2	CP+EQ using the initial set
3	naive 3	CP+EQ using the initial set
4	know 1	CP+EQ using the initial set
5	know 2	CP+EQ using the initial set
6	know 3	CP+EQ using the initial set
7	naive 1	CP+EQ using the naive 1 augmented set
8	naive 2	CP+EQ using the naive 2 augmented set
9	naive 3	CP+EQ using the naive 3 augmented set
10	know 1	CP+EQ using the know. 4 augmented set
11	know 2	CP+EQ using the know. 4 augmented set
12	know 3	CP+EQ using the know. 4 augmented set
13	naive 1	CP+EQ using the final set
14	naive 2	CP+EQ using the final set
15	naive 3	CP+EQ using the final set
16	know 1	CP+EQ using the final set
17	know 2	CP+EQ using the final set
18	know 3	CP+EQ using the final set

Table 1. 18 Comparisons between the EQs created by the curators and by CharaParser+EQ using different ontology sets

METHODS

The key components of the software are CharaParser (Cui, 2012) and an EQ generation module. CharaParser employs unsupervised machine learning and syntactic parsing methods to mark up and associate entity and quality phrases**.** Figure 1 shows an example output of CharaParser in XML (eXtensible Markup Language) format.



```xml
<character_unit source_pdf="Swartz 2012" character_id="states1034">
  <description>
    <statement statement_type="character" character_id="states1034" seg_id="0">
      <text>Body scale [morphology]</text>
      <structure id="o35" name_original="scale" name="scale" constraint="body" />
    </statement>
    <statement statement_type="character_state" character_id="states1034" state_id="state1036" seg_id="0">
      <text>rhomboid with internal ridge</text>
      <structure id="o36" name_original="rhomboid" name="rhomboid" />
      <structure id="o37" name_original="ridge" name="ridge" constraint="internal" />
      <relation id="r5" name="with" from="o36" to="o37" negation="false" />
    </statement>
    <statement statement_type="character_state" character_id="states1034" state_id="state1037" seg_id="0">
      <text>round</text>
      <structure id="o38" name="whole organism" name_original="">
        <character name="shape" value="round" />
      </structure>
    </statement>
  </description>
</character unit>
```

Figure 1. Example output of CharaParser. Given the character text (shown in black), CharaParser parses and marks-up the structures (entities), characters (qualities), and relations (qualities) in XML.

The EQ generation module queries relevant ontologies and uses a set of heuristic rules to convert the XML output of CharaParser to EQ statements as the final output. Identifying E/Q phrases and then translating the phrases into ontology terms is how biocurators create EQ statements, and CharaParser+EQ follows the same logic steps. Figure 2 shows the complete logic flow of CharaParser+EQ.[2]

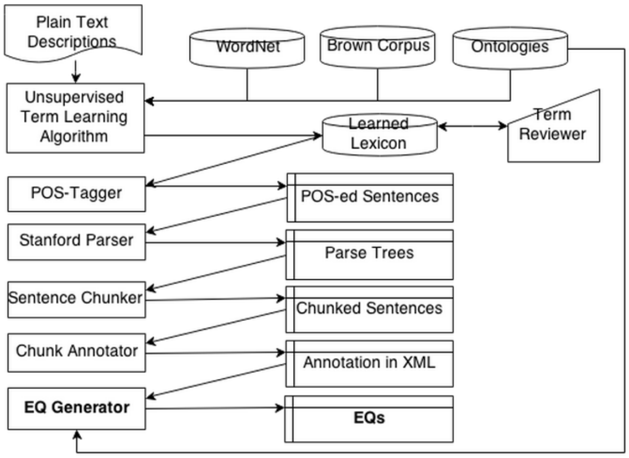


Figure 2. CharaParser+EQ. The components labeled in bold are the EQ Generation modules, the rest are original CharaParser components.

The Unsupervised Term Learning algorithm (Cui, Boufford, & Selden, 2010) efficiently performs bootstrapping-based learning on plain-text descriptions to learn entity (mostly noun) and quality (mostly adjective) phrases , essentially assigning POS (Part of Speech) tags to biological terms but leaving common English words for the Stanford Parser (Klein & Manning, 2003) to tag. The learned lexicon is then reviewed and approved by the user using the Term Reviewer interface. With the corrected lexicon, the POS-Tagger partially tags sentences before they are passed to the Stanford Parser, because the latter often tags biological terms wrong (e.g., mistaking *stems* as a verb). The term learner makes use of the POS information in WordNet when the contextual information in the text is not sufficient to determine the role for a word. It uses Brown Corpus (excluding the "learned" section) (Francis & Kucera, 1964) to identify common English words. Ontologies are also used as an existing lexicon to identify entity phrases, but not yet quality phrases, in sentences.

From the partially POS-tagged sentences, the Stanford Parser produces parse trees using its lexicalized probabilistic context-free grammar (PCFG). Then Sentence Chunker uses a set of heuristic rules to chunk parse trees into different types of entity and quality chunks, which are converted to XML markup as output by Chunk Annotator. Details on chunks and chunk annotations are provided as Appendix B in Cui (2012). The XML markup identifies and associates entity and quality phrases/candidates that are then ontologized by the EQ Generator using relevant ontologies.

The 2012 BioCreative evaluation results suggested that CharaParser was able to identify entity and quality phrases as well as the curators, but the EQ generation module performed at less than half of the curators' performance. Since 2012, the EQ generation module has been the main focus of development.

[2] Source at https://github.com/phenoscape/phenoscape-nlp/releases/tag/v0.1-alpha

The 2012 EQ generation module was designed to generate exact EQs from candidate phrases marked in XML output: it first transforms structure elements to Entity and Related Entity phrases and character and relation elements to Quality phrases, then it searches ontologies for terms that match the E, Q, and RE phrases and generates EQs. The two-step algorithm has the drawback of losing access to all the information provided in the XML markup once the program enters the look up step. The 2013 EQ generation module is modified to allow access to the XML output throughout. It generates candidate E, Q, RE phrases and looks them up in the ontology. If lookup is not successful, it employs various strategies to re-coordinate candidate phrases, apply wildcards, modify candidates or switch between entities and qualities to find matches. It utilizes ontological relations such as *is_a* and *part_of* in ontology lookup and EQ generation. It generates multiple EQ proposals or partially ontologized EQs for uncertain cases. In addition, entity terms in ontologies are provided as input to the software, so pre-coordinated entity terms are matched with ease.

The most important change was made in the ontology lookup methods. The example shown in Figure 1 illustrates some of the thorny issues. The reader probably has noticed that *rhomboid* should be a shape (quality), not a structure (entity). However, *rhomboid* is also a name for a muscle in UBERON so it can also be an entity. On the other hand, *internal ridge* seems to be an entity, but because many different anatomical structures (e.g., bones, fish scales) have ridges, they are not considered an entity in the ontologies, instead they are treated as a feature (i.e., ridged), of some entity (e.g., bones), and indeed *ridged* is included in PATO, a quality ontology. The boundary of entity and quality terms is therefore not always well defined, and is dependent on the context and the design of the ontologies. This led to the change of the algorithm from a 2-step strategy to the current integrated step that takes evidence from different sources to disambiguate entity and quality terms.

We must note here, however, Q: PATO:ridged is a wrong quality for the last example, because PATO:ridged is defined as "an elongated raised *margin or border*"(PATO release 2015-03-15 and before), so *internal ridges* should not be ontologized as PATO:ridged. Reading into human readable definitions is beyond what a computer algorithm can do at this time, and it also presents a challenge for human curators. In the end, *with internal ridge* was left un-ontologized by curators.

Precision and Recall Calculation

Classic precision and recall measures for assessing the effectiveness of information retrieval systems cannot be directly used here, because for one character description, a varied number of EQs may be generated as the character description may express multiple qualities for one or more entities. Further, CharaParser+EQ could generate multiple proposals for one EQ. Figure 3 illustrates the scenario. The measures we need must be able to assess the similarity between two sets of EQs created for the same character.

A Jaccard Similarity based measure proposed in Mistry & Pavlidis (2008) provides a good semantic based similarity measurement for EQs, however, it only works on fully ontologized EQs, and requires super-computing power to create an artificial ontology merging all ontologies used in the EQs with exponential numbers of generated nodes. To take into account the 15% of EQs proposed by CharaParser+EQ that contain one or more un-ontologized components[3], we have developed the following precision/recall-like measures that are used in the current evaluation.

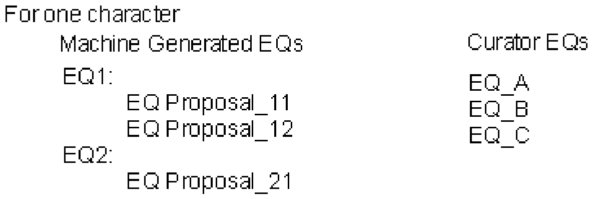


Figure 3. Varied number of EQs and EQ proposals could be generated from one character.

EQ Precision = sum of similarity of proposed EQs to answer EQs for a character / number of proposed EQs for all characters

EQ Recall = sum of similarity of answer EQs to proposed EQs for a character / number of answer EQs for all characters

Similarity of EQs is the average similarity of the proposed entity, quality, and related entity components to those of the curators (i.e., answer EQs). Figure 4 illustrates the procedure of calculating entity similarity. Quality and related entity similarities are calculated in the same way. With entity, quality, and related entity similarity scores, the similarity of a machine proposed EQ to a curator created EQ is the average of the entity, quality, and related entity similarity.

As shown in Figure 4, because a varied number of entities may be created from any character and in any order, the similarity of machine proposed entities to curator created entities is calculated by comparing each pair of machine proposed entity and curator created entity and summing up the greatest similarity of all proposed entities to a curator entity.

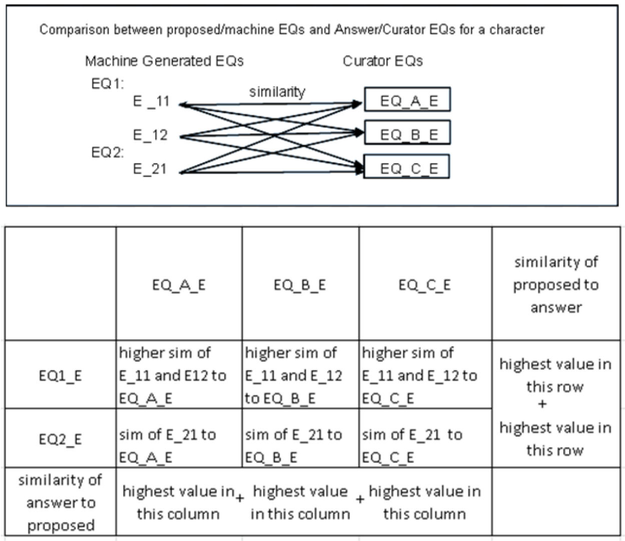
Now we narrowed the problem down to the calculation of the similarity score between a pair of entities, qualities, or related entities, which are often complex expressions post-

[3] Un-ontologized components are included in an EQ when CharaParser+EQ fails to find a matching term in the ontologies.

composed with individual ontology terms. Between two expressions under comparison, the longest common subsequence (LCS) is identified, and

Similarity of E/Q/RE[4] = [(sum of the similarity scores of the terms in the LCS) + 0.5 * (sum of the similarity scores of the terms not in the LCS)]/ average number of terms in the answer and test sentences

Individual ontology terms are scored using the following rubric: the similarity of two terms is 1 if they have the same IDs (i.e., IRIs), 0.75 if they are a partial match, or 0 otherwise. Two terms T1 and T2 are considered a partial match if any of the following is true: *is_a*(T1, T2), *is_a*(T2, T1), *part_of*(T1, T2), *part_of*(T2, T1), that is, if the entity/quality represented by one term *is_a* or is *part of* the entity/quality represented by another term.



	EQ_A_E	EQ_B_E	EQ_C_E	similarity of proposed to answer
EQ1_E	higher sim of E_11 and E12 to EQ_A_E	higher sim of E_11 and E_12 to EQ_B_E	higher sim of E_11 and E_12 to EQ_C_E	highest value in this row + highest value in this row
EQ2_E	sim of E_21 to EQ_A_E	sim of E_21 to EQ_B_E	sim of E_21 to EQ_C_E	
similarity of answer to proposed	highest value in this column +	highest value in this column +	highest value in this column	

Figure 4. Illustration of entity similarity calculation

RESULTS

CharaParser+EQ performance

Given the EQs output by curators and the machine, we calculated precision and recall of the machine's output against the curators'. Table 2 gives the precision and recall scores of the comparisons listed in Table 1, along with the number of EQs created by the curators and by CharaParser+EQ under different settings.

Paired T-tests between the performance (both recall and precision) of the comparisons 1-6 and those of 7-12 show statistically significant differences ($p < 0.05$). Paired T-tests between the performance of the comparisons 7-12 and those of 13-18 do not show statistical significance. Non-parametric tests (Wilcoxon) yield the same results. The average recall of the comparisons 1-6 is 0.42. The average recall of the comparisons 7-13 is 0.49, representing a 16.7% increase. The precision is also increased by 14.3%.

ID	round	# of Curator EQs	# of CP+EQ EQs	Precision	Recall
1	naive1	618	848	0.33	0.39
2	naive2	569	848	0.39	0.46
3	naive3	560	848	0.38	0.45
4	know1	608	848	0.32	0.38
5	know2	592	848	0.36	0.42
6	know3	551	848	0.36	0.43
Avg(1-6)				0.36	0.42
7	naive1	618	846	0.4	0.48
8	naive2	569	836	0.41	0.48
9	naive3	560	833	0.41	0.49
10	know1	608	843	0.4	0.49
11	know2	592	837	0.4	0.47
12	know3	551	836	0.41	0.51
Avg(7-12)				0.41	0.49
13	naive1	618	855	0.4	0.47
14	naive2	569	855	0.4	0.48
15	naive3	560	855	0.41	0.5
16	know1	608	855	0.4	0.49
17	know2	592	855	0.39	0.47
18	know3	551	855	0.4	0.51
Avg(13-18)				0.4	0.49

Table 2. Performance of CharaParser+EQ against curator EQs using different sets of ontologies. IDs refer to the IDs used in Table 1.

Performance comparison between CharaParser+EQ 2012 and 2013

To compare the performance of the 2013 CharaParser+EQ with the 2012 version, we run 2012 version on the same set of 203 characters using the final set of ontologies and compared the resulting EQ with curators' EQs. The results (Table 3) show the average performance of current CharaParser+EQ using the final set ontologies (precision = 0.4, recall=0.49) is significantly better than the 2012 version ($p < 0.005$), with a 21% increase in precision, and 75% increase in recall.

	Precision	Recall
naïve 1	0.32	0.26
naïve 2	0.36	0.31
naïve 3	0.34	0.29
know 1	0.32	0.27
know 2	0.30	0.31
know 3	0.36	0.25
average	0.33	0.28

Table 3: Performance of 2012 CharaParser+EQ against curator EQs using the final set ontologies

Inter-curator agreement

We also calculated precisions and recalls of one curator's output as compared to another and this set of results are presented in Table 4. The scores in bold are the comparison of the EQs created in the Naïve Round to those created in the Knowledge Round by the same curator on the same set of character descriptions.

[4] In calculating the E/Q/RE similarity, individual terms are matched to terms in the answer key once, making the similarity score a value between 0 and 1.

Precision	naïve 1	naïve 2	naïve 3	know 1	know 2	know 3
naïve 1	1	0.56	0.57	**0.78**	0.57	0.59
naïve 2	0.58	1	0.57	0.58	**0.68**	0.56
naïve 3	0.59	0.56	1	0.58	0.56	**0.71**
know 1	**0.79**	0.56	0.56	1	0.58	0.58
know 2	0.59	**0.67**	0.55	0.6	1	0.56
know 3	0.62	0.56	**0.72**	0.61	0.58	1
Recall	**naïve 1**	**naïve 2**	**naïve 3**	**know 1**	**know 2**	**know 3**
naïve 1	1	0.60	0.60	**0.81**	0.60	0.63
naïve 2	0.59	1	0.58	0.58	**0.69**	0.57
naïve 3	0.60	0.59	1	0.59	0.57	**0.72**
know 1	**0.80**	0.59	0.59	1	0.61	0.62
know 2	0.59	**0.69**	0.57	0.61	1	0.58
know 3	0.62	0.57	**0.72**	0.61	0.57	1

Table 4. Comparison of the EQs created by different curators in the Naïve and Knowledge Rounds.

Recalls of curators results as compared with one another range from 0.56 to 0.63 (average 0.59) while recalls of CharaParser+EQ with augmented ontologies range from 0.47 to 0.51 (average 0.49). The recall difference between the software and curators is 0.10 or less, while the recall difference among curators is less than or equal to 0.07 (=0.63-0.56). The precision difference between CharaParser+EQ and curators is much greater. Average precision of curators is 0.58 while the average precision of CharaParser+EQ best performance is 0.41. Performance differences between the machine and curators are statistically significant ($p<0.05$).

		%	avg EQ complexity
naïve_1 vs. know_1	changed	56	5.82
	unchanged	44	3.84
naïve_2 vs know_2	changed	70	5.06
	unchanged	30	3.85
naïve_3 vs. know_3	changed	64	5.23
	unchanged	36	3.69

Table 5. Changes in EQs made by curators between the Naive Round and the Knowledge Round and the effect on CharaParser+EQ precision and recall scores

Curator EQs changes in the Knowledge Round

The self-comparison scores (in bold) shown in Table 4 show that the EQs curators created in the Naïve Round are substantially different from those created in the knowledge round. Table 5 shows the percentages of the characters with changed/unchanged EQs and the complexity[5] of the changed vs. unchanged EQs. Table 5 shows that character states that are changed have an average complexity greater than 5, while the states that are not changed have an average complexity less than 4. Among the EQs that were changed in the Knowledge Round, we observed a mixed effect in terms of EQ complexity: 29% of changed EQs with increase complexity (i.e., with new components added to the EQ), 33% with decreased complexity, and 38% with unchanged complexity (i.e., curators replaced some parts in an EQ without changing its component count).

Impact of External Knowledge

Despite the fact that each curator changed EQs for more than half of the characters in the Knowledge Round, the performance differences between CharaParser+EQ and the curators in the Naive Round and the Knowledge Round were not statistically significant at 0.95 confidence level, as indicated by paired T test or Wilcoxon test on precision or recall scores between IDs 1, 2, 3, 7, 8, 9, 13, 14, 15 and 4, 5, 6, 10, 11, 12, 16, 17, 18 in Table 2. Further, the inter-curator recalls and precisions between the two rounds are not significantly different, as indicated by paired T test or Wilcoxon on precision or recall scores between naïve1-naive2, naive1-naive3, naive2-naive3 and know1-know2, know1-know3, know2-know3 in Table 4.

	Setting	Entity		Quality		Related Entity	
		p	r	p	r	p	R
13	naive 1_CP_best	0.32	0.34	0.46	0.52	0.45	0.14
14	naive 2_CP_best	0.31	0.35	0.51	0.57	0.36	0.18
15	naive 3_CP_best	0.34	0.36	0.45	0.53	0.36	0.18
16	knowledge 1_CP_best	0.32	0.34	0.44	0.52	0.45	0.14
17	knowledge 2_CP_best	0.33	0.36	0.48	0.53	0.37	0.14
18	knowledge 3_CP_best	0.33	0.36	0.45	0.56	0.36	0.17
Avg		0.33	0.35	0.47	0.54	0.39	0.16

Table 6. CharaParser+EQ performance on entity, quality, and related entity

Setting	Entity		Quality		Related Entity	
	p	r	p	r	p	r
naive 1_naive 2	0.46	0.47	0.56	0.59	0.38	0.61
naive 1_naive 3	0.48	0.47	0.54	0.59	0.39	0.58
naïve 2_naive 3	0.47	0.45	0.61	0.63	0.51	0.48
know 1_know 2	0.45	0.45	0.55	0.56	0.45	0.59
know 1_know 3	0.42	0.42	0.54	0.58	0.36	0.56
know 2_know 3	0.44	0.44	0.56	0.58	0.40	0.46
Avg	0.45	0.45	0.56	0.59	0.42	0.55

Table 7. Curator agreement on entity, quality, and related entity

[5] EQ Complexity is measured as the average number of ontology term components in an EQ statement.

Performance on EQ Components

Precision and recall scores on entities, qualities, and related entities were calculated for CharaParser+EQ as well as for curators. Tables 6 and 7 present some representative results. Both curators and CharaParser+EQ scored higher on qualities than entities. CharaParser+EQ recalls on related entity were very low.

Performance without Partial Scores

When precision and recall scores were calculated, we assigned partial match score 0.75 to EQ components that are not an exact match, but have either is_a or part_of relationship to the target component. When discounting partial scores, precisions and recalls of curators were reduced by 0.02 on average, and precisions and recalls of CharaParser+EQ were reduced by 0.04-0.05 on average.

DISCUSSION

The experiment results show that the performance of CharaParser+EQ improves significantly as compared to the 2012 version (recall is improved 75% and precision is improved 21%). CharaParser+EQ reached 0.41/0.49 precision/recall with augmented ontologies used in the experiment. We attribute the performance improvements to the more effective ontology lookup method and the pre-matching of entity terms in descriptions.

The results also suggest that CharaParser+EQ perform better with more complete ontologies. Using individual curator augmented or the final ontologies, CharaParser+EQ precision is improved by 14% and recall by 16% (Table 2).

However, a significant gap between the performances of CharaParser+EQ and expert curators remains. The average precision score of CharaParser+EQ with augmented ontologies is 0.1 lower than the curators' average (=16.95% lower), and recall is 0.17 lower (=29.31% lower). The performance gap is also seen through the larger impact on performance scores of CharaParser+EQ when partial match scores are discounted. We have identified a number of weaknesses of CharaParser+EQ, including its difficulties with related entities (Table 6), that we need to work on in the future.

Curators and CharaParser+EQ created more consistent or better qualities than entities. This is probably due to the complex post-compositions needed to create entities when a pre-composed match cannot be located in or added to the ontologies. Composing qualities is much less common than composing entities. Only 5% of qualities were post-composed in the curator results, while 47% of entities were post-composed. Post-composition of entities is also one main difficulty faced by CharaParser+EQ and a major factor contributing to inter-curator variations.

Consistent with previous research (e.g., Arighi et al., 2013; Camon, 2005; Söhngen 2011; Wiegers 2009), substantial inter-curator variation is also seen in this experiment (59%/58% precision/recall on average when one curator's EQs are compared to those of another). It was unexpected to find that external knowledge did not help curators to converge on their EQs, nor did it widen the gap between curator and machine generated EQs (Table 2, Table 4).

To better understand where CharaParser+EQ fails and to uncover potential causes of inter-curator variation, a manual review of the CharaParser+EQ EQs, generated using the final set of ontologies, and the EQs created by the three curators in the knowledge round has been conducted. This review covers 30% of the 203 characters included in this experiment.

Challenging cases for CharaParser+EQ include:

1. Post-composing complex entities, especially when the component terms are located far apart from the main entity.

Example 3:
Character: *Mandible*, proximal end small, *with stout medial process* and medial-cotyla forming a narrow sulcus: *no*
***Curator E**: UBERON:anatomical projection and (part_of some (BSPO:medial region and (part_of some UBERON:mandible)))*
***Curator Q**:PATO: absent*[6]

Example 4:
Character: *Pelvis, posterior portion, juncture of the ilium and the ischium*, posteriorly directed point: absent
***Curator E**: UBERON:anatomical projection and (part_of some (BSPO:posterior region and (part_of some UBERON:pelvis))) and (PATO:adjacent_to some UBERON:ilium) and (PATO:adjacent_to some UBERON: ischium)*
***Curator Q:** PATO:absent*[7]

In Example 3, the components, *mandible* and *medial process*, needed to post-compose the entity are separated by another clause (*proximal end small*). In Example 4, five components need to be post-composed and *juncture* needs to be translated to *PATO:adjacent_to*.

2. Mapping some text phrases to relational qualities (i.e., qualities that relate one entity with another).

Example 5:
Character: Prootic, entocarotid-fossa (GNC30): present as distinct fossa *within* the recessus-vena-jugularis.
***Curator E:** UBERONTEMP:entocarotid fossa*[8]
***Curator Q:** PATO:located in*
***Curator RE:** UBERONTEMP:recessus vena jugularis*

[6] Showing one curator's result. Curators may not all agree on Q/E/RE for these examples.

[7] This curator deemed *posteriorly directed point* too specific to be ontologized, so it is not expressed in the EQ.

[8] "TEMP" was used in the experiment to indicate the term is newly added by a curator to an ontology.

The word *within* is mapped to *PATO:located in*. As there is not a dictionary supporting such mappings, the EQ Generator module uses WordNet to discover possible mappings, but we acknowledge that is not enough to address the problem.

3. Processing implicit negations. For example, the word *absent* in *fusion absent* and *contact absent* expresses a negative meaning and it should be mapped to *PATO:separated from*.

4. Processing semantics encoded in special syntax or symbols. For example, the plus sign in *distal carpal 1+2* means two bones, *distal carpal 1* and *distal carpal 2,* are fused.

Besides the above issues, at the time of the experiment, searching PATO for multiple-word quality terms was not implemented in CharaParser+EQ, causing the system to fail trying to find qualities such as *PATO:posteriorly directed*, even when the XML output includes *posteriorly* as a modifier for *directed*.

The following set of issues affects both the machine and the curators. They are identified as the main factors contributing to curator variations, and they present even greater challenges to CharaParser+EQ.

A. Ontology search issues. For the machine, lexical-based word matching is the basic method to search an ontology but matches found this way can be semantically wrong matches. Curators sometimes run into the same problem. For example, for *tooth crown distinct from root*, CharaParser+EQ and some curators used *PATO:differentiated* as the quality for *distinct from*, but *PATO:differentiated is_a PATO:cellular potency*, so it is a wrong match semantically. For another example, multiple PATO terms may match the phrase *separate from*, including *PATO:separated from* (*is_a PATO:structure*), *PATO:far from* (*is_a PATO:position*), *PATO:adjacent to* (*is_a PATO:position*), and *not PATO: in contact with* (*is_a PATO:structure*). Differentiating these matches is often not straightforward, even for curators.

B. Vague character descriptions or descriptions requiring *highly* specialized knowledge to interpret. For example, the quality phrase *weak* in *dermal sculpture on skull-roof weak* invited a variety of interpretations from the curators: *PATO:poorly developed*, *PATO:decreased magnitude*, and *PATOTEMP:weakly sculptured surface*. On the other hand, *PATO:decreased strength* has *weak* as its exact synonym. The character shown as Example 1 earlier uses a single *no* to negate several entity and quality combination, it is not clear to an educated non-expert whether the negation applies to all entities (no mandible, or no medial-cotyla), qualities (not small, not stout, not forming, or not narrow) or any specific ones.

C. Incomplete ontologies. When a needed entity or quality term can't be found in an ontology, one can either propose a new pre-coordinated term and add it to the ontology (see Example 5) or post-compose a term using existing terms. Different new terms or post-composed terms may be created for the same entity or quality, so either strategy results in variations. For example, *because dermal sculpture* is not in UBERON, one curator created a new term *surface sculpting* and used it to post-compose *UBERONTEMP:surface sculpting and (part_of some UBERON:dermatocranium)* as the ontological translation for *dermal sculpture on skull-roof*. Another curator used *PATO: sculpted surface* to create a post-composed term *UBERON:dermatocranium and (bearer_of some PATO:sculpted surface)* to represent the same concept. Sometimes different post-composed terms may be semantically equivalent, but in this case EQs constructed using these two post-composed terms have different meanings[9]. Lacking good matching terms in ontologies also contributes to the variations in curation granularity. Because *dermal sculpture* cannot be found in UBERON, the third curator used *dermatocranium* (skull roof) alone as the Entity, leaving out *dermal sculpture*. The EQ resulted from this less granular curation may still meet the needs of the users, but is misleading semantically, changing the meaning of the description from "dermal sculpture weak" to "skull-roof weak".

RELATED WORK

CharaParser+EQ can be categorized as a semi-automated text annotation tool. Many text annotation tools have been developed in the past; some of them are primarily concerned with named entity recognition in general (Cunningham et al., 2013; Maynard et al., 2005; Popov et al., 2004) while others address specific applications (Gao, 2013; Rahman, 2012; Soo, 2004). In biology, the BioCreative Workshops evaluate systems that extract a variety of entities and associations, including gene mentions, normalized human gene/products (gene/products to Gene Ontology terms), protein-protein interaction, diseases, chemicals, gene and disease associations (Arighi, 2013; Hirschman 2005; Krallinger 2008; Leitner 2010).

Tools that extract terms for the purpose of indexing (e.g., Cunningham et al., 2013 and most of the systems evaluated at BioCreative) have different requirements from the tools that produce annotations for machine reasoning. The latter requires higher semantic accuracy, a higher degree of associations among extracted targets, and deeper understanding of logical relations among entities. Most of the latter systems employ a largely manual workflow and work as an interactive annotation tool. Phenex (Balhoff 2010) and Phenote (http://www.phenote.org/) are tools used by human curators to generate EQ annotations for different

[9] *UBERON:dermatocranium and (bearer_of some PATO:sculpted surface)* means "dermatocranium with sculpted surface", which is different from "sculpted surface of dermatocranium" (or *dermal sculpture on skull-roof*).

kinds of phenotype descriptions. CharaParser+EQ attempts to automate some of the steps a human curator undertakes. We are not aware of other automated or semi-automated systems that produce EQs.

Part of the task of the EQ Generator in CharaParser+EQ (Figure 2) is to map a phrase to a term or terms in an ontology. Our mapping method bears similarity with MetaMap (Aronson, 2001; Aronson, 2010) as both perform syntactic parsing to identify candidate phrases and both generate different lexical variations of a phrase to increase the chance of matching. MetaMap includes some steps not included here, for example, assessing character sequence similarity between two words (i.e. measuring the similarity between *apnea* and *apnoea*). Other similar concept recognizing systems include KnowledgeMap (Denny, 2003), MGrep (Dai, 2008), and ConceptMapper (Apache UIMA Development Community 2014). CharaParser+EQ differs from these tools in that the EQ Generator post-composes complex terms in addition to recognizing existing terms (i.e., pre-composed terms) in ontologies. In addition, all these systems focus on concept mapping only, while EQ Generator also attempts to establish correct associations among and within entities, qualities, and related entities. For example, simply mapping *males* to *UBERON:male organisms* (a potential entity) in Example 2 not sufficient to score high in the EQ generation task, as it may produce a spurious EQ, E: *UBERON:male organisms Q:PATO:absent.*

CONCLUSION

We report an experiment that evaluates the performance of CharaParser+EQ on EQ generation. A gold-standard was not created nor used in this evaluation because of the well documented existence of inter-curator variations (Arighi et al., 2013; Camon et al., 2005; Wiegers et al., 2009). Instead, the evaluation was conducted more realistically against three expert biocurators (with and without access to external knowledge).

While CharaParser+EQ's performance has improved since the last formal evaluation in 2012, there is still a long way to go to high-throughput EQ production from text. We will continue to improve software performance and at the same time work to develop (a) more complete ontologies by adding terms identified in CharaParser XML output to ontologies, (b) error-preventing ontology searching mechanisms, for example, signal the user the superclasses of a matched term to prevent incorrect lexical-based matches, (c) clear guidelines and error-preventing mechanisms on the formation of pre-coordinated vs. post-coordinated terms, (d) logical definitions for terms with the duality of entity and quality roles, for example, define *sculptured* as *has_part some sculpture* and (e) consistency promoting mechanisms, for example, cross-check EQs of similar character statements and alert the user the potential inconsistencies.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers DBI-1062542, EF-0849982, and DBI-1147266. The authors acknowledge the constructive suggestions offered by Todd Vision, Hilmar Lapp, Prashanti Manda, Melissa Haendel, and Chris Mungall on experiment design and ontology access and use. We thank the anonymous reviewers for their suggestions on improving the readability of the paper.

REFERENCES

Apache UIMA Development Community. Apache ConceptMapper annotator user Guide (2011). Retrieved July 6, 2015 from https://uima.apache.org/d/uima-addons-current/ConceptMapper/ConceptMapperAnnotatorUserGuide.html.

Arighi, C.N., Carterette, B., Cohen, K.B., et al. (2013). An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)* 2013:bas056.

Aronson, A.R (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of American Medical Informatics Association 2001 Annual Symposium* (pp.17–21).

Aronson, A.R., & Lang, F-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association 17*, 229–236.

Balhoff, J.P., Dahdul, W.M., Kothari, C.R., et al. (2010). Phenex: Ontological Annotation of Phenotypic Diversity. *PLoS ONE, 5* (5), e10500. doi:10.1371/journal.pone.0010500.

Banerjee, S. & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005).* Retrieved July 6, 2015 from http://www.cs.cmu.edu/~alavie/papers/BanerjeeLavie2005-final.pdf.

Blake, J.A., Bult, C.J., Eppig, J.T., et al. (2009). The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Research 37*, D712-719.

Bowes, J.B., Snyder, K.A., Segerdell, E., et al. (2008). Xenbase: a Xenopus biology and genomics resource. *Nucleic Acids Research 36*, D761-767.

Bradford, Y, Conlin, T, Dunn, N, et al. (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Research 39*: D822-829.

Camon, E.B., Barrell, D.G., Dimmer, E.C., et al. (2005). An evaluation of GO annotation retrieval for

BioCreAtIvE and GOA. *BMC bioinformatics 6*, S17. doi:10.1186/1471-2105-6-S1-S17.

Cui, H (2012). CharaParser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology 63*(4), 738-754.

Cui, H., Boufford, D., & Selden, P. (2010). Semantic annotation of biosystematics literature without training examples. *Journal of American Society of Information Science and Technology 61*(3), 522–542.

Cunningham, H., Tablan, V., Roberts, A., et al. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology 9*, e1002854.

Dahdul, W.M., Cui, H., Mabee, P. et al. (2014) The Biological Spatial Ontology: anatomical descriptors for spatial and topological aspects of biological structures. *Journal of Biomedical Semantics. 5* (34), doi:10.1186/2041-1480-5-34.

Dai, M., Shah, N.H., Xuan, W., et al. (2008). An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics* 21

Dececchi, T.A., Balhoff, J.P., Lapp, H. & et al. (in press). Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic Biology*.

Denny, J.C., Smithers, J.D., Miller, R.A., et al. (2003). "Understanding" medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association 10*, 351–362

Francis, W.N. & Kucera, H. (1964). Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. Retrieved July 6, 2015 from http://icame.uib.no/brown/bcm.html

Gao, L., Campbell, H.A., Bidder, O.R., et al. (2013). A Web-based semantic tagging and activity recognition system for species' accelerometry data. *Ecological Informatics 13*, 47–56.

Hirschman, L., Yeh, A., Blaschke, C., et al. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics 6*, S1. doi:10.1186/1471-2105-6-S1-S1.

Howe, D.G., Frazer, K., Fashena, D., et al. (2011). Data extraction, transformation, and dissemination through ZFIN. *Methods in cell biology 104*, 311-325.

Klein, D., & Manning, C.D. (2003). Accurate unlexicalized parsing. In Y. Matsumoto (Ed), *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423–430). Stroudsburg, PA: Association for Camputational Linguistics.

Krallinger, M., Morgan, A., Smith, L., et al. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biology 9*, **S1.**

Leitner, F., Mardis, S.A., Krallinger, M., et al. (2010). An overview of BioCreative II. 5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 7*, 385–399

Mabee, P., Balhoff, J.P., Dahdul, W.M., et al. (2012) 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. *Journal of Applied Ichthyology 28*, 300-305.

Maynard, D. (2005). Benchmarking ontology-based annotation tools for the semantic web. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC)*. Retrieved July 6, 2015 from https://gate.ac.uk/sale/lrec2008/benchmarking.pdf.

Mistry, M. & Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics* 9, 327-337.

Mungall, C.J., Torniai, C., Gkoutos, G.V., et al. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology 13*, R5-R5.

PATO: Phenotypic Quality Ontology. Retrieved July 6, 2015 from http://wiki.obofoundry.org/wiki/index.php/PATO:Main_Page.

Popov, B., Kiryakov, A., Ognyanoff, D., et al. (2004). KIM-a semantic platform for information extraction and retrieval. *Natural Language Engineering 10*, 375–392.

Rahman, F. & Siddiqi, J. (2012). Semantic annotation of digital music. *Journal of Computer and System Sciences 78*, 1219–1231.

Skutschas, P.P. & Gubin, Y.M. (2012) A New Salamander from the Late Paleocene—Early Eocene of Ukraine. *Acta Palaeontologica* Polonica 57, 135-148.

Soo, V-W., Yang S-Y., Chen, S-L., et al. (2004). Ontology acquisition and semantic retrieval from semantic annotated Chinese poetry**.** In *Proceedings of 2004 Joint ACM/IEEE Conference on Digital Libraries* (pp. 345-346).

Wiegers, T.C., Davis, A.P., Cohen, K.B., et al. (2009). Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC bioinformatics 10*, 326-337**.**