Evaluating Open Information Extraction for Ontologization of a Thematic Domain

Gopala Krishna Koduri[1], Siva Reddy Gorla[2], Bharat Ram Ambati[2], and Xavier Serra[1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2] Institute for Language, Cognition and Computation, University of Edinburgh, UK
gopala.koduri@upf.edu

Abstract. In the past decade, domain-independent approaches to information extraction have paved way for its web-scale applications. Adapting them further to acquire knowledge from thematic domains can greatly reduce the need for manual knowledge engineering. This requires understanding how amenable the assertions extracted by such approaches are to ontologization. To this extent, we propose a framework for a comparative evaluation of the open information extraction systems. The first part of the framework compares the volume of assertions along different dimensions with an aim to understand their coverage of the domain quantitatively. In the second part, the assertions are evaluated qualitatively by employing them in three of the fundamental tasks of ontologization: object identification, concept identification and semantic relation extraction. The combined observations lead to useful insights about not only the quality of the assertions, but also the nature of the information extraction approaches.

1 Introduction

The advent of the semantic web and the linked open data movements have not only resulted in a growing number of community-built structured data sources, but also catalyzed the development of domain-independent approaches for extracting information from unstructured text, further enriching them. Open information extraction (OIE) is one such paradigm that has emerged in the past decade, and has been used to extract assertions from unstructured data at web-scale with a considerable success [1]. Until recently, domain-specific approaches to information extraction from text required manual knowledge engineering as a prerequisite [2]. The OIE approaches, however, do not require a prespecified vocabulary and/or relation-specific input. Therefore, adapting them to information extraction from thematic domains would alleviate the need or manual knowledge engineering.

The process of channeling the assertions extracted from these approaches into a coherent knowledge-base itself poses certain challenges. There has been little work so far to identify and address such issues. In this paper, we propose a framework for a comparative evaluation of the open information extraction approaches. The first part of the framework compares the volume of extracted assertions across different aspects of a given domain with an aim to understand the coverage of the domain quantitatively. In the second part of the framework, the assertions are used in three fundamental tasks

of ontologization: object identification, concept identification, and semantic relation extraction. The results from each task are validated against structured content in Wikipedia and/or are manually checked as necessary. The results from the two parts of the framework, when juxtaposed against each other, give us concrete insights into the differences between the performances and the nature of the approaches.

The remainder of the paper is organized as follows. In sec. xx, an overview of the data we work with is presented, and in sec. xx, the OIE approaches that we chose to compare are discussed. In sec. xx, we present the framework with various quantitative and qualitative measures for analysing their performances on the tasks of relation extraction and ontologization, and demonstrate it using a thematic domain. Sec. xx concludes the paper with our observations and remarks.

2 Open Information Extraction

Information extraction is the task of obtaining a set of assertions from the natural language text, featuring the entities and the relations of the corresponding domain. The approaches are diverse ranging from those which learn from the labeled training samples for the desired set of target relations, to those which operate in an unsupervised manner. An easy access to large volume of unstructured text on the web has necessitated approaches that scale appropriately to take advantage of this data. Open information extraction aims to extract the assertions from voluminous data without requiring a pre-specified vocabulary or labeled data for relations [1].

2.1 ReVerb & OpenIE 4.0

For demonstrating our evaluation framework, we choose two state-of-the-art OIE systems: ReVerb [3] and OpenIE 4.0 [4], which are shown to have outperformed the earlier systems such as TextRunner [1], woe^{pos} and woe^{parse} [5]. ReVerb addresses the issue of incoherent and uninformative extractions found with the former approaches, by using few syntactic and lexical constraints. OLLIE is a successor of ReVerb, and includes the noun-mediated relations which are not handled by the latter. It also incorporates the context of the assertions in the form n-ary relations. OpenIE 4.0 employs a similar methodology to that of OLLIE, to retrieve assertions by semantic role labeling, also known as shallow semantic parsing. The implementations for both ReVerb and OpenIE 4.0 are available online[3].

2.2 Semantic parsing

On the other hand, deep semantic parsing is an active research topic in the natural language processing (NLP) community, which aims to obtain a complete logical form of a given sentence. It facilitates applications such as question-answering systems, and further has several direct implications for OIE as it is domain-independent and is shown to be web-scalable [6]. To our knowledge, there is no existing literature that compares it

[3] Available at https://github.com/knowitall/

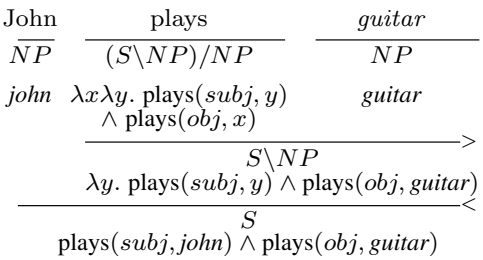


Fig. 1. An example showing the sentence `John plays guitar', parsed using CCG.

with the likes of ReVerb and OpenIE 4.0. We therefore built an information extraction wrapper around a state-of-the-art semantic parser, to be compared with the selected OIE approaches. What follows is a brief description of this system.

We use Combinatory Categorial Grammar (CCG) [7] as our grammatical framework to parse natural language sentences to logical representation. CCG is known for its transparency between syntax and semantics, i.e. given the syntactic structure (CCG derivation) of a sentence, a semantic representation can be built deterministically from its derivation. Each word in a sentence is first assigned a CCG category based on its context. Each category represents the syntactic constraints that the word has to satisfy. For example, in Figure. 1, the word *plays* is assigned a syntactic category $(S\backslash NP)/NP$ implying that *plays* take a noun (NP) argument on its right, and a noun argument (NP) on its left to form a sentence (S). An equivalent semantic category in terms of a lambda function is constructed from the syntactic category, here $\lambda x.\lambda y.\text{plays}(subj,\ y) \wedge \text{plays}(obj,\ x)$ with *plays* representing the predicate, x and y representing the object (guitar) and subject (John) arguments. CCG defines a set of combinators using which the adjacent categories combine to form syntactic categories of larger text units like phrases (*e.g:* plays guitar), from there on leading to parsing a whole sentence. Correspondingly, the lambda functions of the categories compose, eventually leading to semantic representation of the sentence. The advantage with CCG is that the complexity of obtaining a logical representation of the sentence is simplified into the task of assigning categories to words. We use a modified version of Boxer [8] to further convert our sentences of interest to binary assertions of the form <argument 1, relation phrase, argument 2>.

3 Data

A major challenge in developing technologies for the exploration and the navigation of music repertoires from around the world lies in obtaining and using their cultural context. The vocabulary used for describing and relating the entities (musical concepts, roles of people involved etc) differs to a great extent from music to music. Most commercial platforms have a limited view of such context, resulting in poor navigation and exploration systems that fail to address the cultural diversity of the world. Within the music information research community, there is a growing interest for developing culture-

aware approaches to address this problem [9]. Such approaches are diverse in terms of the data they work with (audio, metadata and contextual-data) and methodologies they employ [10].

However, to our knowledge, there are no major attempts that use text, arguably the largest openly available data source. As a first step in this direction, we choose to demonstrate our framework in the music domain. Indian art music traditions: Carnatic and Hindustani, have a very distinct character, especially when compared to the popular music styles that drive the music market worldwide. The terminology and the structuring of the knowledge in these music traditions differs substantially from what people are accustomed to, on most commercial platforms [11]. Therefore, they make a suitable yet challenging thematic domain to analyze the quality of the assertions for ontologization.

Our data consists of the plain text obtained from the Wikipedia pages corresponding to the Carnatic and Hindustani music traditions, after removing the tables, listings, figures, infoboxes and other structured content. Text from each page is tokenized to sentences, which are further filtered using the following constraints: a minimum number of 3 words and a maximum of 21 words per sentence, with each word not exceeding 30 characters in length. These constraints are empirically found to greatly reduce the number of malformed and highly complex sentences.

We observed that a majority of the sentences featured pronouns. The resulting assertions only partially contribute to ontologization. For instance, consider the sentence `She is a Composer'. The resulting assertion would be <She, is a, Composer>. With a few more of such sentences, it is possible to learn that there exists a semantic category called *composer*. However, such assertions are helpless in identifying objects of the corresponding semantic category. Therefore, the pronouns in the text from each page are resolved using the deterministic coreference resolution described in [12][4]. There were a few false assertions as a result. However, there is a substantial rise in the recall of the entities in the domain. Table. 1 list the total number of sentences, and the number of extractions from each of the OIE systems. ReVerb and Open IE 4.0 associate a confidence score with the extracted assertions. We did not however choose to filter these based on this score, as [13] have found that a system with a better recall at the cost of lower precision is actually preferred for knowledge-base population using open information extraction. We ignore the context from the n-ary assertions to convert them to binary form.

Music	#Sentences	#ReVerb	#OpenIE 4.0	#Sem. Parsing
Carnatic	10284	9844	15013	19241
Hindustani	10724	9944	15777	18496

Table 1. The number of sentences for each music, and the number of extractions using different OIE systems.

[4] Available online at http://nlp.stanford.edu/software/dcoref.shtml

4 Evaluation framework

In the information extraction literature, comparing different approaches by their performances on a set of labelled sentences or by employing human judges, ~~is a common practice~~ [3, 4]. Our goal, however, is to evaluate them by the usefulness of the assertions extracted. We quantify this using a series of tasks that help in understanding the coverage of the entities and the relations of a given domain in the extracted assertions, quantitatively and qualitatively. The tasks discussed in the first part of the evaluation comparing the volume of the assertions. While in the second part, we validate to what extent the assertions yield to be structured. We then juxtapose and compare the results from both parts of the evaluation.

4.1 Quantitative assessment

We study the distribution of the extracted assertions across four different aspects to gain an insight into their coverage of the domain with respect to each of them: **sentences**, **objects**, **relation types** and **concepts**. For the purpose of analyses discussed in this section, the first argument in the assertion is taken for an object, and the second argument of assertions featuring a subsumption relational phrase (eg: is a, be etc..) is taken for a concept.

Observations from the distribution of the extractions across sentences give a crude perspective of the modularity of the information extraction approach, which is its ability to identify multiple, distinct relations from a given sentence. The distribution of extractions across the objects allows us to gain an overview of the scope of the extracted relations in identifying the entities in the given domain as well as in describing a given entity. The distribution of extractions across relation types allows us to understand the relevance and coverage of an identified relation-type in the domain.

As we will see, a large majority of the assertions from all the OIE systems correspond to the subsumption relation type, often outnumbering the other relation types by orders of magnitude. Therefore, it is important to further analyze this relation type. These relations mainly inform us about the semantic category membership of the entities. Hence, they assume importance for ontologization as they are resourceful in defining the taxonomy of the given domain. The distribution of extractions per class would reveal to what extent the assertions actually carry the required information.

4.2 Qualitative assessment

The tasks discussed in this section are complementary to those presented in the former, validating whether the quantitative observations correlate with the performances of OIE approaches on various tasks in ontologization. For this, we consider the three fundamental tasks of ontologization: object identification, concept identification and semantic relation extraction [14].

Concept identification. It is the task of identifying the semantic categories in the domain. The second argument from all the assertions featuring the subsumption relation

type are collected. They are disambiguated based on their spellings, mostly automatically using string matching and edit distance measures[5] with minimal manual intervention where necessary. The resulting arguments are taken to be the candidate concepts of the given domain. We compare the coverage of these against the classes in the ontologies manually engineered with the help of music experts, for each of the Carnatic and Hindustani music[6].

Object identification. It concerns with finding the entities of a given domain and assigning a semantic category to them. The set of first arguments from all the assertions are considered as candidates for being objects in the domain. A list of titles of the Wikpedia pages in the domain along with the categories each page belong to, is acquired. The page titles correspond to objects, and the categories are manually mapped to classes in our ontology. This constitutes the reference with which we compare the results from the two subtasks of object identification.

For evaluating the first subtask of object identification, i.e., finding the entities of the given domain, we measure the overlapping (O) and the residual (R) portions of the candidate objects from each approach with respect to the reference set. If X is the set of candidate objects and Y is the reference set, O and R are defined as:

$$O(X,Y) = \frac{|X \cap Y|}{|Y|} \tag{1}$$

$$R(X,Y) = \frac{|X - Y|}{|X|}$$

The second part, semantic category assignment, is evaluated using two methods. In the first method, we manually build a set of rules over subsumption relation type for each semantic category. For instance, for an object to belong to the semantic category *singers*, it must have either of the words *vocalist* or *singer* in the second argument of the corresponding assertions with subsumption relation type. All the objects satisfying the rules for a given semantic category are assigned to it.

In the second method, a given object is reduced to be represented by a term vector corresponding to the words from second arguments of the assertions it is part of. Following this, each semantic category is initiated with a seedset of objects belonging to it. A given semantic category is taken to be an abstract super object, represented by the union of the term vectors of the constituting objects. A bootstrapping mechanism is started, which in a given iteration, finds the closest object to the given semantic category and adds it to the seedset, and recomputes its representation. The distance between given two objects corresponds to the cosine similarity between the term vectors transformed using TF-IDF, followed by Latent Semantic Indexing [15]. Unlike the first approach, which is constrained to assertions with subsumption relation type, this method takes advantage of the full spectrum of relation types. Results from the both methods are evaluated using O and R measures from eq. 1, where X and Y correspond to the candidate set of objects obtained using one of the methods for a given semantic category, and the reference set of objects respectively.

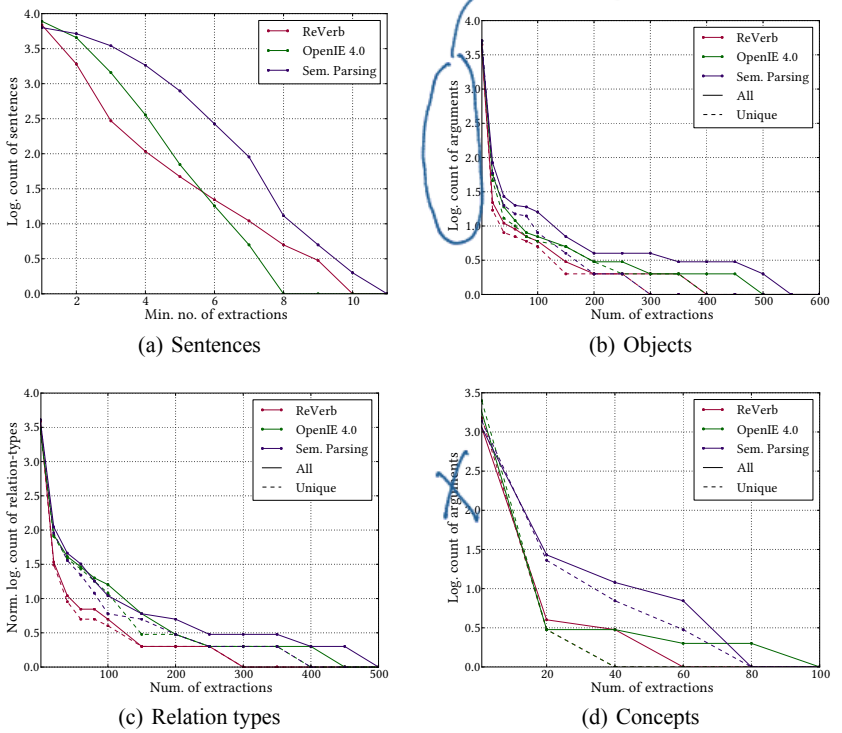
[5] Available at https://github.com/gopalkoduri/string-matching
[6] Available at https://github.com/gopalkoduri/ontologies

Semantic relation extraction. It refers to the relation types other than those which convey concept hierarchies. The assertion shown in Figure xx is one such example, where *plays* is a relation that connects *person* and *musical instrument* semantic categories. We formulate two measures to compare the OIE approaches in this task: breadth (B) and depth (D) of the identified relation types. B corresponds to the absolute number of valid relation types identified for each semantic category, and the D corresponds to the number of valid assertions for a given relation type. This is mostly done manually.

5 Results and discussion

5.1 Quantitative assessment

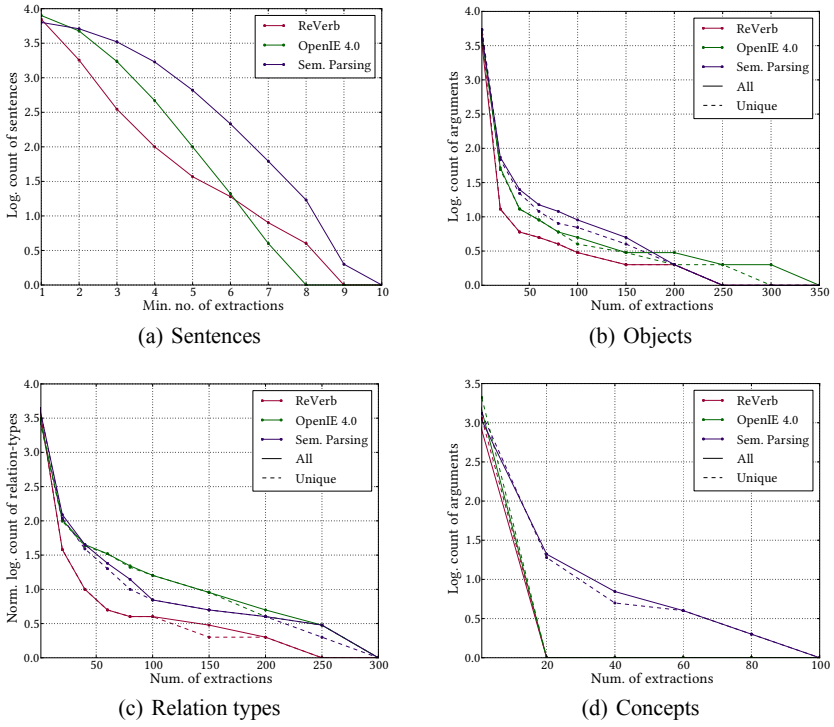
Figs. 2(a) and 3(a) show the distribution of the number of extracted assertions using each of the OIE approaches, across sentences in Carnatic and Hindustani music, respectively. Notice that the y-axis is a log scale. Between ReVerb and OpenIE 4.0, the latter seem to perform better, which can be attributed to the noun mediated relations. The semantic



(a) Sentences (b) Objects

(c) Relation types (d) Concepts

Fig. 2. Distribution of number of extracted assertions from each of the approaches for Carnatic music, across different aspects.



(a) Sentences (b) Objects

(c) Relation types (d) Concepts

Fig. 3. Distribution of number of extracted assertions from each of the approaches for Hindustani music, across different aspects.

parsing based system, however, retrieves substantially more relations per sentence than these two. A tight coupling between syntax and semantics proves to be advantageous in chunking different types of assertions, as well as relating entities far off each other in a given sentence. For instance, in sentences which feature a single subject, but multiple relations (e.g: x is a y, born in z to a and b.), it performed thoroughly well compared to others. The difference between their performances for Carnatic and Hindustani music is negligible, which shows that this result is consistent.

Figs. 2(b) and 3(b) show the corresponding distribution across objects. Recall that we defined an object to be the first argument of an assertion. The few objects with a disproportionately high number of extractions are usually the pronouns (despite resolving most of them), followed by musical terms. The semantic parsing based system retrieves slightly more number of assertions per object compared to OpenIE 4.0, which in turn performs better than ReVerb. We observed some redundancy in assertions for a given object, which is beneficial as this can be used as a measure of confidence in asserting the corresponding relation. In order to analyze this, we have also plotted the distributions with unique extractions across objects (shown in dashed lines in the figures). For

Carnatic music, we can observe that the semantic parsing based system retrieves substantially more number of redundant assertions per object compared to the other two. This difference, however, is less obvious in the case Hindustani music.

Figs. 2(c) and 3(c) show the distribution of the number of extracted assertions across the relation types. The results for semantic parsing based system and OpenIE 4.0 are more or less the same, both of which are substantially better than ReVerb. The redundancy in assertions, seen as the difference between the distributions shown by solid and dashed lines, is not as pronounced as it is for objects. The decline in the total number of relation-types as the number of extractions go higher, is less steep than in the case of objects. Unless the vocabulary in the domain is itself limited, this may indicate a slightly better coverage of relation types compared to that of the objects in the domain.

Figs. 2(d) and 3(d) show the distribution of the number of extracted assertions across the concepts. Recall that a concept is defined to be the second argument of an assertion. The difference between the semantic parsing based system and the other two is quite marked, both for Carnatic and Hindustani music, with the former retrieving more assertions per concept. For Hindustani music, the coverage of concepts in the assertions of ReVerb and OpenIE 4.0 is very low, with no concept having more than 20 assertions.

To summarize, the results indicate that the semantic parsing based system has a better coverage of objects, concepts and relation types of the domain than OpenIE 4.0, which is followed by ReVerb. It also retrieves more assertions per sentence compared to the other two. Note that this section has only provided the quantitative information, which by no means is complete in itself. The results we discuss in the following section complement these providing qualitative observations along the same dimensions (i.e., objects, concepts and relation types).

5.2 Qualitative assessment

In this section, we present the results for various tasks in the ontologization of Indian art music domain: concept identification, object identification and semantic relation extraction.

Concept identification. Recall that we define candidate concepts to be the collection of second arguments from the assertions of a given OIE system. We filter out those candidates which appear less than 5 times. We map the rest to the classes in the ontologies as described in sec. 4. Table. 2 shows the number of classes in the ontolgies for each music, and the number of concepts mapped from the assertions of the OIE systems.

Music	#Ontology	#ReVerb	#OpenIE 4.0	#Sem. Parsing
Carnatic	53	4	4	22
Hindustani	55	1	2	9

Table 2. The number of classes in the ontologies for each music, and the number of concepts mapped from the assertions of the OIE systems.

Object identification. The first subtask in this part is to find the entities in the domain. The candidate entities from each OIE system are defined to be the collection of first arguments from all its assertions. Table. 3 shows the total number of entities in the reference data taken from Wikipedia for each music, and the number of entities in the intersection of these with the candidate entities of each OIE system. There is no marked difference between the results, with nearly all the systems having about 60% of the entities from reference data in their assertions. Further, these correspond to only about 7% of all the candidate entities.

Music	#Reference	#ReVerb	#OpenIE 4.0	#Sem. Parsing
Carnatic	618	349	364	364
Hindustani	697	396	410	399

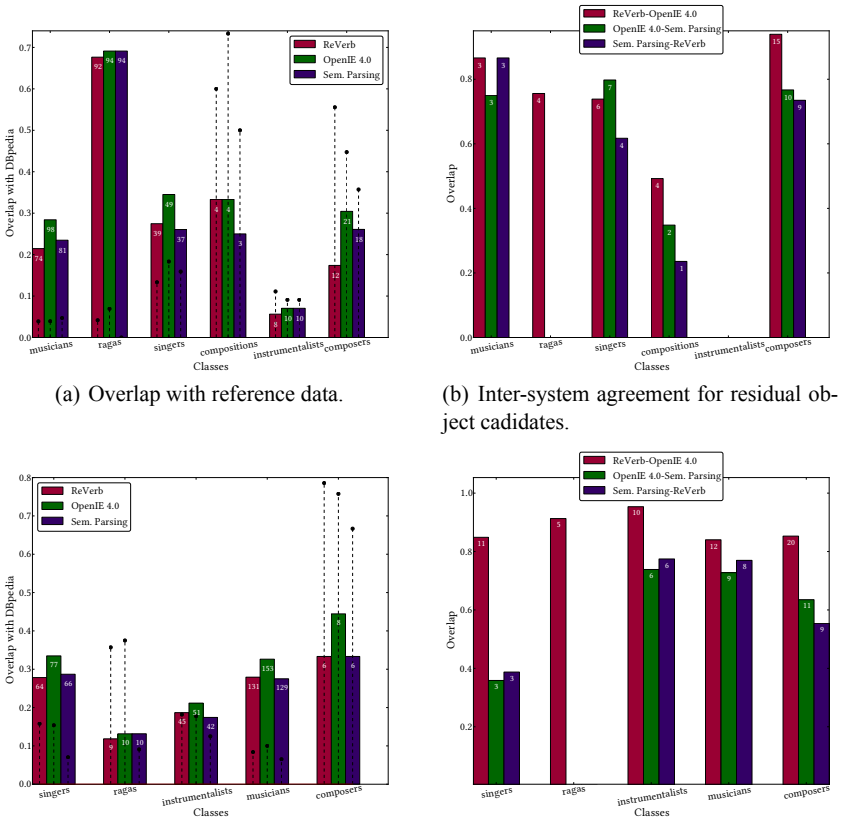
Table 3. The number of entities in the reference data for each music, and those identified using the OIE systems.

For the second subtask of ~~object identification,~~ i.e., assigning objects to semantic categories, we have considered those concepts from our ontology for which there is a corresponding category on Wikipedia, each having at least 20 pages. This was done to avoid manual labeling of objects. We found 5 such semantic categories for Hindustani music: musicians, singers, instrumentalists, composers and ragas[7]. For Carnatic music, in addition to these, we have found another semantic category: compositions. As discussed, we evaluate this task using two methods: rule-based and bootstrapping-based method.

Figs. 4(a) and 4(c) show the overlap (O, see eq. 1) on rule-based semantic category assignment for objects found in Carnatic and Hindustani music using the OIE systems. The stem plots in the figures show residual portion of objects (R). The most notable performances are seen for the raga category in Carnatic music. This can be attributed to two specific reasons: most Carnatic ragas on Wikipedia are described using a template, and the terminology consists mainly of Sanskrit terms which set them apart from the rest (mainly people). On the other hand, the description for Hindustani ragas varied a lot from raga to raga, and often the Sanskrit terms are inconsistently romanized making it hard for OIE systems to retrieve meaningful assertions. In theory, there is a template for almost every category, but there is a lot of variability, such as this, in describing the corresponding objects, except in the case of Carnatic ragas. OpenIE 4.0 seems to perform slightly better in terms of overlap, compared to the semantic parsing based system and ReVerb.

It is noteworthy to observe that residual object candidates are consistently less in number for the semantic parsing based system. There are two possibilities with them: they can be either false positives, or true positives which are not found in the reference data. In most cases, they are false positives. However, there are also a few of the latter. In order to understand them further, we have plotted the inter-system agreement in

[7] Raga is the melodic framework for both the Indian art music traditions.



(a) Overlap with reference data.

(b) Inter-system agreement for residual object cadidates.

(c) Overlap with reference data.

(d) Inter-system agreement for residual object cadidates.

Fig. 4. Results for rule-based semantic category assignment of objects identified in Carnatic (top) and Hindustani (bottom) music.

figs. 4(b) and 4(d), which is given by the cosine similarity between R of different approaches. ReVerb and OpenIE 4.0 agree with each other consistently higher over many of the semantic categories. We have observed that the cases where two or more systems agree on the candidature of a given object, it is highly probable that the object actually belongs to the semantic category. All the figures also show absolute numbers to put into perspective the proportion of residual objects where the systems agree with each other.

The second method for the evaluation of semantic category assignment employs bootstrapping as discussed in sec. 4. This process involves selection of a seedset and determining the number of bootstrapping iterations. For the sake of brevity, we have set the size of seedset to be the same for all the semantic categories, which is 3. The

objects in the seedset are randomly chosen from the ones among the reference data taken from Wikipedia. However, as the bootstrapping process itself can be sensitive to the initial selection of the objects in the seedset, the whole process is repeated 5 times with randomly chosen seedsets. The bootstrapping method is terminated once the size of seedset reaches that of the corresponding category in the reference data. After every 5 instances added during the process, we measure the overlap (O) and residual (R) portions of the seedset with respect to the reference data. Figs. 5.2 and 5.2 show their mean over 5 runs.

In most categories and for all the three systems, it can be seen that R grows quickly over iterations, making the residual portion the majority among the candidate objects, which brings the precision down. The semantic parsing based system consistently outperforms the other two methods, both in terms of having higher O, and lower R. Between ReVerb and OpenIE 4.0, there is no substantial difference in terms of O. For Carnatic singer and instrumentalist categories, however, the latter results in a lower R, and a slightly higher O compared to the former.

Interesting. It must be due to many sparse features in Reverb and IE 4.0, In semantic parsing features are coarse and not sparse.

Semantic relation extraction.

6 Related work

7 Conclusions

References

[1] Etzioni, O., Banko, M.: Open Information Extraction from the Web. Commun. ACM **51**(12) (2008) 68--74

[2] Sarawagi, S.: Information Extraction. Found. Trends Databases **1**(3) (2008) 261--377

[3] Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Empir. Methods Nat. Lang. Process. (2011)

[4] Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open Language Learning for Information Extraction. In: Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. (2012)

[5] Wu, F., Weld, D.: Open information extraction using Wikipedia. In: Assoc. Comput. Linguist. Number July (2010) 118--127

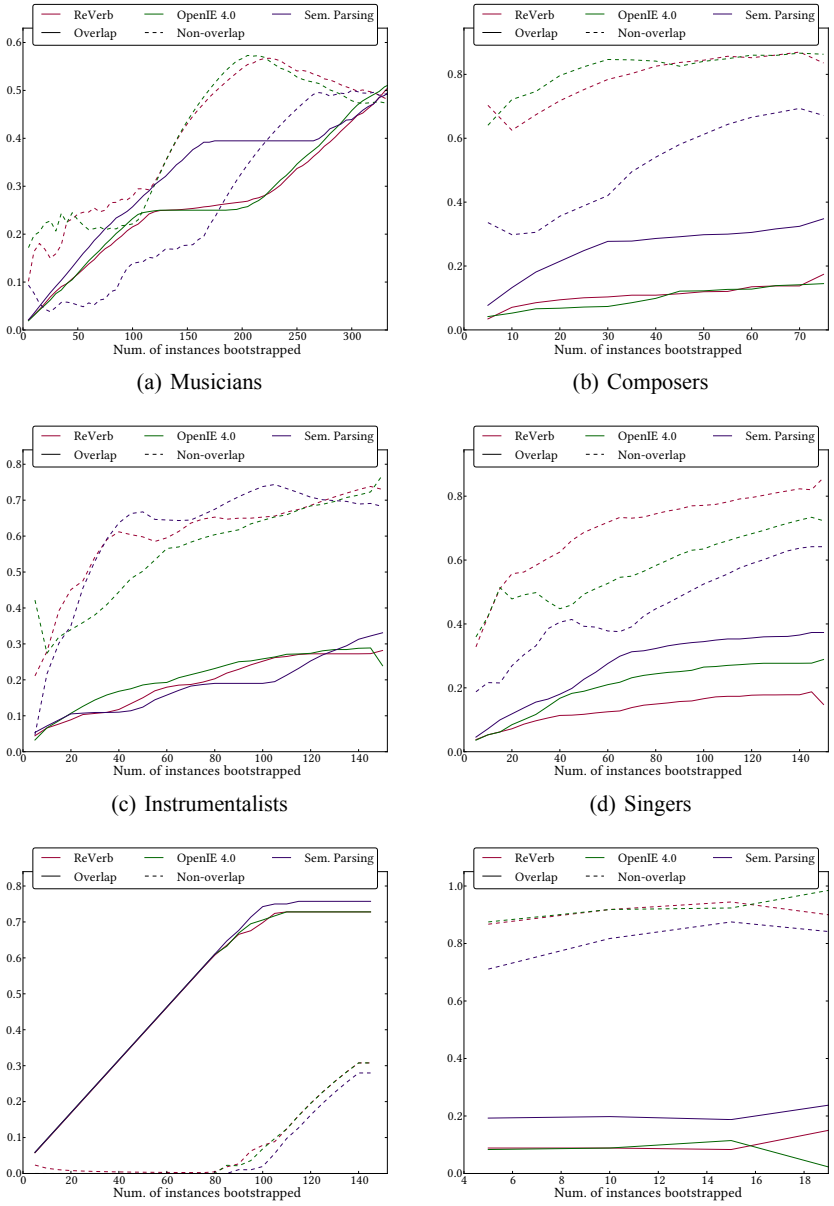
[6] Harrington, B., Clark, S.: Asknet: Automated semantic knowledge network. In: AAAI. (2007) 889--894

[7] Steedman, M.: The Syntactic Process. MIT Press, Cambridge, MA, USA (2000)

[8] Bos, J., Clark, S., Steedman, M., Curran, J.R., Hockenmaier, J.: Wide-coverage semantic representations from a CCG parser. In: COLING, Morristown, NJ, USA, Association for Computational Linguistics (2004) 1240--1246

[9] Serra, X.: A Multicultural Approach in Music Information Research. In: ISMIR. (2011) 151--156

[10] Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., Widmer, G.: Roadmap for Music Information Research. (2013)
[11] Krishna, T.M., Ishwar, V.: Karṇāṭik Music : Svara, Gamaka, Phraseology And Rāga Identity. In: 2nd CompMusic Work. (2012) 12--18
[12] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Comput. Linguist. **39**(4) (December 2013) 885--916
[13] Soderland, S., Roof, B., Qin, B., Xu, S., Etzioni, O.: Adapting open information extraction to domain-specific relations. AI Mag. (2010) 93--102
[14] Petasis, G., Karkaletsis, V.: Ontology population and enrichment: State of the art. In: Knowledge-driven Multimed. Inf. Extr. Ontol. Evol. Springer-Verlag (2011) 134--166
[15] Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Work. New Challenges NLP Fram. Lr., Valletta, Malta, ELRA (May 2010) 45--50

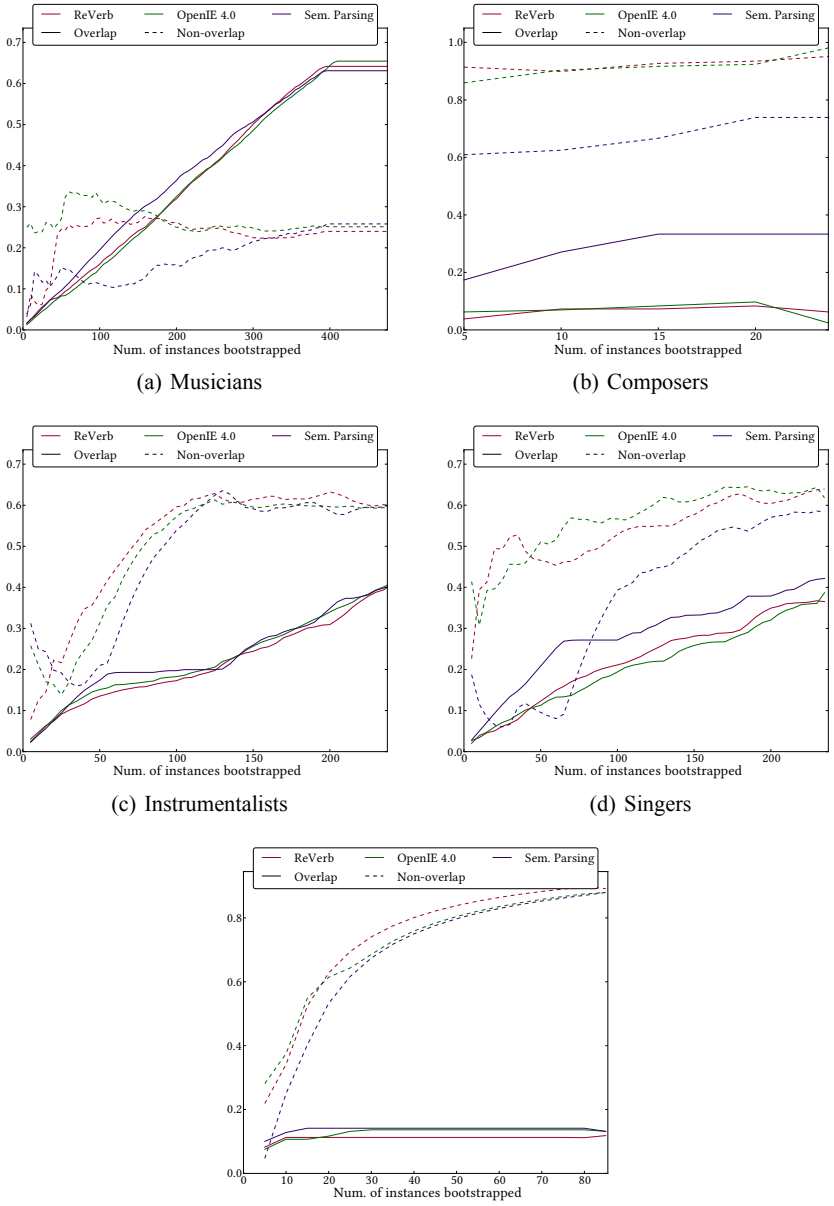


(a) Musicians (b) Composers

(c) Instrumentalists (d) Singers

(e) Ragas (f) Compositions

Fig. 5. Results for bootstrapping-based semantic category assignment of objects identified in Carnatic music



(a) Musicians

(b) Composers

(c) Instrumentalists

(d) Singers

(e) Ragas

Fig. 6. Results for bootstrapping-based semantic category assignment of objects identified in Hindustani music