SUMMARY
Entity descriptions have been exponentially growing in community-generated knowledge databases, such as DBpedia. However, many of those descriptions are not useful for identifying the underlying characteristics of their corresponding entities because semantically redundant facts or triples are included in the descriptions that represent the connections between entities without any semantic properties. Entity summarization is applied to filter out such non-informative triples and meaning-redundant triples and rank the remaining informative facts within the size of the triples for summarization. This study proposes an entity summarization approach based on pre-grouping the entities that share a set of attributes that can be used to characterize the entities we want to summarize. Entities are first grouped according to projected multilingual categories that provide the multi-angled semantics of each entity into a single entity space. Key facts about the entity are then determined through in-group-based rankings. As a result, our proposed approach produced summary information of significant better quality ($p$-value $= 1.52 \times 10^{-3}$ and $2.01 \times 10^{-3}$ for the top-10 and -5 summaries, respectively) than the state-of-the-art method that requires additional external resources.

**key words:** entity summarization, DBpedia, ranking, multilingual joint space, entity grouping

1. Introduction

The rapid increase in the number of triples (i.e., facts) in knowledge bases (KB) has made it imperative to extract essential information from many relevant and similar facts that describe an entity comprising a set of entity–property–value triples. Consider that DBpedia [1] version 2015-04 provides localized editions in 128 languages. This released version contains approximately 38.3 million entities described by 6.9 billion Resource Description Framework (RDF) triples. Each entity is described in an approximate average of 180 triples. Therefore, entity summarization [2], which provides a smaller representative subset from a lengthy description, has attracted much attention in recent years. Entity summarization is concerned with the generation of high-quality entity summaries. Although several approaches have been proposed in [2]–[5], their qualities are still far from ideal, and some approaches rely on external resources.

This study proposes an entity summarization method that uses entity grouping to identify the clusters of entities in a KB that share a set of attributes, which can be used to characterize the entities we want to summarize. KBs have many semantic groups of entities (i.e., DBpedia, such as “Person,” “Company,” and “Film”). For example, an entity group (“Usain Bolt”, “Carl Lewis”, “Michael Johnson”, “Babe Ruth”, “Hyun-jin Ryu”) shares a set of properties related to the “Sports Player” type. More precisely, this group can be subdivided into two groups as follows: “Athlete” = (“Usain Bolt”, “Carl Lewis”, “Michael Johnson”) and “Baseball Player” = (“Babe Ruth”, “Hyun-jin Ryu”). Each type has distinguishing characteristics that can reveal the underlying facts generating the entity summaries. Consider the difference of the typical player characteristics between “Usain Bolt” and “Babe Ruth.” “Usain Bolt” has key facts, such as “sport event” or “medal information,” while “Babe Ruth” has more emphasis on his “position” or “team.”

Consider why a new entity grouping is necessary even if a KB, such as DBpedia, has its own mechanisms for setting the types of each entity. First, its coverage of the type of entities is not sufficient. Moreover, the types for each entity, if they exist, are not stable enough to make a summary for its entity’s description because of the mismatches between the defined type information and the actual entity descriptions. An investigation on DBpedia in the English edition showed that 1/3 of the entities have no type. The English DBpedia has a total of 5.9 million resources as entities, out of which only four million entities are classified by their type. The properties of the facts for each entity lack consistency even if two different entities are identified (i.e., “Hyun-jin Ryu” and “Babe Ruth”) as belonging to the same group (i.e., “Baseball Player”). For example, the property of the batting position “battingSide” for “Babe Ruth” is present in the DBpedia English edition, but there is no such information for “Hyun-jin Ryu.” However, the Korean edition of DBpedia has such information.

We propose a new category-based clustering of entities to overcome the insufficient type in DBpedia. Entity grouping is particularly necessary for clustering entities in their multilingual projection to ensure that no relevant descriptions are biased toward one specific language edition. Category information will be used as a more stable source of clustering entities in the multilingual joint space while observing the unstable manifestation of the entity types in KB’s triples. A multilingual joint space model has the advantage of overcoming the missing categories caused by biases in different languages and can contribute to the creation of more stable and well-formed entity groups than the monolingual condition within it.

Our proposed approach, which consists of an entity
summarization based on an entity grouping in a multilingual projected entity space (Multi-EGS), comprises the following steps: first, we mine all language categories from KBs (DBpedia language editions) to build information on distinct entity groups. Second, each fact is ranked based on the pertinent property–value pairs, which are consecutively called ‘features’, of the in-group. We show the re-ranking for individual entity summarization with correlations between values according to the different ranks of properties even within a single entity group. Finally, we iteratively choose highly ordered and less similar facts by adopting a fact-ranking system.

The remainder of this paper is organized as follows. Section 2 introduces some related works. Section 3 describes the proposed entity grouping method and its application for summarizing entity descriptions. Section 4 illustrates the evaluation method and reports the experimental results. Section 5 discuss the detailed result analysis. Finally, conclusions and the plans for future work are detailed in Sect. 6.

2. Related Works

Entity summarization is studied for its ability to generate a summary of an entity description for a specific task (e.g., browsing and searching [6], [7]) or for generic purposes. Entity summarization is a specific field within the Linked Data research community, which researches on the problem of ‘features ranking (with features denoting Linked Data property–value pairs). Entity summarization deals with selecting features that are most interesting to present to a user.

RELIN [2] is a relatedness- and informativeness-based entity summarization model, which is among the initial works on entity summarization in the literature. RELIN provides a limited-sized summary with the objective of selecting the distinctive information that can identify an entity. In RELIN, the PageRank algorithm [8] is generalized to rank features based on both their relatedness and informativeness. However, the PageRank algorithm measures the relative importance of nodes (entities) according to their connections, making it difficult to determine the necessary information that each entity must summarize.

At the time of this writing, FACeted Entity Summarization (FACES) [5], as a current state-of-the-art method, aims to provide diversity along with the important characteristics of an entity. In FACES, an entity is usually described using a conceptually different set of facts, called facets, to improve the coverage of its summarization. The features within each facet are similar compared to those between facets. For example, the features expressed through the “KnownFor” and “Field” properties are conceptually similar because they both talk about an entity’s professional life. However, features that have the “Field” and “birthPlace” properties are conceptually dissimilar because these represent completely disparate information. FACES provides a faceted summary of a given length by combining at least one feature from each facet. However, several important facts for a summary may be present in one facet. Thus, a summary in each facet unit is not always ideal. Moreover, FACES expands each feature to obtain a set of words that rely on the external resource, WordNet [9] (i.e., hyponyms), in such a way that expanded words can be used to glean higher-level abstract meaning. WordNet is a widely used lexical database of the English language, which can be thought of as a machine readable dictionary. WordNet is useful in finding the semantic similarity between words using its underlying hyponym–hyperonym and meronym–holonym relations. However, WordNet does not always cover concepts in the KB, particularly for relatively less popular concepts in English. For example, “Busan” is South Korea’s second largest city after “Seoul,” but is not indicated as such in WordNet.

Thus, “Busan” cannot be expanded as a “place” or “area” by the method used in FACES. We propose herein a method that does not use external resources other than the KB to be summarized.

Thalhammer et al. [3] proposed an approach that leverages usage data (rating data) to summarize entities in the movie domain in a Linked Open Data (LOD) space [10]. They utilized rating data to support the measurement of the similarities between movie entities. They also identified a set of nearest neighbors for each entity. The neighborhood formation of each entity was based on the log-likelihood ratio scoring with the following four parameters: the number of users who rated both entities; the number of users who rated the first, but not the second entity, and vice versa; and the number of users who rated neither of the two entities. They counted the number of entities, which had the same feature in their nearest neighbors group, for each feature of the entity and selected the top-n features as the summarization for each entity. The intuition of this work is very similar to the intuition of our research because the features shared by the entity group members (neighborhoods) are considered more important for their identity than the features they share with an entity not in their respective neighborhood. However, the neighborhood formation in their work, which is a critical step, solely depends on the rating data. The “user rating data” does not describe the essence of the entity. In addition, a user rating is not possible for many entities, which makes it different from the proposed entity-specific characteristic summarization method herein.

3. Proposed Approach

In this section, we will discuss how to summarize the facts for a given entity by identifying the essential characteristics for it to belong to an entity group and adjusting the order of facts within an identified entity group. Table 1 lists all the notations used in this paper. We define “entity summarization” in this work as follows:

Definition 1: An entity summarization is a process used to identify $\text{Summ}(e_i) \subset \text{ED}_{e_i}$, which contains the selected top facts with the highest scores corresponding to the summary of information concerning a given entity $e_i$, where the length
First, in the preprocessing part, we mine all the language categories \( \mathcal{B} \) to build information on distinct entity groups. The intuition of this study is that an important attribute among entities within the same group would be essential for finding an important factual triple of individual entity units and use it in fact summaries. In other words, the entities in the same group have distinct semantic signatures that enable us to determine whether the facts shared among the members of the group are more relevant than those shared with the entities not in the same group. We utilized the category information to infer the categorizable characteristics of entities and build the entities’ fine-grained semantic group. We particularly projected the different languages’ biased category information into a single joint space that could help overcome the missing categories with the rather biased angle of the language community and detect highly informative keywords for a more stable entity grouping. The categories in the multilingual editions of DBpedia act as a KB source supplement to the complementary characteristics of a given entity. For example, a comparison of the DBpedia English and Italian editions of categories for “Usain Bolt” in Tables 2 and 3 shows that the numbers of the categories used in the two different languages are obviously different. Moreover, several categories are only in one monolingual edition. “Jamaican Roman Catholics” is only in the English edition, while “Natiil 21 agosto” (“born on August 21st”)
is only in the Italian edition. We expect that the projection among different languages' biased category information can help overcome the problem of missing categories by taking a rather biased angle for the language community. The joint space model for the multilingual category projection will then infer the signature to build fine-grained semantic entity groups that fulfill our purposes when used in entity summarization.

This step induces a set of disjoint clusters, where each entity in B is categorized into a cluster (that represents an entity group) by executing a clustering process over the multilingual projected entity space generated by weaving different categories from several DBpedia language editions. The category information is accessed with the dct:subject property, and each category of a language version is linked to another language's category using sameAs-link in DBpedia without any burden of the multilingual alignment of tokens in categories. The vector space of keywords from the categories will be used to identify the characteristics of an entity, like "sprinters," "athletes," and "births" for a given entity "Usain Bolt," as in Table 2. We employed the k-means algorithm to accomplish this step because it is regarded as one of the simplest and most efficient unsupervised learning algorithms for clustering large datasets [11]. The k for k-means is derived from assigned types in B and the vector space of keywords from the categories as features.

3.2 Summary Generation

In the second part of the summarization step, each fact, after the entity group is completed, is ranked according to our proposed in-group triple ranking, which is applied to select the pertinent "properties" and "values" from each group. The working principle behind the fact ranking is that we assign a higher score to the facts containing more relevant properties with high frequencies to reflect the importance of a property to a group. More relevant values have higher correlations between the two nodes for a given fact. Hence, the score of a fact \( f_{p,e} \) is defined as follows:

\[
\text{score}(f_{p,e}) = p\text{-score}(e, p) + v\text{-score}(e, v) + \lambda(p\text{-score}(e, p) \times v\text{-score}(e, v)),
\]

where \( p\text{-score}(e, p) \) is a weight assigned to the property \( p \) for the group of \( e \); \( v\text{-score}(e, v) \) is a correlation weight assigned to the value \( v \) for the entity \( e \), and \( \lambda \) is a tuning parameter that determines the ratio of the synergy indicators. A more detailed description is included in the section that follows.

3.2.1 Property Scoring (p_score)

The property-scoring scheme for the proposed fact-scoring formula is derived using a property-weighting function that obtains the properties interacting most strongly in the in-group space for the frequencies of the labels of the edges. The rationale behind this is that the more frequently a property appears in a group, the more important the property must be to that group. Such a property is less important when it occurs across groups. This scheme is based on the label of an in-group edge specifically influenced by the TF-IDF [12] technique to obtain the top labels of the edge (properties) from each group. The TF-IDF method determines the relative frequency of words in a specific document through an inverse proportion of the words over the entire document corpus [13]. A set of properties \( P(e_i) \) and the governing properties of \( \mathcal{G}^{e} \) may not have a membership relationship even if an entity \( e_i \) belongs to an entity group \( \mathcal{G}^{e} \). For example, "Usain Bolt" and "Michael Johnson" are in the same entity group "Athlete," but "Usain Bolt" has a property "religion," and "Michael Johnson" does not. From the perspective of the "Athlete" entity group, the "religion" property is not proper, but important for describing "Usain Bolt." This is handled by yet another value scoring scheme that can be captured based on how frequently two different entities (i.e., one is \( e \), and the other is \( v \)) are mentioned across the groups. Thus, the essential facts for "Usain Bolt" include some properties that maintain the membership of the entity group "Athlete." Some properties are seen as important to the individual that may be in a different facet from "Athlete" if they are frequently mentioned. The facts that are most relevant and less similar to the other facts are included in the summary. We iteratively select the facts that are ranked highest and less similar to the previously selected ones until we reach the length limit \( \sigma \). The \( p\text{-score} \) is computed as follows:

\[
p\text{-score}(e, p_1) = \| (s, p, o) \in \bigcup_{i \in \mathcal{G}^{e}} \mathcal{E}\mathcal{D}_i \mid p = p_1 \| \times \log\left(\frac{\| \mathcal{G} \|}{\| \mathcal{G}_{p_1} \|}\right),
\]

where \( p_1 \) denotes the property, \( \mathcal{G}^{e} \) is the group that belongs to \( e \), \( \mathcal{G} \) indicates the total groups, \( \mathcal{G}_{p_1} \) indicates the groups containing \( p_1 \), and \( \| x \| \) represents the total number of \( x \). This calculation intuitively determines the relevance of a given property in a particular group.

3.2.2 Value Scoring (v_score)

We determine the significance among properties and acquire the relevance among values to determine which facts are the most essential. Let a star-graph \( \mathcal{E}\mathcal{D} \) exist that contains different lengths for each edge exist, indicating the relatedness of two nodes connected by an edge. Note that the labels of the edges can be regarded as the coordinates of the graph nodes. The length of an edge indicates a low or high score in cases where the entity and the value are more independent or highly correlated. A correlation measure is used in the case of two related nodes. We assume that two entities are more highly correlated when the fraction of facts that are in common with the total number of facts of both entities is higher. \( v\text{-score} \) is computed as follows:
The first two elements connected by the addition operation are the coefficient-based scores of the facts of entities $e$ as a subject and $v$ as an object value. Each coefficient is normalized to the $[0-1]$ range and calculated internally in groups. The scores for the object entities associated with the other groups are not reflected. The third element, which is a multiplication operation, represents the influence of the object entity $v$ within a group.

3.2.3 Redundancy Checking

After obtaining the fact-ranking result, we focus on generating a summary of the fact collection by considering both relevance and anti-redundancy until a given length of summary is reached. Naturally, a KB may contain redundant information. For instance, the same property may occur multiple times and an identical entity as a value may be in-

Algorithm 1 Entity Summary Generation.

Input:
1) a knowledge base $G$ that consists of a set of facts for entities \( \{E_{D_1}, \ldots, E_{D_N}\} \), where $N$ is the total number of entities in $G$
2) given entity to be summarized $e_i$,
3) summary size $\sigma$,
4) tradeoff parameter depending on limitation summary size, $\theta$;

Output: Summary ($S$)

1: $E_{D_i} \leftarrow \{\}$ \quad \triangleright\ \text{an ordered list of all of facts of $e_i$ as entity}$
2: $S \leftarrow \emptyset$;
3: $SP \leftarrow \emptyset$; \quad \triangleright\ \text{a set of properties of the summary $S$}$
4: $SV \leftarrow \emptyset$; \quad \triangleright\ \text{a set of values of the summary $S$}$
5: for each $f_{p,v}$ in $E_{D_i}$ do
6: \hspace{1em} summary_insertable = True;
7: \hspace{1em} if $\text{sim}(p, \forall x \in SP) \leq \theta$ then
8: \hspace{2em} summary_insertable = False;
9: \hspace{1em} end if
10: \hspace{1em} if $\text{sim}(v, \forall y \in SV) \leq \theta$ then
11: \hspace{2em} summary_insertable = False;
12: \hspace{1em} end if
13: \hspace{1em} while $|S| \leq \sigma$ do
14: \hspace{2em} if summary_insertable: then
15: \hspace{3em} $SP = SP \cup p$
16: \hspace{3em} $SV = SV \cup v$
17: \hspace{2em} $S = S \cup f_{p,v}$
18: \hspace{1em} end if
19: \hspace{1em} end while
20: \hspace{1em} end for
21: \hspace{1em} return $S$

4. Experiments and Results

This section describes the experiments conducted to empirically test the proposed entity group-based summarization approach.

4.1 Experiment Setting

The DBpedia dataset is used for the evaluation because it is the benchmark dataset selected in [2], [5] and contains numerous entities that belong to different domains. The same version of the DBpedia dataset (3.9) is used as an input KB to assess the effectiveness of our method for entity summarization against FACES [5], which is the current state-of-the-art method. We also utilize the ten largest languages of DBpedia (i.e., English, French, German, Italian, Spanish, Russian, Dutch, Polish, Portuguese, and Swedish editions) to project multilingual category information into a single space that provides the integrated multi-angled semantics of each entity.

The category label is first tokenized for entity grouping, and stop words are removed. The words are stemmed using Porter’s stemming algorithm [15] to reduce them to their stem forms. All the category labels of the entities are represented as vector stem words. The category labels marked in a different language are translated into English through the built-in OWL property owl:sameAs link in the linked data [16]. The owl:sameAs statement indicates that two Uniform Resource Identifier (URI) references refer to the same thing: the entities have the same identity. For example, we could state that two category URIs connected by an owl:sameAs link actually refer to the same reference. We use it to translate the “1986 年生” category (in the Japanese edition) into “1986_births” in English through the following RDF statement:

\[
\begin{align*}
s & : \text{<http://dbpedia.org/resource/Category:1975_births>} \\
p & : \text{<http://www.w3.org/2002/07/owl#sameAs>} \\
o & : \text{<http://ja.dbpedia.org/resource/Category:1975 年生>}
\end{align*}
\]

We assume herein that the number and the type of defined groups correspond with the assigned types in the KB, meaning that the entities with an unknown type can be mapped onto existing predefined types rather than requiring the definition of a new type. Therefore, a clustering result
was obtained using the \( k \)-means algorithm with the number of clusters \( k \) set to the number of known types \( (k = 322) \) in the DBpedia.

DBpedia has two types of properties: object and data-type. Data-type properties relate individuals to literal data (e.g., strings, numbers, dates, etc.), whereas object properties relate individuals to other individuals. The authors of FACES shared their gold-standard entity summaries, which were given by a group of human experts, that consisted of five and 10 facts (at least seven human-generated summaries) targeting only object properties for each of the selected 50 entities of DBpedia. These summaries are referred to herein as ideal summaries.

4.2 Result Analysis

The evaluations of the summarization systems use an ideal summary provided by multiple human annotators by counting the unit overlaps with the generated summary regarded as the quality. Similar to those in RELIN [2] and FACES [5], we also use the same quality metric, such as in Eq. (4), where \( n \) is the number of human annotators required to produce the individual ideal summaries denoted by \( Summ_i^I(e) \) for \( i = 1, \ldots, n \), and the automatically generated summary is denoted by \( Summ(e) \) for entity \( e \). The summary that achieves the highest quality score is considered as the most similar to the ideal summary. The criterion for choosing an important triple in making the gold standard is different for each participant. Hence, the value of the user agreement should be clarified. Given \( \sigma \in \{5, 10\} \), entity \( e \) and \( n \) ideal summaries received, their agreement is defined by their average overlap, as described by Eq. (5) [2]. The agreements averaged over all entities are 1.9596 and 4.6770 for \( \sigma = 5 \) and 10, respectively.

\[
\text{Quality}(Summ(e)) = \frac{1}{n} \sum_{i=1}^{n} |Summ(e) \cap Summ_i^I(e)| \tag{4}
\]

\[
\text{Agreement} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |Summ_i^I(e) \cap Summ_j^I(e)| \tag{5}
\]

Table 4 shows the performance evaluation results of Multi-EGS compared to FACES. We considered several baselines to analyze the effectiveness of the entity group-based approaches. The simplest baseline is building a group of entities by utilizing the assigned entity types in KB (Typed). Another baseline considered is building entity groups using monolingual categories (EGS). The table clearly shows that our group-based summarization approach outperforms FACES in terms of the summarization quality. A two-tailed paired t-test is also performed to verify the statistical significance of the performance improvement. The respective \( p \)-values for Multi-EGS against FACES for the top five and 10 lists were 0.02013 and 0.00152, respectively. Thus, our approach provides significantly better results than FACES.

We observed that the emphasis on the redundancy removal is sensitive to the desirable length of a summary by testing different \( \theta \) values. The \( \theta \) value can vary from 0 to 1 and is calculated using the simple sequence-matching method for the strings that appear in the properties and the values that make up the triple, as described in Sect. 3.2.3. The fewer the duplicate strings in the two strings compared, the closer the score is to 0. The ideal entity summary for a short summary consisting of five triples does not contain similar facts, such as “\( \lambda \) Usain Bolt, birthPlace, Jamaica”, and “\( \lambda \) Usain Bolt, residence, Jamaica”. Note that these two facts have a word overlap of “Jamaica”. Such information is acceptable for longer summaries.

5. Discussions

This section contains a more detailed result. Parameter \( \lambda \) in the equation used to calculate the importance of triples is analyzed in the first analysis herein. To recall, parameter \( \lambda \) denotes the strength of the interaction between the property and value scorings, which constitute a triple. Figure 2 shows the correlation of the quality score of the overall summary system with the changes in parameter \( \lambda \). This analysis indicates that our system has a better performance than the state-of-the-art in all settings, particularly the strong peak at 4.5. Note that the convergence is found when parameter \( \lambda \) is set to 5 or higher. The property and the value of a single fact are considered together in the ranking computation to obtain an improved result.

A second detailed analysis is then conducted in the case of summarizing the best quality through the proposed method and the case of summarizing that with the lowest quality. Table 5 shows the summary results of the highest and lowest qualities of the proposed method (Multi-EGS) and the summary results of the existing methods (FACES).

![Fig. 2 Analysis of different parameters \( \lambda \).](image)
Table 5  Comparison of the top-five summary results.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Size(stemmed)</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual (EGS) C</td>
<td>44,242</td>
<td>0.6512</td>
</tr>
<tr>
<td>Monolingual (EGS) C+</td>
<td>44,801</td>
<td>0.6508</td>
</tr>
<tr>
<td>2-langs merger: C</td>
<td>44,686</td>
<td>0.6601</td>
</tr>
<tr>
<td>3-langs merger: C</td>
<td>44,991</td>
<td>0.6639</td>
</tr>
<tr>
<td>4-langs merger: C</td>
<td>45,093</td>
<td>0.6659</td>
</tr>
<tr>
<td>5-langs merger: C</td>
<td>45,259</td>
<td>0.6751</td>
</tr>
<tr>
<td>6-langs merger: C</td>
<td>45,445</td>
<td>0.6796</td>
</tr>
<tr>
<td>7-langs merger: C</td>
<td>45,501</td>
<td>0.6734</td>
</tr>
<tr>
<td>8-langs merger: C</td>
<td>45,575</td>
<td>0.6744</td>
</tr>
<tr>
<td>9-langs merger: C</td>
<td>45,658</td>
<td>0.6802</td>
</tr>
<tr>
<td>10-langs merger (Multi-EGS): C</td>
<td>45,732</td>
<td>0.6839</td>
</tr>
</tbody>
</table>

Table 6  Purity values for each feature type. C denotes using categories, while C+ denotes using upper categories.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Feature Size(stemmed)</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Biden [3.0]</td>
<td>44,242</td>
<td>0.6512</td>
</tr>
<tr>
<td>&lt;party, Democratic Party (United States)&gt;</td>
<td>44,801</td>
<td>0.6508</td>
</tr>
<tr>
<td>&lt;state, Delaware&gt;</td>
<td>44,686</td>
<td>0.6601</td>
</tr>
<tr>
<td>&lt;president, Barack Obama&gt;</td>
<td>44,991</td>
<td>0.6639</td>
</tr>
<tr>
<td>&lt;profession, Lawyer&gt;</td>
<td>45,093</td>
<td>0.6659</td>
</tr>
<tr>
<td>&lt;birthPlace, Pennsylvania&gt;</td>
<td>45,259</td>
<td>0.6751</td>
</tr>
<tr>
<td>&lt;termPeriod, Joe Biden 1&gt;</td>
<td>45,445</td>
<td>0.6796</td>
</tr>
<tr>
<td>&lt;religion, Catholic Church&gt;</td>
<td>45,501</td>
<td>0.6734</td>
</tr>
<tr>
<td>&lt;party, Democratic Party (United States)&gt;</td>
<td>45,575</td>
<td>0.6744</td>
</tr>
</tbody>
</table>

for the corresponding entities. “Joe Biden” denotes the case with the highest score for Multi-EGS during the evaluation (quality = 3.0). The quality of the ideal summaries of the two systems is similar. We can see that both systems provide a good summary of the entity. However, FACES still has a higher score (quality = 1.625) in the case that received the lowest score for Multi-EGS, which is “Total Recall (1990 film)”. FACES includes the fundamental value “Jerry Goldsmith” several times in the summary, thereby leading to increased overlaps between each of the ideal summaries. Many duplicate object values are unsuitable entities for summary information.

The third analysis contains a more detailed comparison between multilingual- and monolingual-based groupings. We analyze the effects of using a multilingual category projection for the entity grouping. The quality of the clustering result is evaluated using the purity metric. Purity is one of the primary evaluation metrics that indicate the proportion of correctly clustered samples [17]. Table 6 summarizes the evaluation results of clustering, where the type of information from DBpedia is the gold standard. The dataset used in this study contains approximately 44,000 stemmed tokens as features extracted from the categories used for clustering. These features increase the use of information using the upper-category (i.e., parent category, such as the parent “American singers” of child “American male singers”) structure from an explicit category, but fail to significantly affect the clustering result. Purity is observed to increase from 0.65 by applying EGS to 0.68 when applying Multi-EGS. However, the two clustering processes have an unbalanced number of features, which may have affected the results. Hence, we perform a random sampling analysis to determine the statistical significance of the tokens of clustering as follows: first, we select 10,000 random tokens per system (EGS and Multi-EGS) to partition our original tokens into small- and same-sized token sets between the two approaches (10,000 is approximately 20% of the tokens). We then execute k-means clustering using these features and compute the purity score for the clustering results of each system. The average purity score of 100 random sampling experiments for Multi-EGS (0.4777) is higher than that of EGS (0.4607). A statistical evaluation using a two-sample paired t-test shows a p-value equal to 2.28726×10^{-5}. Multi-EGS exhibits a 0.03% improvement for the summary quality compared to the EGS method for a top-five summary (Table 4).

Figure 3 shows an example of the comparison of the specifications of the multilingual and monolingual conditions of entity grouping. Many entities in one group, as obtained by entity grouping, can belong to different types. Grouping may sometimes be incorrect because of the inclusion of unexpected entities. Through the multilingual projected entity grouping (Fig. 3), the entities in “Usain Bolt”’s group belong to the following nine types: Athlete, Agent, Person, BasketballPlayer, CollegeCoach, AmericanFootballPlayer, OfficeHolder, Criminal, and Politician. These concepts represent an individual. While the monolingual group contains a vast number of types in great disorder, this monolingual grouping does not express the accurate characteristics of the entities compared to the multi-
lingual grouping. In other words, “Usain Bolt” is a popular entity across multiple languages, thereby implying that extensive categories can be found on this entity in various languages. Thus, multilingual grouping comprises a signature to describe the main features of an entity in a group. In addition, it can help entities that are hidden in the long tail in a monolingual space.

In the fourth analysis, we discuss the limitations of the proposed method. The proposed entity-grouping algorithm uses categories in DBpedia to classify the entities into several distinguished groups. The categories are generally intended to group together pages on similar subjects. However, Wikipedia articles are often categorized without specific policies. Hence, the proposed entity-grouping algorithm sometimes clusters the entities differently from what is expected. For example, three entities (i.e., “Joe Biden”, “Barack Obama”, and “Texas”) of the evaluation data have been generated as a single group, but we know that “Texas” and “Joe Biden” (or “Barack Obama”) are intuitively different types of entities. Following shows the top-five summary of Multi-EGS about “Texas”.

Even though no grouping can clearly represent one signature, it does not adversely affect the overall ranking because the important facts are determined by the adjustment of both the property and value ranking. However, we can expect better results if clustering can be done more effectively. In this study, we use a method of dividing entities without using external resources other than the target KB to perform entity summarization. However, for a more robust entity grouping, it is expected to be scalable to the external resources available in the linked data [10], such as YAGO [18], which is expected to help improve the performance of the overall entity summarization system.

Finally, we discuss how to use the proposed technology as an end-application. The Google Knowledge Graph (GKG) is one example, where the current entity summarization technology is actually being serviced. The GKG is a service that outputs summarized important knowledge about entities used in search engine queries in the form of a table in the upper right corner of the Google search result. However, GKG only outputs the values for predefined properties. For example, the date of birth, birthplace, height, and weight are generally included in the top-five information if the entity used as a search term is a person. Thus, distinguishing between the entities becomes difficult. Table 7 shows the comparisons of the Multi-EGS and GKG results on the essential information of two given entities, namely “Cristiano Ronaldo” and “Usain Bolt”. Suppose that you are given the question “Guess who” by reversing the summary information for a given entity. Based on the five facts provided, Multi-EGS can undoubtedly and more quickly determine a given subject in this case. We can perform a Multi-EGS on the whole KB to create a summary white paper of the entity, which could be used in kids’ learning.

### 6. Conclusions and Future Work

This study proposes an approach for configuring a summary within entity groups for the entities of a dataset in a knowledge base (i.e., DBpedia). The approach primarily comprises two steps: 1) entity group formation based on the multilingual joint space model followed by the selection of ingredients for the clustering algorithm from the merger information describing the entity semantic categorizable characteristics among different language datasets; 2) ranking of facts according to their scoring formula derived from the pertinent features of the in-group that combines the importance of properties and any correlations between the values in the facts; the essential facts for an entity complementarily include some properties that maintain the membership of the entity group and some properties that are important to the individual that may be in a different facet from the entity group; and selection of the ordered facts in the summary that includes the facts that are more relevant and less similar to the other facts.

Our proposed approach was evaluated against the state-of-the-art method. Consequently, the proposed approach improved the quality of the summary information compared to the user-created benchmark. The proposed approach improved the efficiency of the entity summarization system using only triples without utilizing external resources, such as

<table>
<thead>
<tr>
<th>Multilingual</th>
<th>Google Knowledge Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>Title</td>
<td>Title</td>
</tr>
<tr>
<td>Sport</td>
<td>Sport</td>
</tr>
<tr>
<td>Event</td>
<td>Event</td>
</tr>
<tr>
<td>Years</td>
<td>Years</td>
</tr>
<tr>
<td>Club</td>
<td>Club</td>
</tr>
<tr>
<td>Award</td>
<td>Award</td>
</tr>
</tbody>
</table>

### Table 7 Comparison of the Multi-EGS and Google Knowledge Graph results on the essential information of two given entities.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Team: Real Madrid C.F.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position: Forward (association football)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clubs: Sporting Clube de Portugal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BirthPlace: Madeira</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationalteam: Portugal national football team</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usain Bolt</th>
<th>Born: August 1, 1986 (age 30), Sherwood Content, Jamaica</th>
<th>Height: 195cm</th>
<th>Weight: 95kg</th>
<th>Club: Racers Track Club</th>
<th>Awards: Laureus World Sports Award for Sportsman of the Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthplace: Jamaica</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event: Sprint (running)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sport: Track and field</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title: Track &amp; Field News Athlete of the Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years: 2012 Summer Olympics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
WordNet. Moreover, the proposed entity summarization can extract a particular group’s stable signatures using multilingual projection without redundancies, thereby providing a useful strategy for identifying the nature of a described entity. This result would be of great interest in utilizing the diversity made available by the different languages to provide a more specific summary for each language.

The current approach calculates the fact ranking in-group. Therefore, creating a suitable summary is correspondingly simpler if a better clustering solution is produced. Various clustering algorithms, such as soft clustering and consensus clustering, will be examined in the near future. We plan to extend our approach to applications involving link prediction in cases where a new (unknown) entity is added to the group. The group-bound signatures are expected to identify the essential and critical facts for entities.

Acknowledgments

This work was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (2015M3A9A7029735), the Industrial Strategic technology development program (10072064, Development of Novel Artificial Intelligence Technologies To Assist Imaging Diagnosis of Pulmonary, Hepatic, and Cardiac Diseases and Their Integration into Commercial Clinical PACS Platforms) funded by the Ministry of Trade Industry and Energy (MI, Korea) and Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

References