MAGES: A Multilingual Angle-integrated Grouping-based Entity Summarization System

Eun-kyung Kim and Key-Sun Choi
Semantic Web Research Center
Korea Advanced Institute of Science and Technology (KAIST)
Republic of Korea
{kekeeo,kschoi}@world.kaist.ac.kr

Abstract

This demo presents MAGES (multilingual angle-integrated grouping-based entity summarization), an entity summarization system for a large knowledge base such as DBpedia based on a entity-group-bound ranking in a single integrated entity space across multiple language-specific editions. MAGES offers a multilingual angle-integrated space model, which has the advantage of overcoming missing semantic tags (i.e., categories) caused by biases in different language communities, and can contribute to the creation of entity groups that are well-formed and more stable than the monolingual condition within it. MAGES can help people quickly identify the essential points of the entities when they search or browse a large volume of entity-centric data. Evaluation results on the same experimental data demonstrate that our system produces a better summary compared with other representative DBpedia entity summarization methods.

1 Introduction

The rapid increase in the number of triples in knowledge bases (KBs) 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 (e.g., <Usain Bolt, nationality, Jamaican>, <Usain Bolt, birthPlace, Spanish Town>, <Usain Bolt, birthPlace, Jamaica>, <Usain Bolt, placeOfBirth, Jamaica>, <Usain Bolt, residence, Jamaica>, etc.). Therefore, entity summarization (Cheng et al., 2011), which creates a short summary from a set of triples from the description of an entity, has attracted much attention in recent years. This is a method designed to help people quickly identify the essential points of entities when searching or browsing a large volume of entity-centric data. Although several approaches have been proposed in (Cheng et al., 2011; Thalhammer and Rettinger, 2014; Gunaratna et al., 2015), their qualities are still far from ideal, and some approaches rely on external resources such as WordNet (Fellbaum, 1998).

This demo presents a multilingual angle-integrated grouping-based entity summarization system (MAGES), which is an entity summarization system for the DBpedia (Lehmann et al., 2014) based on the entity-group-bound ranking in an entity space. The intuition of this study is that property–value pairs—consecutively also called features—shared by an entity's group's members (neighborhoods) are considered more important for their identity than for the features they share with an entity that is not in their respective neighborhood. For example, there are two distinct groups: A = {"Usain Bolt", "Carl Lewis", "Michael Johnson"} and B = {"Babe Ruth", "Hyun-jin Ryu"}. Each group has distinguishing characteristics that can reveal underlying triples that generate entity summaries. Consider the difference between "Usain Bolt" in A and "Babe Ruth" in B for their typical player characteristics. "Usain Bolt" has essential properties such as "sport event" or "medal information," whereas "Babe Ruth" would have more emphasis on his "position" or "team."

There are many predefined semantic groups (i.e. types) of entities in DBpedia such as "Baseball Player," "Company," and "Film." However, although DBpedia has its own mechanisms for setting entity types, its coverage of the entity types is not sufficient. Moreover, the types of each entity, if they

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 203–207, Osaka, Japan, December 11-17 2016.

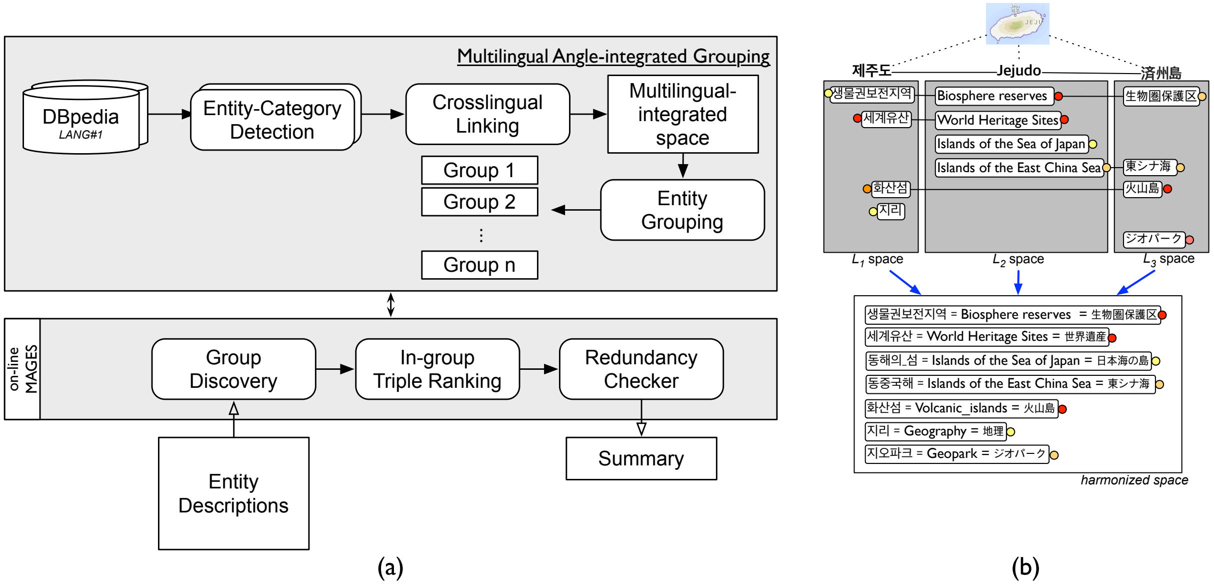
Multilingual Angle-integrated Grouping
DBpedia
LANG#1
Entity-Category Detection
Crosslingual Linking
Multilingual-integrated space
Group 1
Group 2
⋮
Group n
Entity Grouping
on-line MAGES
Group Discovery
In-group Triple Ranking
Redundancy Checker
Entity Descriptions
Summary
(a)
Jeju
제주
JEJU
제주도
Jejudo
済州島
생물권보전지역
세계유산
화산섬
지리
Biosphere reserves
World Heritage Sites
Islands of the Sea of Japan
Islands of the East China Sea
生物圏保護区
東シナ海
火山島
ジオパーク
L_1 space
L_2 space
L_3 space
생물권보전지역 = Biosphere reserves = 生物圏保護区
세계유산 = World Heritage Sites = 世界遺産
동해의_섬 = Islands of the Sea of Japan = 日本海の島
동중국해 = Islands of the East China Sea = 東シナ海
화산섬 = Volcanic_islands = 火山島
지리 = Geography = 地理
지오파크 = Geopark = ジオパーク
harmonized space
(b)

Figure 1: (a) The outline of the proposed multilingual angle-integrated grouping-based entity summarization system, (b) An example of multilingual topic integration: Collecting the scattered data (due to the different perspective) together to obtain the enriched knowledge of an entity in the DBpedia.

exist, are not stable enough to make a summary for the entity's description because of mismatches between defined type and actual entity descriptions. Therefore, new entity grouping is necessary, especially the grouping of entities in their multilingual angle integration to ensure that no relevant information is biased towards one specific language edition. DBpedia currently serves over 120 language editions extracted from language-specific Wikipedias that can contain different information from one language to another (Lehmann et al., 2014). In particular, language-specific editions can provide 1) more details about certain topics and 2) information missed in other DBpedia editions according to a specific cultural viewpoint. Unlike the prior studies of entity summarization, we particularly focus on methods for incorporating a variety of different lingual information scattered in Linked Open Data (Bizer et al., 2009) to enhance entities' topic detection. A multilingual angle-integrated space model can contribute to the creation of entity groups that are well-formed and more stable than the monolingual condition within it.

2 System Description

Figure 1 (a) shows the outline of the proposed system. It comprises the following steps: i) we mine all lingual category tags from multiple DBpedia language editions to create information about distinct entity groups, ii) Each feature is ranked based on the pertinent features of the in-group, and iii) We iteratively choose highly ordered and less similar facts by adopting a feature-ranking system.

2.1 Multilingual Angle-integrated Grouping

While observing the unstable manifestation of entity types in a KB's triples, category information will be used as a more stable source of clustering entities in a multilingual integrated space. We utilized the category tags to infer the topics of entities to build the entities' fine-grained semantic group. We integrated different languages' biased category tags into a single space that could help overcome missing categories and could help detect highly informative keywords for more stable entity grouping. For example, comparing the DBpedia Korean, English and Japanese editions of categories for "Jejudo" in Figure 1 (b), several categories are only in one monolingual edition: "Islands of the Sea of Japan" is only in the English edition and "지리[jili]" ("Geography") is only in the Korean edition.

Then, we induced a set of disjointed clusters in which each entity in DBpedia is categorized into a cluster (that represents an entity group) by executing a clustering process over the multilingual integrated entity space generated by weaving different category tags from several DBpedia language editions. The vector space of tokens from categories was used to identify the characteristics of an entity, such as "islands," "sea," and "japan," for a given entity "Jejudo" as in Figure 1 (b). We employed the *k*-means algorithm to accomplish this, because it is regarded as one of the simplest and most efficient unsupervised learning algorithms for clustering large data sets (J. A. Hartigan, 1979). The value of *k* for the *k*-means algorithm is determined by the number of types that exist in DBpedia.

2.2 In-Group Triple Ranking

In this step, all triples about each entity are ranked according to the in-group-relevance scoring formula. The working principle behind triple ranking is that we assign a higher score to triples that contain more relevant properties with high frequencies to reflect the importance of a property to a group, and more relevant values have higher correlations between two entities for a given triple. Hence, the score of a triple $t^e_{p,v}$ is defined as

$$\begin{aligned} score(t^e_{p,v}) &= p\_score(e,p) + v\_score(e,v) \\ &\quad + \lambda(p\_score(e,p) \times v\_score(e,v)), \end{aligned} \tag{1}$$

where $p\_score(e,p)$ is a weight assigned to the property p for the group of e, the $v\_score(e,v)$ is a correlation weight assigned to the value v for the entity e, and λ is a tuning parameter that determines the ratio of the synergy indicators. The p_score is derived by a property-weighting function that obtains the properties that interact most strongly in the in-group space for frequencies of labels of properties. This scheme is based on the label of an in-group property specifically influenced by the term frequency–inverse document frequency (TF-IDF) technique to obtain the top labels from each group. The v_score is derived by a correlation measure that is used in the case of two related entities; we assume that two entities are highly correlated when the fraction of triples that are in common with the total number of triples of both entities is higher.

2.3 Redundancy Checker

After the obtention of the triple ranking results, we focus on generating a summary of the triple collection by considering both relevance and anti-redundancy, until a given length of summary is reached. We attempt to iteratively measure the similarity of the next candidate triple to previously selected ones, and select a candidate if its similarity is below a threshold (user parameter) until the length limit of summary (σ) is reached. Given that a triple is much shorter than a sentence, most terms are specified within the KB. Therefore, a sequence matching procedure (Mount, 2004) provides the similarity measure among the words that appear in triples.

3 Experiments and Evaluation

We utilize the 10 largest languages in DBpedia—English, French, German, Italian, Spanish, Russian, Dutch, Polish, Portuguese, and Swedish—to project multilingual category information into a single space that provides integrated multi-angled semantics of each entity. All the category tags of the entities are tokenized and represented as vector stem words for entity grouping. Category tags marked in a different language are translated into English through the `owl:sameAs` link in the linked data. We assume that if two category tags are connected by means of this link, those categories can be considered to be the same.

As a current state-of-the-art method, FACeted Entity Summarization (FACES) (Gunaratna et al., 2015) aims to improve the coverage of its summarization using a conceptually different set of facts, called facets of an entity. The authors of FACES shared gold-standard entity summaries given by a group of human experts that consisted of 5 and 10 triples for each of the selected 50 entities in DBpedia. These are referred to as ideal summaries in our study.

Evaluations of the summarization systems use an ideal summary provided by multiple human annotators by counting the unit overlaps with the generated summary, which is regarded as the quality such as in Equation 2 (Cheng et al., 2011), 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 the entity e. The summary that achieves the highest quality score is considered to be the most similar to the ideal summary. Given $\sigma \in 5, 10$, an entity e and n ideal summaries received, their agreements (Cheng et al., 2011) averaged over all entities are 1.9596 and 4.6770 for σ = 5 and 10, respectively.

$$Quality(Summ(e)) = \frac{1}{n}\sum_{i=1}^{n} |Summ(e) \cap Summ_i^I(e)| \tag{2}$$

Table 1 shows the performance evaluation results of MAGES compared to FACES and other baselines. We considered several baselines to analyze the effectiveness of the entity group-based approaches. The simplest baseline was to build a group of entities utilizing the assigned entity types in KB (Typed). Another baseline that we considered was to build entity groups using monolingual categories (GES). It is clear from the Table 1 that our group-based summarization approach outperformed FACES in terms of the summarization quality. Moreover, a two-tailed paired t-test was performed to verify the statistical significance of the performance improvement. Significance was accepted at $p <$ 0.05. For the top-5 and top-10 lists, the respective p values for MAGES against FACES were 0.02013 and 0.00152. Thus, our approach provides significantly better results than FACES. FACES provides a faceted summary of a given length by incorporating 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 (e.g., hypernyms). However, WordNet does not always cover concepts in the KB, particularly relatively less popular concepts in English. For example, "Busan" is South Korea's second largest city after "Seoul," but the former is not indicated as such in WordNet. Thus "Busan" cannot be expanded as a "place" or "area" by the method used in FACES.

We also performed a random sampling analysis to verify the statistical significance of the integration of multiple lingual entity spaces, because an unbalanced number of tokens for clustering could affect the overall result. First, we selected 10,000 random tokens per system (GES and MAGES) to partition our original tokens into small- and same-sized token sets for the two approaches. Then, we executed clustering with these ingredients, and computed the Purity score (Amigó et al., 2009) for the clustering results of each system, in which the type information from DBpedia is gold standard. The average score of 100 random sampling experiments for MAGES (0.4777) was higher than that of experiments for GES (0.4607). A statistical evaluation using a two-sample paired t-test showed a p value equal to 2.28726×10^{-5}. MAGES exhibited a 0.03% improvement for the summary quality compared to the GES method for a top-five summary, as shown in Table 1. In other words, 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 of a monolingual space.

Systems	σ **= 5**	σ **= 10**
FACES (state-of-the-art)	1.4611	4.3641
MAGES	**1.7082**	**4.5523**
GES	1.6727	4.4191
Typed	1.4651	4.1120

Table 1: Evaluation of the quality of summaries ($\lambda = 4.5$).

4 Conclusion

In this demo, we have presented MAGES, which is a system for configuring a summary within entity groups for entities of a data set in DBpedia. Our evaluation shows that the MAGES approach to summary generation outperforms another DBpedia entity summarization system when compared to the user-created benchmark. Moreover, MAGES can extract a particular group's stable signatures using multilingual angle integration, which can provide a useful strategy for identifying the nature of a described entity.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP)(No. R0101-16-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform); the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP(2015M3A9A7029735).

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Gong Cheng, Thanh Tran, and Yuzhong Qu. 2011. RELIN: Relatedness and informativeness-based centrality for entity summarization. In *Lecture Notes in Computer Science*, pages 114–129. Springer Berlin Heidelberg, Berlin, Heidelberg, October.

Christiane Fellbaum, editor. 1998. *WordNet - An Electronic Lexical Database*.

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2015. Faces: Diversity-aware entity summarization using incremental hierarchical conceptual clustering. In Blai Bonet and Sven Koenig, editors, *AAAI*, pages 116–122. AAAI Press.

M. A. Wong J. A. Hartigan. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

David W. Mount, 2004. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.

Andreas Thalhammer and Achim Rettinger. 2014. Browsing DBpedia entities with summaries. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 511–515. Springer International Publishing, Cham, May.