i18n-CKG: Considerations in Building Internationalization Contextualized Knowledge Graphs

Niel Chah[1][0000−0002−3377−7823]

University of Toronto, Canada
niel.chah@mail.utoronto.ca

Abstract. This paper proposes a methodology for creating an Internationalization Contextualized Knowledge Graph (i18n-CKG) that is designed to capture the wide range of international concepts and concerns. This methodology first surveys how well known entities are represented through different properties in various language locales. Next, the i18n-CKG is constructed as a combination of the i18n properties with inter-language translations and a scoring system to measure the universality of the properties. As a brief exercise, an i18n-CKG that models the i18n differences for a popular entertainer in English, Korean, Russian, and Chinese is presented.

Keywords: Internationalization · i18n · Knowledge graph

1 Introduction

Internationalization, often abbreviated as “i18n” (18 represents the number of letters between the first and last), refers to the adaptation of a product or service based in a certain language or locale for another language or locale. As different countries, cultures, and communities have different priorities and ways of viewing the world, it is important for knowledge graphs to be able to flexibly represent these i18n differences. In this paper, a methodology for creating a contextualized knowledge graph that is sensitive to i18n issues, an i18n-CKG, is presented.

First, a brief literature review will show how i18n considerations have been previously addressed in existing knowledge graphs, such as DBpedia. Next, the methodology for creating an i18n-CKG will be described with a case study using representative knowledge graph applications from different language locales, English (en), Korean (ko), Russian (ru), and Chinese (zh). The paper ends with a discussion of future work.

2 Related Works

A number of prior works have expanded the i18n capabilities of Linked Data. There have been case studies of i18n for the Greek DBpedia [4], Bengali DBpedia

[8], multilingual ontology matching [6] and guidelines for multilingual linked data [2], in addition to the ongoing i18n language DBpedia versions and others such as YAGO [10] and Wikidata [11].

It is also important to recognize that people may underestimate the difficulties of representing international knowledge in Linked Data. Often, it is not as simple as mapping ontologies from different languages together. In software development, a number of notable resources have pointed out these false assumptions with names [5], languages [3] and code [1].

3 i18n-CKG Methodology

The intuition for this methodology can be expressed as follows. Although the English Wikipedia[1] is a widely used website, people from different languages, communities, and cultures maintain their own Wikipedia or equivalent repository of open knowledge. Similarly, although the Google search engine holds a significant market share globally, other search engines are prominent in certain locales, such as Naver in South Korea and Yandex in Russia. By analyzing these i18n data sources and search services as a "ground truth", it is possible to compare how different locales represent relationships between entities.

First, the i18n data sources and popular local services for delivering structured information should be identified. Wikipedia[2] consists of many language versions. The infoboxes in certain articles also display the structures (or ontologies) in the respective language Wikipedia. These multilingual infoboxes act as up-to-date sources of i18n data that may not be captured in older releases of linked data such as Wikidata and DBpedia. In the case of search services, Google is widely popular and holds a significant market share worldwide [7, 9]. In South Korea, Naver is a competitor to Google, while Baidu and Yandex dominate in China and Russia respectively [9]. To the extent that these local search services provide results catered to their language's audience, it is useful to compare how popular search queries and entities are represented differently.

For an implementation of this methodology, a notable entity that would be displayed in the search services of all four language locales (English (en), Korean (ko), Russian (ru), Chinese (zh)) was selected. As a popular Korean pop music entertainer with appearances in international music, television, and films, South Korean singer "Rain" was compared across the aforementioned services to determine which relationships (properties) were most prominently displayed in the structured infoboxes and search engine results. With the structured data from the various sources collected computationally, machine translation of non-English texts and human-supervised reconciliation of complex properties were used to create the final set of triples from each i18n source.

As part of the i18n-CKG, an additional measure is proposed to score how frequently the i18n properties are present across language locales. This expresses how "local" a specific property is for a language. Such scores would also be

[1] https://en.wikipedia.org/wiki/Main_Page

[2] https://www.wikipedia.org/

represented as triples: `(p, i18nPropScore_en, value)`. A simple implementation is to use $s_l = \frac{r}{t}$ where r is the presence of the property in a language (0 or 1) and t is the total times the property is present across l locales. Higher scores indicate more "local" or locale-specific properties, while lower scores mean more "universal" properties, unless the value is zero. The code to generate the proposed i18n-CKG is maintained on GitHub[3].

For a complete translation of the ontology across languages (ontology localization), this can be accomplished through methods outlined in existing guides [2]. The properties, $p \in P$, that were discovered in the languages, $l \in L$, each have a language specific label represented by a RDF triple as `(p, label, "string"@lang)`. For each p across l languages, additional triples on the order of a maximum of $(|p|(|l| - 1))(|l|)$ should be newly created, assuming no property labels are already available outside the local language. For Table 1, this would be up to $348 = ((29)(4 - 1))(4)$ additional triples.

4 Results of i18n-CKG Construction

Table 1 shows how the four language locales represent the properties of the entity "Rain". For instance, the "date of birth" and "place of birth" properties are the most consistently displayed across all data sources, suggesting that these properties are universally interesting across locales. A number of properties represent less well known attributes, such as weight, blood type, and zodiac sign. These properties are likely to be of interest in the language locales where they are represented, but not as relevant or interesting for locales that omit them.

A selection of the preliminary calculations of the i18n property scores are noted below. The higher scores indicate that a property is more relevant in a particular language locale, such as "Blood_type" for zh, while lower scores are attributed to properties that are widely found in many language locales. Once all additional triples have been created, the resulting i18n-CKG data may be loaded into an existing triplestore or knowledge graph to provide i18n context.

- `(Date_of_birth, i18nPropScore_en, 0.25 (= 1/4))`
- `(Date_of_birth, i18nPropScore_zh, 0.25 (= 1/4))`
- `(Blood_type, i18nPropScore_en, 0.0 (= 0/1))`
- `(Blood_type, i18nPropScore_zh, 1.0 (= 1/1))`

5 Conclusions and Future Work

A number of extensions to the methodology proposed in this paper are possible. Qualitatively, the differences between the properties in the different language ontologies can be compared more deeply. For example, for the property label "Born", there may be a combination of "place of birth", "date of birth", and/or "birth name" that are actually found in the data. In addition, there are also

[3] https://github.com/nchah/i18n-ckg

Table 1. Properties featured in i18n data sources and search services

Properties	en WP - Google		ko WP - Naver		ru WP - Yandex		zh WP - Baidu		Ct.
Native name	✓		✓		✓		✓		4
Date of birth	✓	✓	✓	✓	✓	✓	✓	✓	8
Place of birth	✓	✓	✓		✓	✓	✓	✓	7
Residence	✓								1
Occupation	✓		✓		✓		✓	✓	5
Years active	✓				✓		✓		3
Spouse	✓	✓	✓			✓	✓		5
Children	✓	✓	✓				✓		4
Also known as	✓						✓		2
Genres	✓		✓				✓		3
Instruments	✓						✓		2
Labels	✓		✓	✓	✓		✓		5
Associated acts	✓		✓				✓		3
Website	✓		✓	✓			✓		4
Height		✓		✓		✓		✓	4
Albums		✓						✓	2
Artist name			✓						1
Religion			✓						1
Religious name			✓						1
Weight				✓					1
Education				✓			✓		2
Awards				✓					1
Hanja name					✓		✓		2
Country					✓		✓		2
Ethnicity							✓		1
Language							✓		1
Debut date							✓		1
Blood type								✓	1
Zodiac sign								✓	1
Count	14	6	13	7	8	4	19	7	-

opportunities to expand the computational methods in future iterations of this research to scale with the amount of entities (on the order of millions) and languages (on the order of hundreds).

At an implementation level, the creation of i18n labeled properties requires international language translation capabilities that can also scale. A combination of reliable machine translation tools and linguistic verification by bilingual experts may need to be implemented to maintain accuracy. While this is straightforward for some properties (e.g. "date of birth"), difficulties may become apparent as more obscure or i18n-specific properties are examined. Improvements in these processes should be undertaken while also building on similar ontology localization and mapping efforts in such cases as DBpedia[4].

Limitations to the i18n-CKG approach are also important to consider. For instance, *biases* in the data may affect the coverage of i18n-CKG. Certain languages and communities that are not well represented online would not be in-

[4] http://mappings.dbpedia.org/index.php/Main_Page

cluded to the same degree as more widely used languages such as English and French. There may also be divergent representations of the same entity or property within the same language or locale, which require additional work to resolve.

As more and more locales are covered, it will also be necessary to develop processes for dealing with data that conflict between i18n locales. Conflicting data between i18n locales may arise due to political issues (e.g. territorial disputes), cultural differences, or unreliable i18n sources, among other reasons. Further work is needed to handle such potentially sensitive matters.

In this paper, a preliminary methodology for creating a contextualized knowledge graph that is sensitive to i18n issues, an i18n-CKG, was described. The two stage approach included a survey of i18n differences using various i18n sources of "ground truth" and the construction of a knowledge graph that accounted for the i18n differences. Future research in building contextualized knowledge graphs with i18n considerations in mind should prove fruitful as new methods and improved processes are developed.

References

1. Computerphile: Internationalis(z)ing code - computerphile. https://www.youtube.com/watch?v=0j74jcxSunY (2014)
2. Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., Aguado-de Cea, G.: Guidelines for multilingual linked data. In: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics. p. 3. ACM (2013)
3. Hamill, B.: Falsehoods programmers believe about language. http://garbled.benhamill.com/2017/04/18/falsehoods-programmers-believe-about-language/ (2017)
4. Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., Metakides, G.: Internationalization of linked data: The case of the greek dbpedia edition. Web Semantics: Science, Services and Agents on the World Wide Web **15**, 51–61 (2012)
5. McKenzie, P.: Falsehoods programmers believe about names. https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/ (2010)
6. Meilicke, C., GarcíA-Castro, R., Freitas, F., Van Hage, W.R., Montiel-Ponsoda, E., De Azevedo, R.R., Stuckenschmidt, H., ŠVáB-Zamazal, O., Svátek, V., Tamilin, A., et al.: Multifarm: A benchmark for multilingual ontology matching. Web Semantics: Science, Services and Agents on the World Wide Web **15**, 62–68 (2012)
7. NetMarketShare: Search engine market share. https://www.netmarketshare.com/search-engine-market-share.aspx (2018)
8. Sarkar, A., Marjit, U., Biswas, U.: Towards bengali dbpedia. Procedia Technology **10**, 890–899 (2013)
9. Statcounter: Search engine market share. http://gs.statcounter.com/search-engine-market-share (2018)
10. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
11. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)