Zhishi.lemon: On Publishing Zhishi.me as Linguistic Linked Open Data*

Zhijia Fang[1], Haofen Wang[1], Jorge Gracia[2], Julia Bosque-Gil[2], and Tong Ruan[1]

[1] East China University of Science and Technology, Shanghai, 200237, China
fzjacky@mail.ecust.edu.cn,{whfcarter,ruantong}@ecust.edu.cn
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte 28660 Madrid (Spain)
{jgracia,jbosque}@fi.upm.es

Abstract. Recently, a growing number of linguistic resources in different languages have been published and interlinked as part of the Linguistic Linked Open Data (LLOD) cloud. However, in comparison to English and other prominent languages, the presence of Chinese in such a cloud is still limited, despite the fact that Chinese is the most spoken language worldwide. Publishing more Chinese language resources in the LLOD cloud can benefit both academia and industry to better understand the language itself and to further build multilingual applications that will improve the flow of data and services across countries. In this paper we describe Zhishi.lemon, a newly developed dataset based on the *lemon* model that constitutes the lexical realization of Zhishi.me, one of the largest Chinese datasets in the Linked Open Data (LOD) cloud. Zhishi.lemon combines the *lemon* core with the *lemon* translation module in order to build a linked data lexicon in Chinese with translations into Spanish and English. Links to BabelNet (a vast multilingual encyclopedic resource) have been provided as well. We also present a showcase of this module along with the technical details of transforming Zhishi.me to Zhishi.lemon. The dataset is accessible on the Web for both humans (via a Web interface) and software agents (with a SPARQL endpoint).

Keywords: linked data, translation, multilingualism

1 Introduction

With the development of the Semantic Web, a growing number of structured data in the form of RDF triples have been published and interlinked together as Linked Open Data (LOD) on the Web. Among them, Zhishi.me [10] constitutes

* This work was supported by the 863 plan of China Ministry of Science and Technology (No: 2015AA020107), the National Science Foundation of China (No: 61402173), Software and Integrated Circuit Industry Development Special Funds of Shanghai Economic and Information Commission (No: 140304), the LIDER FP7 EU project, and the Spanish Government through the Juan de la Cierva and the FPU programs.

the first effort to publish Chinese knowledge into the LOD cloud at a large scale. It gathers RDF triples from the three largest Chinese encyclopedic Web sites: Baidu Baike[3], Hudong Baike[4] and Chinese Wikipedia[5].

Recently, there has been a growing trend in publishing language resources (LRs) and interlinking them as part of the Linguistic Linked Open Data (LLOD) cloud[6]. The motivation is to have richer linguistic information accessible on the Web of Data, so that it can be consumed by a new generation of linked data-aware natural language processing (NLP) tools and services. In this context, the *lemon* model (LExicon Model for ONtologies) [9] was designed to bridge the gap between lexical and conceptual information, being now a de-facto standard for representing and publishing lexical resources as linked data on the Web. The *lemon* model has been used to expose bilingual and multilingual dictionaries on the Web of Data [5, 2], such as the bilingual from the Apertium initiative [4]. Another example is the RDF version of WordNet [8], created and structured according to *lemon*. Also BabelNet [3], a huge multilingual lexical and encyclopedic resource, has been modeled in *lemon* and published as linked data.

Compared to English and other prevalent languages in the LLOD cloud, resources in Chinese are scarce. In this paper, we move a step towards increasing the presence of lexical information in Chinese in the LLOD cloud while linking it to data in other languages (Spanish and English in particular). Concretely, we have built a new dataset, *Zhishi.lemon*, which constitutes the lexical realization of Zhishi.me. The work closest to ours is Chinese WordNet (CWN) [7]. The main difference between the two efforts is that we focus on entity level word translation while CWN emphasizes conceptual word alignment. Further, Zhishi.me contains a larger number of entities than CWN, usually local denominations, making it a suitable candidate to enrich the LLOD cloud with additional Chinese entries.

In our approach, DBpedia[7] and BabelNet[8] are used as a bridge to help identify correspondences between lexical entries in different languages. We combine the *lemon* core with its translation module [6] to build linked data lexicons in Chinese with translations into Spanish and English. Additional descriptions and lexical relations can be obtained by querying these linked data LRs to enrich the information in Zhishi.me. One advantage is that all the lexical information and translations are external to the original resource (Zhishi.me) so that there is no need to modify it whenever new lexical information is added into Zhishi.lemon. This is consistent with the "semantics by reference" principle followed in *lemon*.

The rest of this paper is organized as follows. Section 2 introduces technical details about linking Zhishi.me to the resources in the LLOD cloud. Section 3 gives an overview of the ontology we designed. Section 4 shows the access mechanisms and experiment results and we conclude the paper in section 5.

[3] `http://baike.baidu.com/`

[4] `http://www.baike.com/`

[5] `https://zh.wikipedia.org/`

[6] `http://linguistic-lod.org/llod-cloud`

[7] http://dbpedia.org/

[8] http://babelnet.org/

2 Linking Zhishi.me to DBpedia and BabelNet

In this section, we introduce the approach used to link Zhishi.me to two widespread resources in the LLOD cloud, namely DBpedia (its Spanish and English portions in particular) and BabelNet. Detecting equivalences among them can help identify translations between entity labels[9] expressed in different languages (Chinese, Spanish and English). However, it is impossible to manually align these three large datasets: not only it requires experts who are proficient in all the three languages, but it also needs great human labor due to the numerous resources to be aligned. We present an automatic way to tackle this problem.

DBpedia is making a major impact on the LLOD. In order to link Chinese resources in Zhishi.me to Spanish and English DBpedia, we turn to the cross-lingual equivalence relations in DBpedia to retrieve the corresponding translations. Since Zhishi.me includes Chinese Wikipedia as one of its sources and interlinks its resources to the equivalent ones in the other two sources (Baidu Baike and Hudong Baike), Chinese Wikipedia serves as a bridge to help detect links from Chinese resources to both their Spanish and English equivalences.

The highly multilingual nature of BabelNet can also be exploited to discover additional equivalences between resources in different languages. However, more than one "Babel synset" can be found for every ambiguous Chinese term. In order to identify the correct "Babel synset", we use the category labels in both BabelNet and Zhishi.me to find the disambiguation result for each Chinese term. BabelNet extracts categories of Wikipedia pages and maps them to WordNet. In [11], we publish the Chinese Linked Open Schema, which has further refined the existing categories in Zhishi.me, so that categories in both Zhishi.me and BabelNet are well-organized and of good quality. Therefore, we can collect a set of category labels from both sources for a given term in Zhishi.me. The larger the overlap between the category set of the term in Zhishi.me and that of the "Babel synset" in BabelNet, the higher the probability that the term can be mapped to that "Babel synset". The "Babel synset" with the largest overlap is then selected as the disambiguation result.

3 Ontology Overview

We create a new lexical dataset, *Zhishi.lemon*, which constitutes the lexical realization of Zhishi.me and contains its translations into other languages. In order to explain the dataset concretely, we divide it into two parts (namely the Chinese lexicalization module and the multilingual translation module) and provide a detailed analysis for each of them in the following two subsections.

3.1 Chinese Lexicalization Module

According to the design of the *lemon* core, each entity label in Zhishi.me is modeled as a *lemon: LexicalEntry* whose *lemon: LexicalSense* points to the appropri-

[9] In the following sections, we are using the expressions "labels" of entities and "terms" in an interchangeable way.

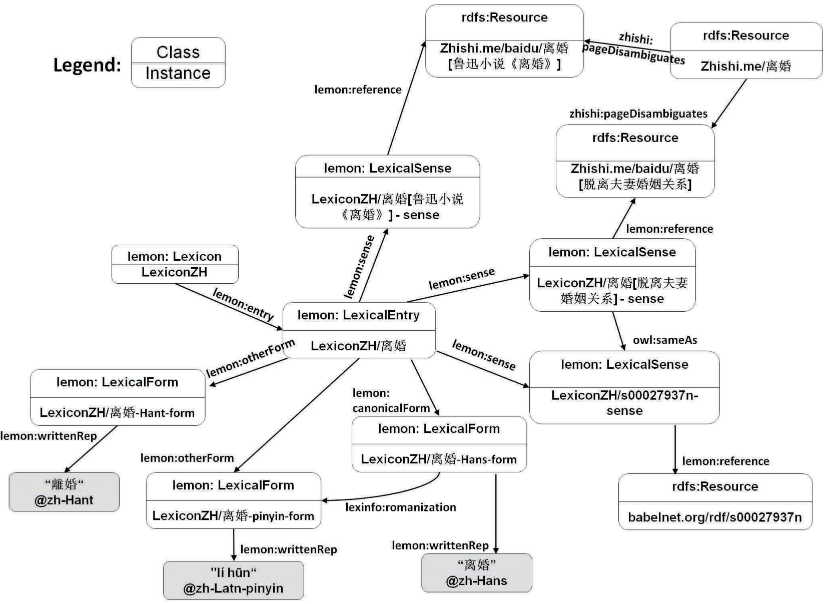

Fig. 1. Chinese Lexicalization Module

ate ontology reference. Since Zhishi.lemon includes translations among Chinese, Spanish and English, we create a monolingual lexicon for each language which will gradually grow when more resources of the same language are added into Zhishi.lemon. As shown in Figure 1, we create an instance of *lemon:Lexicon* called "LexiconZH" to gather all Chinese lexical entries.

The *lemon* model assumes that a lexical entry is not semantically disambiguated until an ontology reference provides the semantics of the entry. In Zhishi.me, the title of an article in an encyclopedia site is used as the label of its corresponding entity, and this is the label that we model as a lexical entry. However, labels may be ambiguous, i.e, they can be linked to more than one possible entity. In order to deal with these ambiguities, Zhishi.me uses *zhishi:pageDisambiguates* to represent that a single label refers to more than one entity. As shown in the example presented in Figure 1, "离婚" can refer to a Chinese novel as well as to a Chinese word with the same sense as "divorce (the legal dissolution of a marriage)" in English. Such word sense disambiguation has been captured in Zhishi.me, so that the label of the subject in a triple describing the *zhishi:pageDisambiguates* relation can be used to create the *lemon:LexicalEntry*, while the object of the triple is its ontology reference.

Given the fact that we are not modifying either Zhishi.me or the other sources (BabelNet, DBpedia), equivalent ontology entities retrieved from DBpedia and BabelNet will be declared at the lexico-semantic layer that Zhishi.lemon describes. In particular, we link the lexical senses that associate a common lexical entry with two semantically equivalent ontology descriptions by using a *owl:sameAs* relation. Figure 1 exemplifies that the Chinese lexical entry “离婚” has three possible lexical senses, two of which are describing the same meaning. Accordingly, we treat both of them as equivalent to one another.

We also integrate some features of Chinese itself into our lexicalization model. First of all, both simplified and traditional Chinese characters are standard character sets of contemporary written Chinese. Therefore, it is necessary to model the two different Chinese script variants in Zhishi.lemon. To that end, we propose the use of different language codes in order to distinguish these two different written forms, according to the guidelines provided by W3C[10]: “zh-Hans” to represent simplified Chinese characters while “zh-Hant” for traditional ones.

On the other hand, romanization is another interesting phenomenon presented in the Chinese language. It is a process of transcribing a language into the Latin script. Today, most Chinese use “Hanyu Pinyin” (simply as “pinyin”) as a common romanization standard. Since the “pinyin” forms should be included in the model as well, we propose the code “zh-Latn-pinyin” as the language code, which follows the W3C internationalization standards[11].

3.2 Multilingual Translation Module

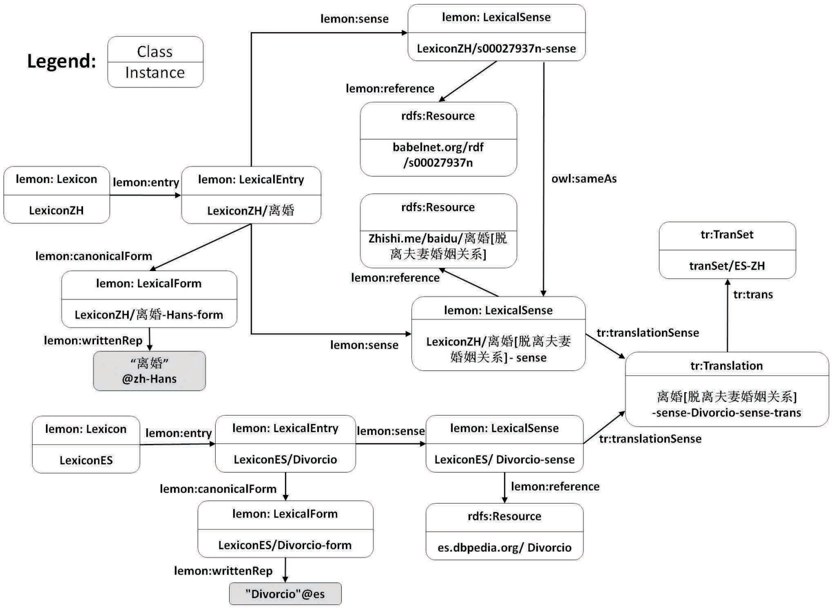

Fig. 2. Multilingual Translation Module

In Section 2, we enrich the multilingual information by discovering alignments with DBpedia and BabelNet. The aligned entity pairs have been used to

[10] `http://www.w3schools.com/tags/ref_language_codes.asp`
[11] `https://www.w3.org/International/questions/qa-choosing-language-tags`

construct the Zhishi.lemon translation module. Concretely, translation relations can be inferred between terms in different languages when they refer to the same ontology entity. Those lexical senses with an equivalent ontology reference have been regarded as a translation pair to be modeled.

To support the representation of such multilingual information, we use the classes *Translation* and *TranslationSet* in the *lemon* translation module to describe the translation relation. Figure 2 shows our proposed diagram. Lexical entries and their associated properties are used to account for the lexical information, which has been discussed in details in Section 3.1. The *TranslationSense* puts the lexical entries from different languages in connection through their lexical senses. *TranslationSet* is designed to group a set of translations, which facilitates querying. For instance, if someone wants to retrieve the Spanish terms of a certain Chinese term, he only needs to query the translation set *tranSet/ES-ZH* instead of searching through the whole dataset.

4 Zhishi.lemon Publication

In this section, we first introduce the online Web access and then give a data statistics of Zhishi.lemon. The data dump is also available via datahub[12].

4.1 IRI Naming Strategy

According to the Linked Data principles, Zhishi.lemon creates IRIs (Internationalized Resource Identifiers) for all resources and provides sufficient information when someone looks up an IRI via the HTTP protocol. Table 1 gives a general view of designed IRI patterns of the dataset. We have followed well established recommendations for this activity [1]. Since Zhishi.lemon consists of a series of linked data lexicons in Chinese, Spanish and English, with translations among them, we use “lexicon” and “tranSet” to indicate the nature of the different resources. In the pattern, [Lang] refers to three possible language marks [ZH], [ES] and [EN] while [LangTag] distinguishes the different character sets used in Chinese. In order to construct the IRIs of the rest of lexical elements, we preserved the labels of the original data in Zhishi.me, denoted as [label], whenever possible, propagating them into the RDF representation. In addition, some other suffixes have been added for the sake of readability: “-form” for lexical forms, “-sense” for lexical senses, and “-trans” for translations. In addition, we have made all the generated information accessible on the Web [13] for both humans (via a Web interface) and software agents (with a SPARQL endpoint).

4.2 Data Statistics

We first provide a general statistics about Zhishi.lemon. As shown in Table 2, the whole dataset contains 364,765 translations and after its conversion into the

[12] `https://datahub.io/dataset/zhishi-lemon`

[13] `http://lemon.zhishi.me/`

Table 1. IRI Patterns

Class	IRI Patterns
Lexicon	`http://zhishi.me/id/lemon/lexicon[Lang]`
Lexical Entry	`http://zhishi.me/id/lemon/lexicon[Lang]/[label]`
Lexical Sense	`http://zhishi.me/id/lemon/lexicon[Lang]/[label]-sense`
Lexical Form	`http://zhishi.me/id/lemon/lexicon[Lang]/[label]-[LangTag]-form`
Translation	`http://zhishi.me/id/lemon/tranSet[Lang1]-[Lang2]/[label1]-sense-[label2]-sense-trans`
Translation Set	`http://zhishi.me/id/lemon/tranSet[Lang1]-[Lang2]`

lemon representation model, 7,036,338 RDF triples were created. Among them, 229,606 resources in Zhishi.me have been found at least one cross-lingual target, in which 218,654 resources come from the English DBpedia, 77,392 from the Spanish DBpedia as well as 16,424 from BabelNet. Since we use a precision-oriented approach for link discovery between Zhishi.lemon and BabelNet, it results in a small number of links. That is to say, the overlap of categories is a hard constraint, so it may filter possible equivalent resources. The precision, however, remains relatively high, which will be evaluated in the next paragraph.

Table 2. Data Statistics

Items	Value
links:BabelNet	16,424
links:DBpedia-en	218,654
links:DBpedia-es	77,392
links:Zhishi.me	229,606
Total Translations	364,765
Total Triples	7,036,338

Table 3. Comparison with CWN

	CWN	Zhishi.lemon
Word/Lexical Entry	12,726	**215,608**
Sense/Lexical Sense	34,358	**523,585**
Lexical Relation/Translation	47,250	**364,765**

Then, we analyze the quality of the links established between Zhishi.me and BabelNet (with regard to the links to DBpedia, they have been inferred from the multilingual re-directions that Wikipedia contains, so we assume a high quality of such data). Among the 16,424 retrieved translation pairs between Zhishi.me and BabelNet, we select a random subset of 8,536 pairs and ask three students in our laboratory to manually check the quality. Since BabelNet has already provided "sameAs" links between different LRs, we only need to verify whether the Chinese terms in Zhishi.me could be aligned to the corresponding ones in BabelNet, which does not need the annotators to be proficient in all three languages. The experiment shows positive results with an extremely high precision (more than 0.98). After analyzing negative cases, we find that some resources in Zhishi.me may be associated with more than one lexical sense. Here, different lexical senses are included in one resource in Zhishi.me, while its corresponding senses are separated in BabelNet, which leads to a mismatch.

Finally, we compared Zhishi.lemon and CWN. Results are shown in Table 3. First, Zhishi.lemon focuses on translations at entity level instead of conceptual word alignment, as opposed to CWN. We believe that cross-lingual alignment among real-world objects will better benefit the LLOD community. Also, Zhishi.lemon achieves a larger scale in all the three types of elements, namely lexical entries, lexical senses and translations. In this sense, we are convinced that it will greatly help fill the gap between Chinese LRs and the LLOD cloud.

5 Conclusion and Future Directions

In this paper, we introduced Zhishi.lemon, a newly developed dataset that constitutes the lexical realization of Zhishi.me. On the basis of the *lemon* core and its translation module, we built a linked data lexicon in Chinese, with translations into Spanish and English. Links to both DBpedia and BabelNet have also been created. Experiments showed the high quality of Zhishi.lemon, which makes it a promising starting point to respond to the lack of Chinese lexical resources in the LLOD cloud. In the future, we plan to transform more Chinese resources and integrate them into Zhishi.lemon. Identifying new translations to other prevalent languages would be another possible direction. Furthermore, we plan to leverage Zhishi.lemon to build more real-world multilingual applications.

References

1. P. Archer, S. Goedertier, and N. Loutas. Study on persistent URIs. Technical report, Dec. 2012.
2. J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda, and G. Aguado-de Cea. Modelling multilingual lexicographic resources for the web of data: the k dictionaries case. In *Proc. of GLOBALEX'16 workshop at LREC'15, Portoroz, Slovenia*, May 2016.
3. M. Ehrmann, F. Cecconi, D. Vannella, J. P. McCrae, P. Cimiano, and R. Navigli. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proc. of LREC'14*, Reykjavik, Iceland, May 2014. ELRA.
4. M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.
5. J. Gracia. Multilingual dictionaries and the web of data. *Kernerman Dictionaries News*, (23):1–4, June 2015.
6. J. Gracia, E. Montiel Ponsoda, D. Vila Suero, and G. Aguado de Cea. Enabling language resources to expose translations as linked data on the web. 2014.
7. C.-Y. Lee and S.-K. Hsieh. Linguistic linked data in chinese: The case of chinese wordnet. *ACL-IJCNLP 2015*, page 70, 2015.
8. J. McCrae, C. Fellbaum, and P. Cimiano. Publishing and linking wordnet using lemon and rdf. In *Proc. of LDL'2014*, 2014.
9. J. P. McCrae, G. Aguado de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the semantic web. *LREJ*, 46(4):701–719, 2012.
10. X. Niu, X. Sun, H. Wang, S. Rong, G. Qi, and Y. Yu. Zhishi.me - weaving chinese linking open data. In *Proc. of ISWC'2011*, pages 205–220, 2011.
11. H. Wang, T. Wu, G. Qi, and T. Ruan. On publishing chinese linked open schema. In *Proc. of ISWC'2014*, pages 293–308, 2014.