Linguistic Linked Data in Chinese: The Case of Chinese Wordnet

Chih-Yao Lee
Graduate Institute of Linguistics,
National Taiwan University, Taiwan
chihyaolee@gmail.com

Shu-Kai Hsieh
Graduate Institute of Linguistics,
National Taiwan University, Taiwan
shukai@gmail.com

Abstract

The present study describes recent developments of Chinese Wordnet, which has been reformatted using the *lemon* model and published as part of the Linguistic Linked Open Data Cloud. While *lemon* suffices for modeling most of the structures in Chinese Wordnet at the lexical level, the model does not allow for finer-grained distinction of a word sense, or meaning facets, a linguistic feature also attended to in Chinese Wordnet. As for the representation of synsets, we use the WordNet RDF ontology for integration's sake. Also, we use another ontology proposed by the Global WordNet Association to show how Chinese Wordnet as Linked Data can be integrated into the Global WordNet Grid.

1 Introduction

Although the rationale underlying synsets remains disputable (Maziarz et al., 2013), the practical value of wordnet as lexical resource is undeniable, particularly that of the first and foremost of its kind, Princeton WordNet (PWN) (Fellbaum, 1998). According to a search run by Morato et al. (Morato et al., 2004) on some major bibliographic databases like LISA, INSPEC and IEEE, the decade between 1994 and 2003 saw a wide range of wordnet applications, including conceptual disambiguation, information retrieval, query expansion and machine translation, among others. At present, more than another decade after the survey, wordnets not only continue to assist in a variety of NLP tasks, but plays an important role in shaping the Semantic Web (Berners-Lee et al., 2001) along with other major language resources (De Melo, 2008).

Central to the practice of the Semantic Web is the use of Linked Data to harmonize and interlink resources and datasets on the Web. This idea has found its way into the world of linguistics and led to the emergence of the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al., 2011). Among the models available for lexicon representation, the *lemon* model (McCrae et al., 2012) is chosen. In adopting *lemon*, we intend not only to render Chinese Wordnet more accessible as Linked Data, but also to examine to what extent the model can express linguistic features peculiar to Chinese languages. On the other hand, we represent synsets using the WordNet RDF ontology designed by Princeton for use in the context of *lemon*. Finally, another ontology consisting of 71 Base Types proposed by the Global WordNet Association is used to illustrate how in the long run Chinese Wordnet can be integrated into the Global WordNet Grid (Pease et al., 2008).

2 Chinese Wordnet

Chinese Wordnet (CWN) is a lexical-conceptual network for Mandarin Chinese, its contents structured along the same lines of PWN. First constructed based on translational equivalents of PWN mapped to Suggested Upper Merged Ontology (Huang et al., 2004), CWN has been reconstructed from scratch in 2014 and released with an open-source license. As with most wordnets CWN provides knowledge about lexicalized concepts, including their representing lexical item's part-of-speech, definition, and a set of other lexicalized concepts with which they form a synset. To date, CWN contains more than 28,000 word-sense pairs that are organized in some 20,000 synsets. In addition to the synonymy implicitly present in synsets, CWN includes other lexical-semantic relations to connect the lexicalized concepts, meronomy and hypernymy-hyponymy in particular.

What distinguishes CWN from its counterparts for other languages are primarily the distinction of meaning facets (Ahrens et al., 1998; Hsieh, 2011)

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 70–74,
Beijing, China, July 31, 2015.

and a newly conceived type of relation termed *paranymy* (Huang et al., 2007). However, it is to be revealed that the current design of *lemon* does not allow for the representation of meaning facets and that the vocabulary of WordNet RDF ontology does not include *paranymy*.

3 Converting CWN into Linked Data with *lemon*

To improve its interoperability with other lexical resources, CWN is converted in RDF format using the *lemon* model. The following subsections provide a general introduction to *lemon* and Linked Data, followed by a discussion of the idiosyncrasies of Mandarin (as reflected in CWN) to be considered for a thorough conversion to a linked, *lemon*ized version of CWN.

3.1 The *lemon* Model and CWN

lemon (McCrae et al., 2011) is an ontology-lexicon model for representing lexical resources whose semantics is given by an external ontology. Following the principle of semantics by reference (Buitelaar, 2010), the model is meant to allow for linguistic grounding of a given ontology via supplementing the ontology with information about how the elements in the ontology's vocabulary are lexicalized in a given natural language. With the lexical and semantic layers separated as such, the same *lemon*-based lexicon can describe elements belonging to different ontologies; conversely, the same ontology can describe the semantics of all lexical resources in *lemon* format. As shown in Figure 1, the core of *lemon* includes:

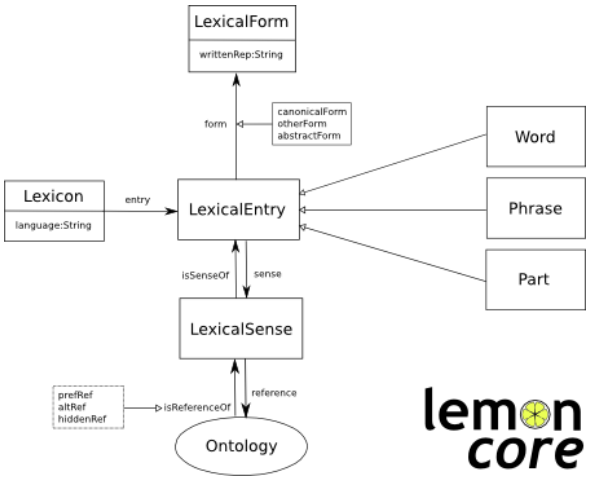

Figure 1: Core modules of the *lemon* model. (Taken from `http://lemon-model.net/`.)

- a *lexical entry*, which represents a single word or multi-word unit,
- a *lexical sense*, which represents the usage of a word as a *reference* to a concept in the ontology, and
- *forms*, which are inflected versions of the lexical entry, and associated with a string *representation*.

While *lemon* has proven adequate for modeling well-documented languages as those found in major lexical resources like PWN (McCrae et al., 2014) and Open Multilingual Wordnet (Bond and Foster, 2013), it remains to be seen whether the model is comprehensive enough for describing less privileged languages too. For instance, it is claimed that "the morphology module of *lemon* may serve less for Bantu languages lexica" (Chavula and Keet, 2014). In our case, while *lemon* suffices for modeling most of the structures in Chinese Wordnet at the lexical level, it does not allow for the representation of meaning facets. Consider the different uses of the lemma *shu1* "book" in the following sentences adapted from Bond et al. (2014):

(1) bang1 wo3 na2 na4 ben3 shu1
help me take that CL book
'Pass me that book.'

(2) ta1 zai4 du2 na4 ben3 shu1
he PROG read that CL book
'He is reading that book.'

(3) na2 yi4 ben3 shu1 gei3 wo3 kan3
take one CL book give me read
'Pass me a book to read.'

The same lemma *shu1* "book" refers to a physical object in (1) but to the information contained in (2). While the two readings may be referred to as different word senses, there exist contexts that allow the co-existence of both readings, as in (3), where the lemma can be interpreted as a physical object as well as the information contained in that object. Meaning distinction as such is therefore considered a facet rather than a sense.

Within the *lemon* model, however, there is no module for modeling meaning facets as there are for representing word forms and word senses. As a result, as many as 6,000 meaning facets identified in Chinese Wordnet cannot be published as part of the Linked Data for the time being.

3.2 Linked Data and Chinese Languages

Linked Data refers to data accessible on the Web and compiled such that it is machine-readable, its meaning is defined explicitly, and it is interlinked with other external data sets. Berners-Lee (2006) provides a set of guidelines for publishing Linked Data:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

Straightforward as the instructions may seem, the first rule regarding URI-naming already poses problems for languages whose writing system is not the Latin alphabet. Consider the URI scheme for identifying lemmas of a specific part-of-speech in the online RDF version of WordNet by Princeton[1]:

`http://.../wn31/` {**lemma**}-{**pos**}

If CWN adopts the same scheme and fills in the lemma slot with Chinese characters and specifies a lexical category, URIs as such will be generated:

`http://.../cwn/lod/` **詞目-n**

While multilingual addresses are well supported in modern web browsers, such URIs mean little to non-Chinese reading users and can hinder other resource providers from mapping CWN entries with their own. Another solution is to romanize the characters and number their tones:

`http://.../cwn/lod/` **ci2mu4-n**

Due to the prevalence of homophones in Chinese, however, the alternative leads to another issue: there exist many heterographs distinguishable only by their logographic representations when no context is given. A romanized form like *ci2mu4* can be interpreted nominally as "shrine-tomb" (祠墓) or "Ibaraki city" (茨木) as well as "lemma" (詞目). As a result, the design of such URI scheme is not effective in identifying a specific lexical entry, at least not for Chinese.[2] On the other hand, the RDF version of WordNet found in lemonUby[3] points to lemmas using the following URI scheme:

`http://.../WN_LexicalEntry_` {**id**}

By contrast, lemonUby makes use of unique IDs in combination with the prefix *WN_LexicalEntry_* to ensure one-to-one correspondence between URIs and lexical entries. Truly unique lemma identifiers are derived as such, even though the scheme observes the first rule for serving Linked Data only loosely, in the sense that with the prefix as the sole meaningful component part and without a lexical form embedded in the URI, the naming does not shed much light on the entry being linked to.

To uniquely identify lemmas without trading off URI readability on the part of the end user, CWN points to lemma entries using both a romanized lexical form and a unique ID. Take for example the following URI:

`http://.../cwn/lod/` **ai4** / **067081**

While the ID 067081 alone suffices to pinpoint its associated lexical entry, *ai4* "love" helps indicate the phonetic form of the lemma being referred to. When the trailing ID is not specified, however, all the entries with the romanization *ai4* will be listed along with their respective IDs. The optionality of the ID component part enables the user (or agent) to begin a query with a romanized form and then narrow it down to a specific lexical entry. Moreover, the path to a lemma can be further appended by a hash tag and a number to point to one sense of the lemma.[4] As for URIs of synsets, since a synset typically contains more than one sense and therefore cannot be represented with one single lexical form, CWN uses only IDs to identify a synset, as the RDF version of WordNet does in lemonUby.

While the first two rules address the scheme and the type of URIs to be used, the last two concern the contents to be served when a URL is dereferenced. In adopting the RDF-native *lemon* model,

[1] `http://wordnet-rdf.princeton.edu/`

[2] Note that the same situation is observed with URIs embedded with lexical forms of alphabetic languages when homophony occurs. For example, The URL `http://wordnet-rdf.princeton.edu/wn31/bank-n` points to both "river bank" and "financial bank" in PWN.

[3] `http://lemon-model.net/lexica/uby/wn/`

[4] Fore example, `http://lope.linguistics.ntu.edu.tw/cwn/lod/biao3/041141#11` points to the eleventh sense of the lemma *biao3* "show".

CWN meets the third rule of using standard formats at the outset. As for the fourth rule that requires the inclusion of other URIs, links to PWN's synsets are included that correspond to those of CWN. This last rule is to be addressed in more detail in Section 4.

3.3 CWN as Linked Data

Chief among the threads of information to be converted in RDF are the word senses and synsets of CWN. While the former correspond readily to *lemon*'s lexical senses, their lemmas to *lemon*'s lexical entries, the latter require special treatment. To comply with the aforementioned principle of separating linguistic realizations from underlying concepts, synsets are regarded as ontological references with which word senses are associated. Using the WordNet RDF ontology[5] introduced by McCrae et al.(2014) for use in the context of *lemon*, we represent CWN's synsets as a subclass of *Concept* in SKOS (Miles and Pérez-Agüera, 2007), expressing synsets without describing them with a formal ontological type. Figure 2 depicts a *lemon* representation of the first sense of the lemma *dong4wu4* "animal" in Turtle format.[6]

```
@prefix owl: <http://www.w3.org/2002/07/
    ↪ owl#> .
@prefix rdf: <http://www.w3.org
    ↪ /1999/02/22-rdf-syntax-ns#> .
@prefix lemon: <http://www.lemon-model.
    ↪ net/lemon#> .
@prefix wordnet-ontology: <http://
    ↪ wordnet-rdf.princeton.edu/
    ↪ ontology#> .
<http://lope.linguistics.ntu.edu.tw/cwn/
    ↪ lod/dong4wu4/052268> a lemon:
    ↪ LexicalEntry ;
    lemon:canonicalForm <#CanonicalForm>
        ↪ ;
    lemon:sense <#1> ;
    wordnet-ontology:part_of_speech
        ↪ wordnet-ontology:noun .
<#CanonicalForm> a lemon:Form ;
    lemon:writtenRep "@cmn .
<#1> a lemon:LexicalSense ;
    lemon:reference <http://lope.
        ↪ linguistics.ntu.edu.tw/cwn/
        ↪ lod/2068> ;
    wordnet-ontology:gloss
        ↪
        ↪ "@cmn ;
    owl:sameAs <http://wordnet-rdf.
        ↪ princeton.edu/wn31/100015568-
        ↪ n> .
```

Figure 2: The first sense of *dong4wu4* in Turtle.

In the WordNet RDF ontology, however, there is no vocabulary for describing the relation between coordinate terms that share the same classificatory criteria, or *paranymy*. Take *season (of the year)* for example. Except when referring to a tropical climate, a first impression about the term is oftentimes the categorization of *spring*, *summer*, *fall* and *winter*. Other terms such as *dry season* and *rainy season* are not thought of as parallel as the four seasons, even though all of them share the same immediate superordinate concept (Huang et al., 2008). While CWN attends to this syntagmatic relation between different groupings of hyponyms, it can only be expressed when PWN adopts this type of relation or when a tailor-made ontology for *lemon*-CWN is in place.

4 Interlinking *lemon*-CWN on the Web

As shown in Figure 2, there can be an outward link to PWN if the synset referenced by a lexical sense has a comparable entry in PWN. By way of synset mapping, *lemon*-CWN is not only linked to PWN, but also indirectly interlinked with other wordnets via PWN. Besides using PWN as key to the LLOD cloud and interface with other linguistic resources, *lemon*-CWN can be integrated into the Global WordNet Grid when organized, along with other wordnets, by the ontology consisting of 71 Base Types proposed by the Global WordNet Association.[7] An initial mapping has identified 169 synsets comparable to the Base Types.[8]

5 Conclusion

We have described a *lemon*ized version of CWN to be integrated in the LLOD cloud and the Global WordNet Grid. In converting CWN into Linked Data, we have established a URI scheme optimal for encoding Chinese lemmas alternatively written in the Latin alphabet. Also, we have pointed out two aspects of CWN that cannot be expressed using *lemon* and the WordNet RDF ontology, respectively the unit of meaning facets and the relation of paranymy. Future work thus includes finding another model that allows for the representation of meaning facets and designing an ontology for *lemon*-CWN that has vocabulary for paranymy.

[5] http://wordnet-rdf.princeton.edu/ontology

[6] http://www.w3.org/TR/turtle/

[7] http://w.globalwordnet.org/gwa/ewn_to_bc/BaseTypes.htm

[8] http://lope.linguistics.ntu.edu.tw/cwn/gwn/

References

Kathleen Ahrens, Li-Li Chang, Ke-Jiann Chen, and Chu-Ren Huang. 1998. Meaning representation and meaning instantiation for chinese nominals. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 1, February 1998: Special Issue on the 10th Research on Computational Linguistics International Conference*, pages 45–60.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5):34–43.

Tim Berners-Lee. 2006. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`. [Online; accessed 28-April-2015].

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362. The Association for Computer Linguistics.

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease, and Piek Vossen. 2014. A multilingual lexico-semantic database and ontology. In Paul Buitelaar and Philipp Cimiano, editors, *Towards the Multilingual Semantic Web*, pages 243–258. Springer Berlin Heidelberg.

Paul Buitelaar. 2010. Ontology-based semantic lexicons: Mapping between terms and object descriptions. *Ontology and the Lexicon*, pages 212–223.

Catherine Chavula and C. Maria Keet. 2014. Is *lemon* sufficient for building multilingual ontologies for bantu languages? In *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014) co-located with 13th International Semantic Web Conference on (ISWC 2014), Riva del Garda, Italy, October 17-18, 2014.*, pages 61–72.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, pages 245–275.

Gerard De Melo. 2008. Language as a foundation of the semantic web.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Shu-Kai Hsieh. 2011. Sense structure in cube: Lexical semantic representation in chinese wordnet. *International Journal of Computer Processing of Languages*, 23:243–253.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Chu-Ren Huang, I-Li Su, Pei-Yi Hsiao, and Xiu-Ling Ke. 2007. Paranyms, co-hyponyms and antonyms: Representing semantic fields with lexical semantic relations. In *Proceedings of the 8th Chinese Lexical Semantics Workshop 2007*, Hong Kong: Hong Kong Polytechnic University.

Chu-Ren Huang, I-Li Su, Pei-Yi Hsiao, and Xiu-Ling Ke. 2008. Paranymy: Enriching ontological knowledge in wordnets. In *Proceedings of the Fourth Global Wordnet Conference*, pages 221–228, Szeged, Hungary: University of Szeged.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications - Volume Part I*, ESWC'11, pages 245–259, Berlin, Heidelberg. Springer-Verlag.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3 rd Workshop on Linked Data in Linguistics*.

Alistair Miles and José R. Pérez-Agüera. 2007. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83.

Jorge Morato, Miguel Angel Marzal, Juan Lloréns, and José Moreiro. 2004. Wordnet applications. *GLOBAL WORDNET CONFERENCE*, 2:270–278.

Adam Pease, Christiane Fellbaum, and Piek Vossen. 2008. Building the global wordnet grid. *Proceedings of the CIL-18 Workshop on Linguistic Studies of Ontology*.