Direct and Indirect Linking of Lexical Objects for Evolving Lexical Linked Data

Yoshihiko Hayashi

Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama, 5600043 Toyonaka, Japan
hayashi@lang.osaka-u.ac.jp

Abstract. *Servicization* of language resources in a Web-based environment has opened up the potential for dynamically combined virtual lexical resources. *Evolving lexical linked data* could be realized, provided being recovered/discovered links among lexical resources are properly organized and maintained. This position paper examines a scenario, in which lexical semantic resources are cross-linguistically enriched, and sketches how this scenario could come about while discussing necessary ingredients. The discussions naturally include how the existing lexicon modeling framework could be applied and should be extended.

Keywords: lexical linked data, lexicon models, multilingual lexical resources, cross-lingual semantic similarity

1 Introduction

Servicization of language resources provides the potential of a dynamic lexical resource [4], which realizes a virtual yet composite lexical resource by combining servicized resources with a service workflow. Furthermore, it is expected that the recovered/discovered relationships among lexical objects in existing language resources can be organized as a secondary language resource, and hence can be effectively reused [6]. This direction could harmonize with the recent trend of Linked Data, as the derived relationships are being overplayed as *links* on top of the primary lexical resources. We would call such a lexical space *evolving lexical linked data* as a whole.

This position paper argues that by opportunistically associating different lexical resources across a language barrier, relevant portion of the lexical resources can be gradually enriched and could be made public by standing on the Linked Data mechanism. This paper also argues more relationships could be acquired, when there exists a lexical semantic disparity.

2 Basic Lexicon Model

The presented work concentrates on WordNet-type semantic lexicons. Their fundamental information structures are represented by the following lexical class objects.

- A `Lexical Entry` comprises of `Forms` and `Senses`.
- A `Form` can be a `Lemma` or a `Phrase`; the latter comprises of more than one `Lemmas`.
- A `Sense` denotes a `Synset`.
- A `Synset` is denoted by one or more `Senses`.
- `Synsets` are linked by one of the predefined `Conceptual Relations`.

3 Conceptual Framework of Evolving Lexical Linked Data

Below we introduce a motivating example, where an English query term *gadget* is issued to search for a set of corresponding Japanese translations, each hopefully grounded in a Japanese conceptual system. Suppose we get two translations, under the same sense division, for *gadget* by using an appropriate translation resource: *t1*:”ガジェット” (*gajetto*), which is the transliteration of *gadget*, and *t2*:”有用な機器” (*yuuyounakiki*), which actually is a two-word phrase.

3.1 Direct Linking of Lexical Objects

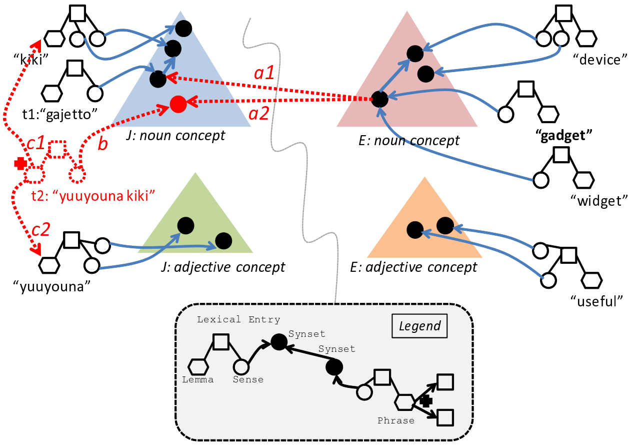

Fig. 1. Evolving Lexical Linked Data: direct linking has been conducted.

Figure 1 illustrates the relevant portion of the lexical linked data just after the query was entered, in which newly introduced lexical objects are indicated by dotted lines. First, cross-lingual synset-to-synset links *a1* and *a2* are introduced. Introduction of *a1* may require sense disambiguation, because *t1*, which is supposed to reside in the Japanese lexical space, could have more than one senses. A `Lexical Entry` node as well as a `Synset` node are, on the other hand, introduced for accommodating *t2*. As *t2* should be morpho-syntactically parsed into [有用な (yuuyouna)/Adj, 機器 (kiki)/Noun], a `Phrase` node is introduced to associate this two-word phrase with its constituents by the *c1* and *c2* links.

These successive operations are invoked directly while handling the query; we thus call them *direct linking* of lexical objects. Note that the ad-hoc `Synset` node is yet to ground in the Japanese conceptual system at this time.

3.2 Indirect Linking of Lexical Objects

While the structure around *t1* has been settled in the current configuration, that of around the ad-hoc `Synset` node for *t2* can be further enriched, again by seeking cross-lingual correspondences. Figure 2 summarizes the outcomes.

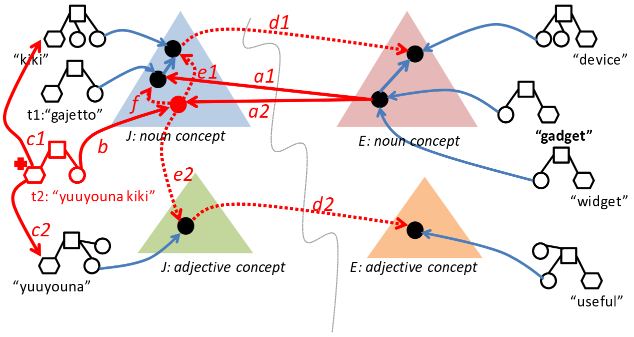

Fig. 2. Evolving lexical linked data: indirect links are introduced.

Two cross-lingual synset-to-synset links (*d1* and *d2*) are first introduced by associating a sense of "機器" (*kiki*) with a sense of *device* and a sense of "有用な" (*yuuyouna*) with a sense of *useful* respectively. By establishing *d1*, the semantic head of the ad-hoc synset for *t2* is then identified and represented by the link *e1*. The same story holds for the semantic modifier of *t2*, and the link *e2* is introduced to represent this semantic relationship. These operations also enable the introduction of the link *f*, which, in a sense, shows "ガジェット" (*gajetto*) can be rephrased as "有用な機器" (*yuuyounakiki*).

The evolving story so far signifies us the possibility of lexical knowledge enrichment that takes advantage of the opportunity to interrelate lexical objects across a language barrier. Let us remind that a semantic gap brought about by differences in the lexicalization would provide us a further opportunity to enrich relevant range of the existing lexical structures.

We could acquire more correspondences as illustrated in Figure 3 by further pursuing this strategy. In the figure, another ad-hoc `Synset` node in the English lexical space, and two semantic links (*g1* and *g2*) to label the semantic head/modifier of the ad-hoc synset are introduced. Besides, the ad-hoc `Synset` node is linked to that of *gadget* by the link *h*; this is in parallel with the link *f* in the Japanese lexical space. Notice again that almost instant introduction of these links is originated from the cross-lingual synset-to-synset matching that is invoked for establishing the correspondences represented by *d1* and *d2*.

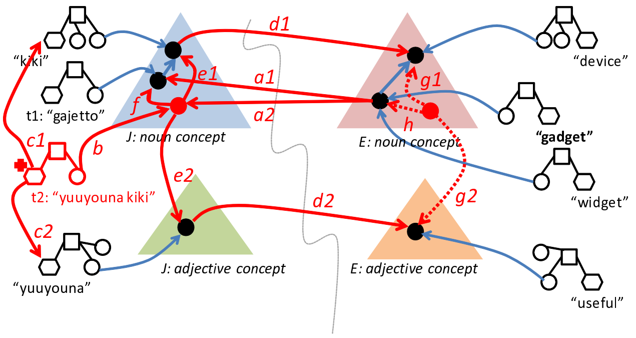

Fig. 3. Evolving lexical linked data: indirect links are further introduced.

We would call these secondary operations initiated after the direct linking as *indirect linking*. The lexical objects introduced in this motivating example are examined in more detail in the next section to sort the necessary elements to realize the scenario.

4 Enabling Direct and Indirect Linking

4.1 Modeling lexical information structure

The basic lexicon model described in section 2 has to be extended in some ways.

First, in the motivating example, two ad-hoc `Synset` nodes were introduced to accommodate the two-word translation phrase *t2*, and the corresponding virtual phrase (could be verbalized as *useful device*) in English. These nodes, in their nature, may be ad-hoc and represent a kind of complex concept that may lexicalize to a phrase rather than a single word in one language. Therefore an instance of the ad-hoc `Synset` class should have an attribute to indicate the instance is typed `complex`, and could have `Morpho-syntactic Head/Modifier` links (like *c1,c2*) as well as `Semantic Head/Modifier` links (like *e1,e2,g1,g2*).

Second, some of the introduced links should be typed differently from the existing lexicon model. Table 1 classifies the links introduced in the motivating example. The link type #1 is of intrinsic important in the presented framework. As the correspondence between synsets in different languages, in a sense, is rarely equivalent [7], it is necessary to label the relation type for each cross-lingual synset-to-synset link instance. We could develop a proper label inventory, presumably by basing on the one developed by EuroWordNet [9], while considering more bilingual characteristics. The link type #5, in a sense, is a variant of the link type #1; the difference is that the correspondence is cross-lingual or not. Therefore we can assume an upper class that subsumes these link types.

Table 1. Classification of the links introduced in the motivating example.

#	link instances	source node type	destination node type	relation type	computational process
1	*a1,a2,d1,d2*	`Synset`	`Synset`	cross-lingual correspondence	synset matching
2	*b*	`Sense`	`Synset`	denotation	–
3	*c1,c2*	`Phrase`	`Lemma`	morpho-syntactic decomposition	morpho-syntactic analysis
4	*e1,e2,g1,g2*	ad-hoc `Synset`	`Synset`	semantic decomposition	–
5	*f,h*	ad-hoc `Synset`	`Synset`	near-synonym	–

The link type #3 represents morpho-syntactic head/modifier relationships, whereas link type #4 represents semantic head/modifier relationships. As far as semantic compositionality holds, these two link types exhibit a kind of parallel structure as illustrated in the example: the semantic links (*e1* and *e2*; typed #4) were eventually introduced, corresponding to the already existing morpho-syntactic links (*c1* and *c2*; typed #3).

On the other hand, in cases where the semantic compositionality does not hold, we should demur the introduction of these semantic links, even each of the Japanese synsets could find their mates in the English lexical space. In such a case, we have to devise an independent method to check the semantic compositionality, or we should seek more semantic constraints to apply, probably from the English lexical space; but this issue largely remains as a future issue.

As for the actual modeling and representation of lexical resources, we can rest with the existing frameworks, including the ISO standard lexical markup framework (LMF) [5], and Lemon [3].

4.2 Matching synsets across a language

One of the most important elements is obviously a computational process for finding a synset mate in another language. We are now studying a method to calculate semantic similarity between synsets across a language, by simply employing bilingual translation resources and probability distributions acquired from a sense-tagged corpus in the target language.

We can also apply and/or combine previously proposed methods. For example, the method reported highly accurate [1] may be applicable with modifications, even it computes similarity between words rather than between synsets; the gloss-overlap-based method presented in [2] would also be readily applied, if we could translate the gloss in one language to another with a reasonable accuracy. However even with a highly promising method at hand, any synset-to-synset relation has to be established by choosing among computationally proposed candidates. The underlying process thus has to incorporate human intervention, where a collaborative operational environment plays a role.

4.3 Further issues

The following issues have to be considered in implementing an effective operating environment. First, we need to have a global mechanism to control the indirect linking operations. As shown in the example, indirect links can be introduced upon establishment of a direct link. However who/what should decide to initiate the indirect linking process is unclear. Moreover, to what extent the indirect linking should be propagated remains uncertain. Second, we are in need of having a proper vocabulary to annotate the lexical objects that participated in direct/indirect linking operations. For example, we would need to know when and how a particular link was established. We thus need to have a sort of ontology for describing linking events, which naturally includes references to the linguistic processes that were actually applied, as well as the human approvals.

5 Concluding Remarks

This position paper presented a notion of evolving linked data, in which recovered/discovered relationships among lexical objects would be published as *links*. It also argued that the associated lexical resources could be enriched further, in particular cases where a sort of lexical semantic disparity exists.

Acknowledgments. The presented work was supported by KAKENHI (21520401) provided by MEXT, Japan, and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan.

References

1. Agirre, E., Alfonseca, E., et al.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: *NAACL-HLT2009*, pp.19–27 (2009)
2. Banerjee, S., and Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: *IJCAI 2003*, pp.805–810 (2003)
3. Buitelaar, P., Cimiano, P., et al.: Towards Linguistically Grounded Ontologies. In: *ESWC 2009*, pp.111–125 (2009)
4. Calzolari, N.: Approaches towards a 'Lexical Web': the Role of Interoperability. In: *ICGL 2008*, pp.34–42 (2008)
5. Francopoulo, G., Bel, N. et al.: Multilingual Resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, Vol.43, No.1, pp. 57–70 (2009)
6. Hayashi, Y.: A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences. In: *IWCS 2011*, pp.155–164 (2011)
7. Hirst, G.: Ontology and the Lexicon. In: Staab, S., and Studer, R. (eds.): *Handbook of Ontologies, Second Edition.* Springer, pp.269–292. (2009)
8. Isahara, H., Bond, F., et al.: Development of the Japanese WordNet. In: *LREC 2008*, pp.2420–2423 (2008)
9. Vossen, P.: EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-lingual Index. *International Journal of Lexicography*, Vol.17, No.2, pp.161–173 (2004)