Cross-Language Hybrid Keyword and Semantic Search with Linguistically Grounded Extraction Ontologies

(Chapter Proposal for inclusion in *Towards the Multilingual Semantic Web*)

David W. Embley, Stephen W. Liddle, Deryle W. Lonsdale
Brigham Young University, Provo, UT, USA

Byung-Joo Shin
Kyungnam University, Kyungnam, Korea

Yuri Tijerino
Kwansei Gakuin University, Kobe-Sanda, Japan

1 Motivation for Chapter Inclusion

Based on our prior work in multilingual extraction ontologies [EZ10, ZELS12, ELL+12, ELLT11, LEL10], we address all three major topics in the call for book chapters:

Principles Our architecture consists of linguistically grounded extraction ontologies with support for free-form queries that generate SPARQL queries over OWL/RDF semantic indexes in multiple languages. Specifically, we address these "Principles" subtopics:

- models for the integration of linguistic information with ontologies, i.e., models for multilingualism in knowledge representation, in particular OWL and RDF(S)
- extensions of state-of-the-art query languages (SPARQL) to account for multilinguality
- web architecture to support multilinguality

Methods Our approach is essentially about retrieving information with cross-language search and query over document collections. Specifically, we address these "Methods" subtopics:

- multilingual and cross-lingual aspects of semantic search and querying of knowledge repositories
- cross-lingual information retrieval
- automatic integration and adaptation of (multilingual) lexicons with ontologies
- multi- and cross-lingual ontology-based information extraction and ontology population

Application We describe innovative and relevant applications and solutions for use cases in the area of the Multilingual Semantic Web.

In compliance with the request to select only one of the major sections in the book for the chapter, we can best classify our proposed chapter as being in the **Methods** section. Although we do have a specific and strong model for the integration of linguistic information with ontologies and although we do have a specific and strong architecture to support multilinguality, neither the model nor the architecture is OWL/RDFS-based. Neither do we extend SPARQL to account for multilinguality. Instead, our method maps the results of applying our linguistically-grounded, extraction-ontology model to generate RDF triples with multiple-language content to which we apply ordinary SPARQL queries generated from free-form queries stated in any one of several languages. Even so, in the **Methods** section we can still explain our model and architecture as we describe our method for cross-language search and query. As for applications, the examples we use to illustrate our method can also serve as use cases for the Multilingual Semantic Web.

2 Proposed Table of Contents

1. Introduction

   Motivating example and problem characterization
   *Our tool: ML-HyKSS (**M**ulti-**L**ingual **Hy**brid **K**eyword and **S**emantic **S**earch)*
   Introduction to the principles, methods, and applications embodied in ML-HyKSS

2. Language-Independent Extraction Ontologies

   Formal model of linguistically grounded extraction ontologies [EZ10]
   Language independent: works for any language
   Languages we've experimented with: Arabic, English, French, Japanese, Korean

3. Language-Independent Semantic and Keyword Indexing

   Methodology for building both a keyword index and a semantic index [ZELS12]
   Indexing for both web pages and other document collections (e.g., OCRed historical documents [ELL+11])

4. Language-Independent Free-Form Query Processing

   Free-form query processing with linguistically grounded extraction ontologies over semantic- and keyword-indexed document collections [ZELS12]
   Advanced form queries to allow for negations and disjunctions [ZELS12]

5. Cross-Language Query Processing

   Query translation at the conceptual level (rather than the more common language level) resulting in more accurate translations, as originally proposed in [LEL10] and formally defined in [ELLT11] with use-case evaluation in [ELL+12]
   Star architecture for cross-language lexicon translation, allowing for new languages to be quickly and efficiently integrated into ML-HyKSS

6. Pragmatics

   Issues that need resolution to make the ideas in ML-HyKSS work in practice
   Implementation status and outlook
   Application use cases: car ads (representing semi-structured sources) and obituaries (representing unstructured sources)
   Many additional application use cases (so far, extraction only—not yet cross-language): e.g., video games, genomics, campground facilities, prescription drugs, gem stones, computer monitors, genealogy, apartment rentals, restaurants, movies, country data, ... (basically any domain that is data rich and narrow in scope)

   (a) Extraction Accuracy for Semantic Indexing

   Experimental evaluations we have conducted [ELLT11, ELL+12]
   Ideas for improving extraction accuracy

   (b) Cross-Language Query Accuracy

   Experimental evaluations we have conducted [ELL+12]
   Ideas for improving cross-language query accuracy

   (c) Efficient Construction of Extraction Ontologies

   Use of off-the-shelf recognizers and language resources
   Automatic construction of query extraction-ontologies (new work in [SE13])

(d) Efficient Construction of Cross-Language Mappings

Star architecture, making new language additions $O(n)$ rather than $O(n^2)$
Use of web services, off-the-shelf code libraries, and language resources to semi-automatically instantiate mappings [ELL+ 12]

7. Conclusion

Contributions in terms of principles, methods, and applications
Expectations, limitations, and practicalities

3 Sketch of Technical Details

A *linguistically grounded extraction ontology* (see Figure 1) is a 4-tuple (O, R, C, L):

O : Object sets—one-place predicates (represented by named rectangles in Figure 1)

R : Relationship sets—n-place predicates, $n \geq 2$ (represented by lines connecting object-set rectangles and by black-triangle aggregation symbols connecting holonyms to meronyms)

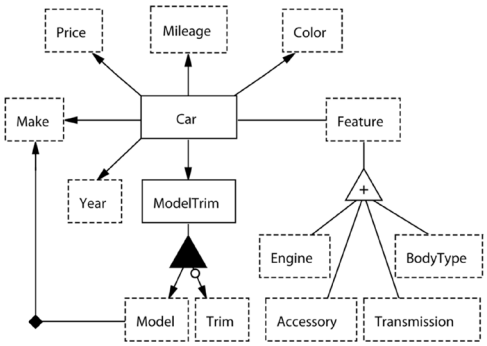
C : Constraints—closed formulas (e.g., $\forall x(Accessory(x) \Rightarrow Feature(x))$—one of the many hypernym/hyponym constraints denoted by the triangle, with the enclosed "+" symbol denoting mutual exclusion among the hyponym object sets; $\forall x(Car(x) \Rightarrow \exists! y(Car\text{-}Year(x, y))$—one of the many functional constraints denoted by the arrowhead on the range side of the Car-$Year$ relationship set; ...)

L : Linguistic groundings—text recognizers for populating object and relationship sets (e.g., partial *Price* and *Make* concept recognizers on the right side of Figure 1)

The conceptual foundation for an extraction ontology is a restricted fragment of first-order logic, but its most distinguishing feature is its linguistic grounding [BCHS09], which turns an ontological specification into an extraction ontology. Each object set has a *data frame* [Emb80], which is an abstract data type augmented with linguistic recognizers that specify textual patterns for recognizing instance values, applicable operators, and operator parameters. Figure 1 shows partial data frames for the object sets *Price* and *Make*. Although any kind of textual pattern recognizer is possible, our current implementation supports only regular expressions as exemplified in the *Price* data frame and lexicons as exemplified in the *Make* data frame, or combinations of regular expressions and lexicons. Relationship sets may also have data-frame recognizers or may be pre-populated with a fixed set of relationships (denoted by a black diamond on the edge, e.g., *Make-Model* in Figure 1). Recognizers for operators, such as for the *LessThan* operator in Figure 1, provide, in addition to keywords such as *less than* or *under* that indicate applicability, references to operand recognizers (e.g., *p2* in Figure 1, which references the *Price* instance recognizer).

Applying an extraction ontology to a text snippet (e.g., the Korean car ad in Figure 2) yields a populated ontology. When augmented by links to the text in a text snippet from which stored facts are extracted, a populated ontology becomes a semantic index. Before executing queries, ML-HyKSS crawls source documents and indexes facts with respect to its extraction ontologies. For example, Figure 2 indicates that ML-HyKSS has extracted and indexed several facts from the Korean car ad. In our implementation, we store our semantic index as RDF triples.

Given semantic and keyword indexes in various natural languages, ML-HyKSS processes cross-language queries as Figure 2 shows. In the example, ML-HyKSS applies a French extraction ontology to the query "Honda moins de 8000 en «excellent état»" and interprets it, discovering two semantic constraints (marquee = Honda and prix ≤ 8000€) and two keywords (Honda and «excellent état»). Given the resulting interpretation, ML-HyKSS translates the conceptualized French query into a




```
Price
  internal representation: Double
  external representation: \$[1-9]\d{0,2},?\d{3}
    |\d?\d [Gg]rand|...
  context keyword: price|asking|obo|neg(\.|otiable)|...
  ...
  unit: dollars|[Kk]|...
  canonicalization method: toUSDollars
  comparison method:
    LessThan(p1: Price, p2: Price) returns (Boolean)
    external representation: (less than|<
      |under|...)\s*{p2}|...
    ...
  output method: toUSDollarsFormat
  ...
end

Make
  ...
  external representation: CarMake.lexicon
  ...
```

Figure 1: Linguistically Grounded Extraction Ontology.

Korean conceptualization. We ensure that cross-language conceptualizations are structurally identical; and therefore since the semantic concepts and constraints have a one-to-one correspondence, the implied select-project-join operations for the query will be the same in both conceptualizations. As for instance values in semantic constraints and for keywords, ML-HyKSS uses existing services for currency conversions, unit conversions, and transliterations (all direct from language to language) and uses existing language resources and pay-as-you-go construction for lexicon, keyword, and commentary translations (all indirect through a central language-agnostic conceptualization so that adding new languages is less costly than it would be if translations between every pair of languages would need to be constructed). Once translated conceptually, ML-HyKSS can immediately generate a SPARQL query over the target language triple store. ML-HyKSS also takes into account keywords and computes a combined semantic and keyword score from which it ranks results. As Figure 2 shows, ML-HyKSS displays results as a relational table, which is augmented with keyword matches, ordered by ranked result, and linked to source documents—linked, for example, so that clicking on Honda (2) brings up the Korean text snippet in Figure 2 and highlights the two instances of 혼다, the translated keywords that appear on the page. Figure 2 also shows that cross-language query processing is symmetrical: a Korean extraction ontology interprets a Korean query (한국어), translates it conceptually into French and processes it against a pre-indexed français semantic and keyword repository.

The key ideas in the ML-HyKSS approach to cross-language query processing are: (1) semantic indexing with native-language extraction ontologies, (2) query interpretation with native-language extraction ontologies, (3) cross-language translation at the conceptual level (rather than at the query level), and (4) use of language resources and web services to quickly assemble extraction ontologies, while still allowing for more costly pay-as-you-go improvements by human experts.

References

[BCHS09] P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference (ESWC'09)*, pages 111–125, Heraklion, Greece, May/June 2009.

[ELL+11] D.W. Embley, S.W. Liddle, D.W. Lonsdale, S. Machado, T. Packer, J. Park, and N. Tate. Enabling search for facts and implied facts in historical documents. In *Proceedings of the International Workshop on Historical Document Imaging and Processing (HIP 2011)*, pages 59–66, Beijing, China, September 2011.

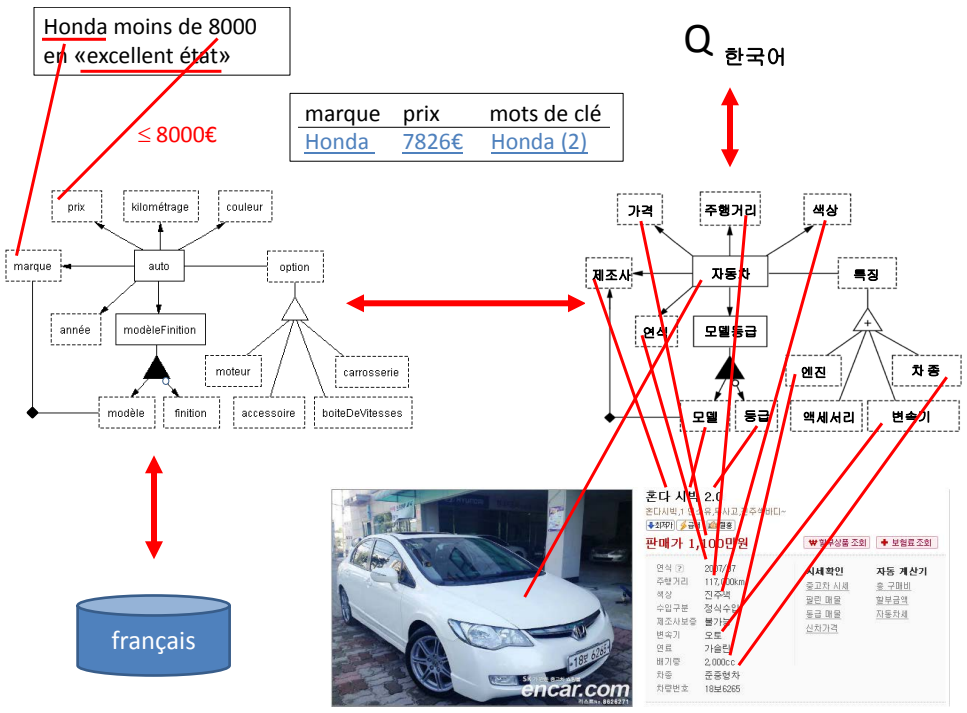


Figure 2: Cross-Language Query Processing.

[ELL+12] D.W. Embley, S.W. Liddle, D.W. Lonsdale, J.S. Park, B.-J. Shin, and A. Zitzelberger. Cross-language hybrid keyword and semantic search. In *Proceedings of the 31st International Conference on Conceptual Modeling (ER 2012)*, pages 190–203, Florence, Italy, October 2012.

[ELLT11] D.W. Embley, S.W. Liddle, D.W. Lonsdale, and Y. Tijerino. Multilingual ontologies for cross-language information extraction and semantic search. In *Proceedings of the 30th International Conference on Conceptual Modeling (ER 2011)*, pages 147–160, Brussels, Belgium, October/November 2011.

[Emb80] D.W. Embley. Programming with data frames for everyday data items. In *Proceedings of the 1980 National Computer Conference*, pages 301–305, Anaheim, California, May 1980.

[EZ10] D.W. Embley and A. Zitzelberger. Theoretical foundations for enabling a web of knowledge. In *Proceedings of the Sixth International Symposium on Foundations of Information and Knowledge Systems (FoIKS'10)*, pages 211–229, Sophia, Bulgaria, February 2010.

[LEL10] D.W. Lonsdale, D.W. Embley, and S.W. Liddle. Ontologies for multilingual extraction. In P. Buitelaar, P. Cimiano, and E. Montiel-Ponsoda, editors, *Proceedings of the World Wide Web Conference's 1st International Workshop on MultiLingual Ontologies (MSW 2010)*, pages 1–4, Raleigh, North Carolina, April 2010.

[SE13] B.-J Shin and D.W. Embley. Reverse engineering relational databases for free-form query answering on the semantic web, 2013. (in preparation).

[ZELS12] A. Zitzelberger, D.W. Embley, S.W. Liddle, and D.T. Scott. HyKSS: Hybrid keyword and semantic search. 2012. (in preparation).