Ontologies for Multilingual Extraction

Deryle W. Lonsdale
Linguistics & English Lang.
Brigham Young University
lonz@byu.edu

David W. Embley
Computer Science
Brigham Young University
embley@cs.byu.edu

Stephen W. Liddle
Information Systems
Brigham Young University
liddle@byu.edu

ABSTRACT

In our global society, multilingual barriers sometimes prohibit and often discourage people from accessing a wider variety of goods and services. We propose multilingual extraction ontologies as an approach to resolving these issues. As envisioned, our ontologies provide a conceptual framework for a narrow domain of interest. Grounding narrow-domain ontologies linguistically enables them to map relevant utterances and text to meaningful concepts in the ontology. Our prior work includes leveraging large-scale lexicons and terminology resources for grounding and augmenting ontological content [12]. Linguistically grounding ontologies in multiple languages enables cross-language communication within the scope of the various ontologies' domains. Technically, we can gauge the success of linguistically grounded ontologies by measuring precision and recall of extracted concepts, and we can gauge the success of automated cross-linguistic-mapping construction by measuring the speed of creation and the accuracy of generated lexical resources.

1. INTRODUCTION

Though English has so far served as the principal language for Internet use (with currently 28.7% of all users), its relative importance is rapidly diminishing. Chinese users, for example, comprise 21.7% of Internet users and their growth in numbers between 2000 and 2009 has been 1,018.7%; the growth in Spanish users has been 631.3% over the last decade. Since more people want to access web information in more languages, this poses a substantial challenge and opportunity for research and business organizations whose interest is in providing multilingual access to web content.

The BYU Data Extraction research Group (DEG)[1] has worked for years on tools—such as its Ontology Extraction System (OntoES)—to enable access to web content of various types: car advertisements, obituaries, clinical trial data, and biomedical information. The group to date has focused on English web data, while understanding the eventual need to extend OntoES to other languages. This appears to be an opportune time for our group to enter the area of multilingual information extraction and show how the DEG infrastructure is poised to make significant contributions in this area as it has already has in extracting English information.

There are currently a few efforts in the area of multilingual information extraction. Some focus on very narrow domains, such as technical information for oil drilling and exploration in Norwegian and English. Others are more general but involve more than two languages, such as accessing European train system schedules. The U.S. government (NIST TREC), the European Union (7th Framework CLEF), and Japan (NTCIR) all have initiatives to help further the development and evaluation of multilingual information retrieval and data extraction systems. Of course, Google and other companies interested in web content and market share are working on ways to provide multilingual access to the Internet.

Almost all of the existing efforts involve a typical scenario that includes: collecting a query in the user's language, translating that query into the language of the web pages to be searched, locating the answers, and then translating the relevant content back into the user's language. This approach is fraught with problems since machine translation (MT), a core component in the process, is still a developing technology.

For reasons discussed below, we believe that our approach has technical and linguistic merit, and can introduce a fresh perspective on multilingual information extraction. Our ontology-based techniques are ideal for extracting content in various languages without having to rely on MT. By carefully developing the knowledge resources necessary, we can extend DEG-type processing to other languages in a modular fashion.

2. THE ONTOLOGY-BASED APPROACH

2.1 Extraction Ontologies

Just over a decade ago, the BYU Data-Extraction research Group (DEG) began its work on information extraction. In a 1999 paper, DEG researchers described an efficacious way to combine ontologies with simple natu-

[1] This work was funded in part by U.S. National Science Foundation grants for the TIDIE (IIS-0083127) and TANGO (IIS-0414644) projects.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
.

ral language processing [4].[2] The idea is to declare a narrow domain ontology for an application of interest and augment its concepts with linguistic recognizers. Coupling recognizers with a conceptual modeling turns a conceptual ontology into an extraction ontology. When applied to text, an extraction ontology recognizes linguistic elements that identify concept instances for the object and relationship sets in the ontology's conceptual model. We call our system *OntoES*, *Ontology-based Extraction System*.

Consider, for example, a typical car ad. Its content can be modeled with a conceptual ontology such as that shown in Figure 1. With linguistic recognizers added for concepts such *Make*, *Model*, *Year*, *Price*, and *Mileage*, the domain ontology becomes an extraction ontology. We have developed a form-based tool [13] that helps users to develop ontologies including declaring recognizers and associating them with ontological concepts. It also permits users to specify regular expressions that recognize traditional value phrases for price such as "$15,900", "7,595", and "$9500"—prices between $100 and $99,999 with optional dollar signs and commas. Users can also declare additional recognizers for other expected price expressions such as "15 grand". To help make recognizers more precise, users can declare exception expressions, left and right context expressions, and units expressions. Users can add keyword phrases such as "MSRP" and "our price" to help sort out various prices that might appear. Applying the recognizers of all the concepts in the car-ads extraction ontology illustrated in Figure 1 to a car ad annotates, extracts, and organizes the facts from that ad.

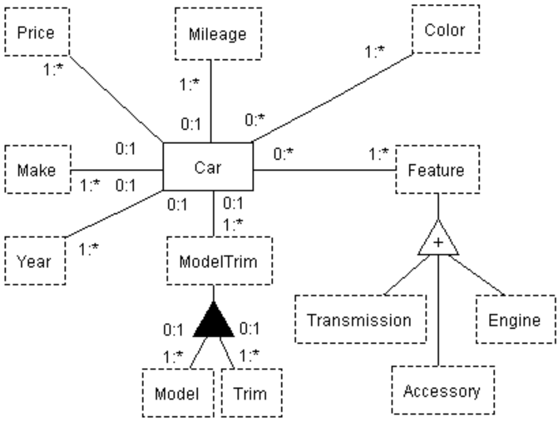

Figure 1: Extraction Ontology for Car Ads.

The result is a machine-readable cache of facts that users can query or use to perform data analysis or other automated tasks. To verify that a carefully designed extraction ontology for car ads can indeed annotate, extract, and organize facts for query and analysis, DEG researchers conducted experiments with hundreds of car ads from various on-line sources containing thousands of fact instances. The OntoES car-ads extraction ontology was able to correctly extract fact instances for concepts with recall measures of almost 95% and precision measures nearing 100% [5].

Recently, DEG researchers have experimented with information extraction in Japanese. Figure 2 shows an OntoES extraction ontology that can extract information from Japanese car ads analogous to the English one shown earlier. The concept names are in Japanese as are the regular-expression recognizers. Yen amounts range from 10,000 yen to 9,999,999 yen rather than $100 to $99,999. The critical observation, however, is that the structure of the Japanese ontology is identical to the structure of the English ontology. This provides a cross-linguistic bridge through concepts rather than through traditional means of translation.

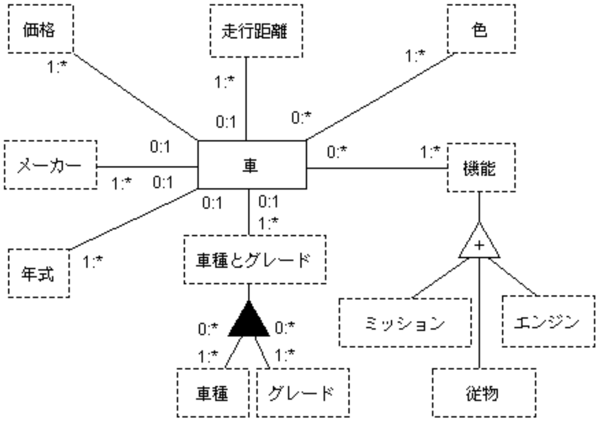

Figure 2: Japanese Extraction Ontology for Car Ads.

As currently implemented, OntoES extraction ontologies can "read" and "write" in any single language. The car-ad examples here are in English and Japanese, but extraction ontologies work the same for all languages. To "read" means to recognize instance values for ontological concepts, to extract them, and to appropriately link related values together in accord with the interrelationships among the concepts in the ontology and in accord with the constraints of the ontology. To "write" means to list the facts recorded in the ontological structure. Having "read" a typical car ad, OntoES might write:

Year: 1984
Make: Dodge
Model: W100
Price: $2,000
Feature: 4x4
Feature: Pickup
Accessory: 12.5x35" mud tires

[2] Recently, others have begun to combine ontologies with natural language processing [10, 2]. The combination has become known as "linguistically grounding ontologies."

In addition, based on the constraints, OntoES knows and can write several meta statements about an ontology. Examples: "an $Accessory$ is a $Feature$" (the white triangle denotes a hyponym/hypernym is-a constraint); "$Trim$ is part of $ModelTrim$" (the black triangle denotes a meronym/holonym is-part-of constraint), "Car has at most one $Make$" (the participation constraint 0:1 on Car for $Make$ denotes that Car objects in car ads associate with $Make$ names between 0 and 1 times, or "at most once").

As currently implemented, however, OntoES cannot read in one language and write in another. This cross-linguistic ability to read in one language and then translate to and write in another language is the essence of our multilingual-oriented development. For example, we expect to be able to read the price in yen from a Japanese car-ad and write "Price: $24,124" and to read the Kanji symbols for the make and write "Make: Mitsubishi". To assure this level of functionality, we need to encode unit or currency conversion routines for values like $Price$ and to encode cross-linguistic lexicons for named entities such as $Make$. In principle, encoding this cross-linguistic mapping is currently possible, but represents a fair amount of manual effort. We are currently finding ways to largely automate this mapping.

Before discussing ideas for semi-automatically creating cross-linguistic mapping, however, we mention some ongoing research work on OntoES itself that will enable it to more fully play its role in the overall goal of facilitating cross-linguistic information extraction and query processing. Two additions appear immediately useful: compound recognizers and patterns.

- Compound Recognizers. We are augmenting OntoES to not only directly recognize ontological concepts but also to directly recognize ontological relationships. Relationship recognition requires the addition of compound recognizers—recognition expressions that depend on other recognition expressions. For example, consider extracting the *between* constraint from the request "Find Nissans for sale with years between 1995 and 2005." Recognizing the *between* constraint requires not only recognizing the relationship designator *between* but also its referents. Recognizing the referents requires a year recognizer. Thus, the full *between* recognizer is compound since successful recognition depends on successful recognition for its referents. DEG researchers have considered compound recognizers for operators in free-form queries [1], but much research remains to fully linguistically ground ontological relationships.

- Patterns. We are augmenting OntoES to identify and extract from patterned text. For example, car ads are often structured as a table with $Price$ in one column, $Year$ in another column, and $Make$ and $Model$ in a third column. After recognizing a patterns in documents, we can apply specialized extraction rules and likely improve extraction accuracy. We have worked some with table patterns [7], but much remains to be done to fully exploit patterns in text.

2.2 Multilingual Mappings

We are extending in a principled way the cross-linguistic effectiveness of our OntoES system by adapting it for users of non-English languages. Though the OntoES system was originally designed to handle English-language documents, it was implemented according to state-of-the-art software engineering principles and best practices. Consequently, we anticipate that internationalization of the system should be relatively straightforward, not requiring wholesale rewrites of crucial components. For example, the character representation used throughout the OntoES system is UTF-8 (a standard encoding for Unicode, a representation designed for almost all known human writing systems). This should allow us to handle web pages in any language, given appropriate linguistic knowledge sources. Since OntoES does not need to parse out the grammatical structure of webpage text, only lower-level lexical (word-based) information is necessary for linguistic processing.

The system's lexical knowledge is highly modular, with specific resources encoded as user-selectable lexicons. The information used to build up existing content for the English lexicons includes a mix of implicit knowledge and existing resources. Some lexicon entries were created by students during class and project work; other entries were developed from existing lexical resources (e.g. the US Census Bureau for personal names, the World Factbook for country names, Ethnologue for language names, etc.). We are developing analogous lexicons for other languages, and adapting OntoES as necessary to accommodate them in its processing. As was the case for English, this involves some hand-crafting of relevant material, as well as finding and converting existing data sources in other languages for targeted types of lexical information. Often this is relatively straightforward: for example, WordNet is a sizable and important component for English OntoES, and similar and compatible resources exist for other languages. However, we also need to rely on linguistic knowledge and experience to find, convert, and implement appropriate cross-linguistic lexical resources.

In the realm of cross-linguistic extraction systems, OntoES has a clear advantage. We claim that ontologies, which lie at the crux of our extraction approach, can serve as viable interlinguas. We are currently substantiating this claim. Since an ontology represents a conceptualization of items and relationships of interest (e.g. interesting properties of a car, information needed to set up a doctor's appointment, etc.), a given ontology should be appropriate cross-linguistically with perhaps occasionally some slight cultural adaptation. For example, in our prior work on extraction from obituaries [4] we found that worldwide cultural and dialect differences were readily apparent even in English material. Certain terms for events like "tenth day kriya", "obsequies", and "cortege" were found only in English obituaries announcing events outside of America. Since our lexical resources serve as a "grounding" of the lowest-level concepts from ontologies with the lexical content of the web pages, substituting one language's lexicon for another's provide OntoES a true cross-linguistic capability. There is no need for MT, the most currently used technique for cross-linguistic information retrieval and is at best only helpful for gisting webpage content.

2.3 Ongoing Work

The work we are engaged in as described in this position paper involves several separate but related tasks. We are locating annotated corpora in other languages that would be amenable for evaluation purposes, and collecting and annotating interesting multilingual web material of our own. We are also developing prototype lexicons and recognizers for these target languages. Of course, our work requires us to develop and adapt prototype ontologies for target languages for sample concepts in data-rich domains.

In addition, we are enhancing extraction ontologies by enabling them to (1) explicitly discover and extract relationships among object instances of interest, and (2) discover patterns of interest from which they can more certainly identify and extract both object instances and relationship instances of interest. This involves devising, investigating, designing, coding, and evaluating algorithms for compound recognizers and for pattern discovery and patterned information extraction.

Finally, we will be evaluating performance of the system using standard metrics and gold-standard annotated data.

3. CONCLUSION

Though an interesting effort in its own right, we expect our multilingual extraction work to also contribute to our larger effort to create a Web of Knowledge [6, 8]. Our research centers around resolving some of tough technical issues involved in a community-wide effort to deploy the semantic web [14] and in concert with efforts at Yahoo!, Google, and elsewhere to extract information from the web and integrate it into community portals to enable community members to better discover, search, query, and track interesting community information [3, 9, 11]. Multilingual extraction ontologies have the far-reaching potential to play a significant role as semantic-web work finds its way into mainstream use in global communities.

4. REFERENCES

[1] M. Al-Muhammed and D. Embley. Ontology-based constraint recognition for free-form service requests. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE'07)*, pages 366–375, Istanbul, Turkey, April 2007.

[2] 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.

[3] P. DeRose, W. Shen, F. Chen, A. Doan, and R. Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. In *Proceedings of the 33rd Very Large Database Conference (VLDB'07)*, pages 23–28, Vienna, Austria, September 2007.

[4] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y.-K. Ng, and R. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, November 1999.

[5] D. Embley, D. Campbell, S. Liddle, and R. Smith. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98)*, pages 52–59, Washington D.C., November 1998.

[6] D. Embley, S. Liddle, D. Lonsdale, G. Nagy, Y. Tijerino, R. Clawson, J. Crabtree, Y. Ding, P. Jha, Z. Lian, S. Lynn, R. Padmanabhan, J. Peters, C. Tao, R. Watts, C. Woodbury, and A. Zitzelberger. A conceptual-model-based computational alembic for a web of knowledge. In *Proceedings of the 27th International Conference on Conceptual Modeling (ER08)*, pages 532–533, Barcelona, Spain, October 2008.

[7] D. Embley, C. Tao, and S. Liddle. Automating the extraction of data from HTML tables with unknown structure. *Data & Knowledge Engineering*, 54(1):3–28, July 2005.

[8] D. 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 (FoIKS10)*, Sophia, Bulgaria, February 2010. (to appear).

[9] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, March/April 2009.

[10] L. Hunter, Z. Lu, J. Firby, W. B. Jr., H. Johnson, P. Ogren, and K. Cohen. OpenDMAP: An open source, ontology-driven, concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9(8), 2008.

[11] R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *Proceedings of the 2009 Symposium on Principles of Database Systems*, pages 1–12, Providence, Rhode Island, June/July 2009.

[12] D. Lonsdale, D. W. Embley, Y. Ding, L. Xu, and M. Hepp. Reusing ontologies and language components for ontology generation. *Data & Knowledge Engineering*, 69:318–330, 2010.

[13] C. Tao, D. Embley, and S. Liddle. FOCIH: Form-based ontology creation and information harvesting. In *Proceedings of the 28th International Conference on Conceptual Modeling (ER 2009)*, pages 346–359, Gramado, Brazil, November 2009.

[14] W3C (World Wide Web Consortium) *Semantic Web Activity Page*. http://www.w3.org/2001/sw/.