Multilingual Extraction Ontologies

David W. Embley, Department of Computer Science, BYU
Stephen W. Liddle, Information Systems Department, BYU
Deryle W. Lonsdale, Department of English and Linguistics, BYU
Yuri Tijerino, Department of Applied Informatics,
Kwansei Gakuin University, Kobe-Sanda, Japan

Abstract

In our global society, multilingual barriers sometimes prohibit and often discourage international travelers from receiving desired goods and services. In a multilingual setting, easing linguistic-based tasks such as finding a restaurant, learning how to use local mass transit, and making simple business transactions would help remove barriers and encourage international travelers to seek and enjoy a wider variety goods and services. We propose multilingual extraction ontologies accessible via mobile devices as an aid to resolving these issues. Multilingual ontologies can communicate in multiple languages with the wireless local web as well as the international traveler and can thus break through language barriers and allow access to local markets and services. As envisioned, 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. Linguistically grounding ontologies in multiple languages enables cross-language communication. 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 crosslinguistic-mapping construction by measuring the speed of creation and the accuracy of generated lexical resources. From a business perspective, we can measure success in at least two ways. First, tech-transfer success ultimately generates measurable royalty revenue and a user base that demonstrates whether the technology is successful at a consumer level. Second, we can see how well a business plan built on our core technology performs in student competitions hosted at BYU and other universities.

We envision an interdisciplinary mentored research environment as an ideal way to resolve the myriad of technical problems involved in building the envisioned collection of multilingual extraction ontologies. While a computer science student leverages new algorithms and data structures to construct linguistically grounded domain ontologies and a linguistics student finds cost-effective ways to mesh multiple languages with respect to various ontologies, an entrepreneurial information-systems student can investigate tech-transfer opportunities and resolve technical issues between a research prototype and an industrial-strength, deployable system. In addition to being interdisciplinary the PIs have expert foreign language knowledge. Each BYU PI speaks one or more foreign languages (German, Spanish, French, and Japanese). Further, the involvement of Professor Tijerino, who will be on sabbatical leave at BYU from Kwansei Gakuin University in Japan, along with his two native Chinese graduate students provide the necessary international constituents that comprise the mentoring research team.

1 Introduction

Although 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) 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, such as the multilingual European train schedules, are coded to directly support several languages. 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 also 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, a core component in the process, is still a developing technology.

For reasons discussed below, we believe that the DEG approach holds much promise 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 machine translation. By carefully developing the knowledge resources necessary, we can extend DEG-type processing to other languages in a modular fashion.

We believe that our approach has technical and linguistic merit, but we also think that it has potential to underpin new products that can compete in the rapidly expanding mulilingual information access market. Part of the work in this project will be to quantify our expectations and build up a business plan around the technology we have developed.

Students will be able to do much of this work, and the mentoring environment we propose is ideal to carry it out. The mentoring environment includes faculty advisors from the three interdisciplinary areas needed for success (computer science, linguistics, and business), and it also has needed international collaborators. A colleague, Professor Yuri Tijerino, from a Kwansei Gakuin University in Kobe, Japan, will be on-campus next Summer as a visiting scholar. Currently, we are already collaborating with him and his students (some of whom recently visited BYU). They have expertise that will nicely complement those of the DEG group, making the mentoring environment international and truly ideal for the proposed project.

2 Project Description

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 natural-language processing [ECJ+99].[1] 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 model 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, the car ads in Figure 1 and the conceptual ontology modeling the facts typically found in a car ad in Figure 2. With linguistic recognizers added for concepts such as *Make*, *Model*, *Year*, *Price*, and *Mileage*, the domain ontology becomes an extraction ontology. Figure 3 shows a screenshot of a tool that lets users declare recognizers and associate them with ontological concepts. The screenshot shows a regular expression that recognizes 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, as the labeled, empty, text fields in Figure 3 indicate. Users can also 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 Figures 2 and 3 to the car ads in Figure 1 annotates, extracts, and organizes the facts in Figure 1. The result is a machine-readable bundle 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% [ECLS98].

Recently, DEG researchers have experimented with information extraction in Japanese. Figures 4 and 5 show an OntoES extraction ontology that can extract from Japanese car ads like the one in Figure 6. 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 crosslinguistic bridge through ontological concepts rather than through traditional means of translation.

As currently implemented, OntoES extraction ontologies can "read" and "write" in any single language. Although the examples here are in English and Japanese, 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 them in accord with the ontological constraints and interrelationships among concepts. To "write" means to list facts recorded in the ontological structure. Having "read" the first car ad in Figure 1, OntoES could "write":

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

[1] Recently, others have begun to combine ontologies with natural-language processing [HLF+08, BCHS09]. The combination has become known as "linguistically grounding ontologies."

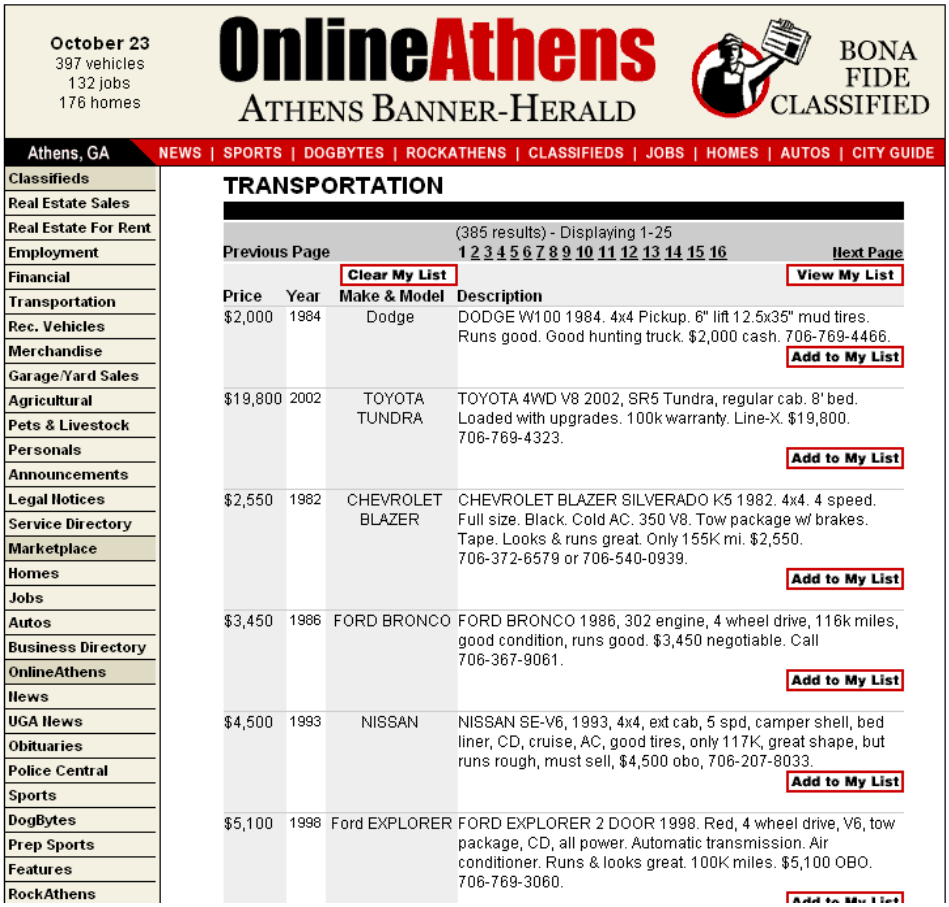

Figure 1: Online Car Ads from the Athens Banner-Herald.

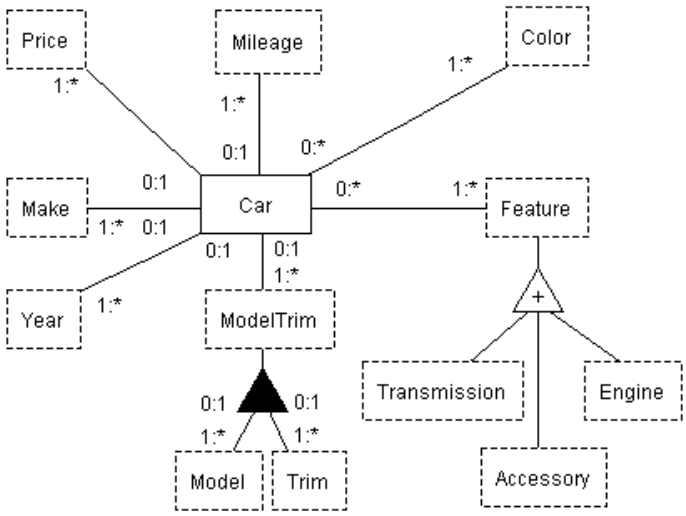

Figure 2: Extraction Ontology for Car Ads.

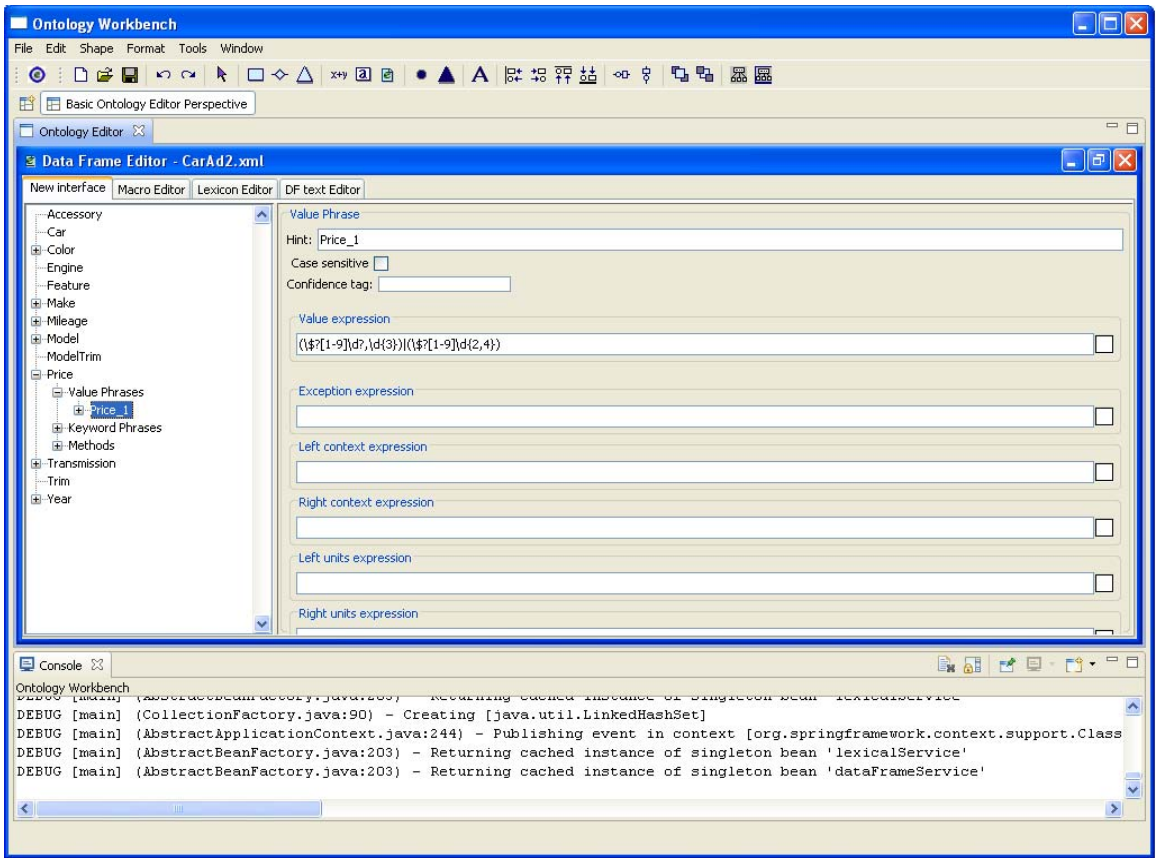

Figure 3: Linguistically Grounding the Price Concept with a Value Recognizer.

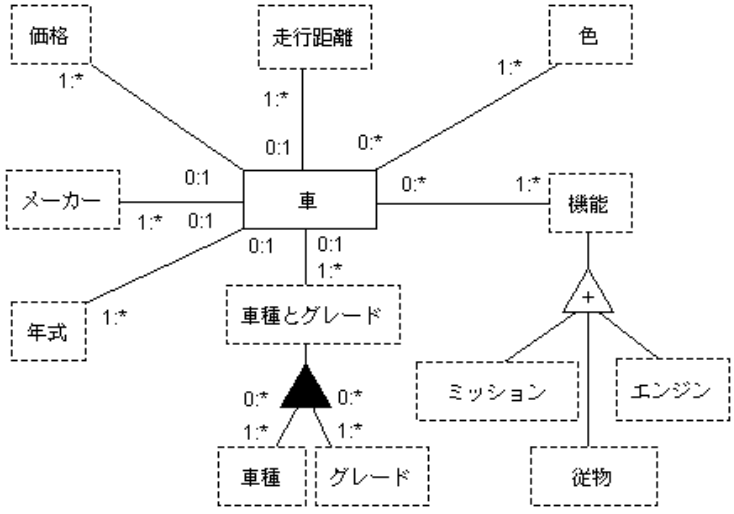

Figure 4: Japanese Extraction Ontology for Car Ads.

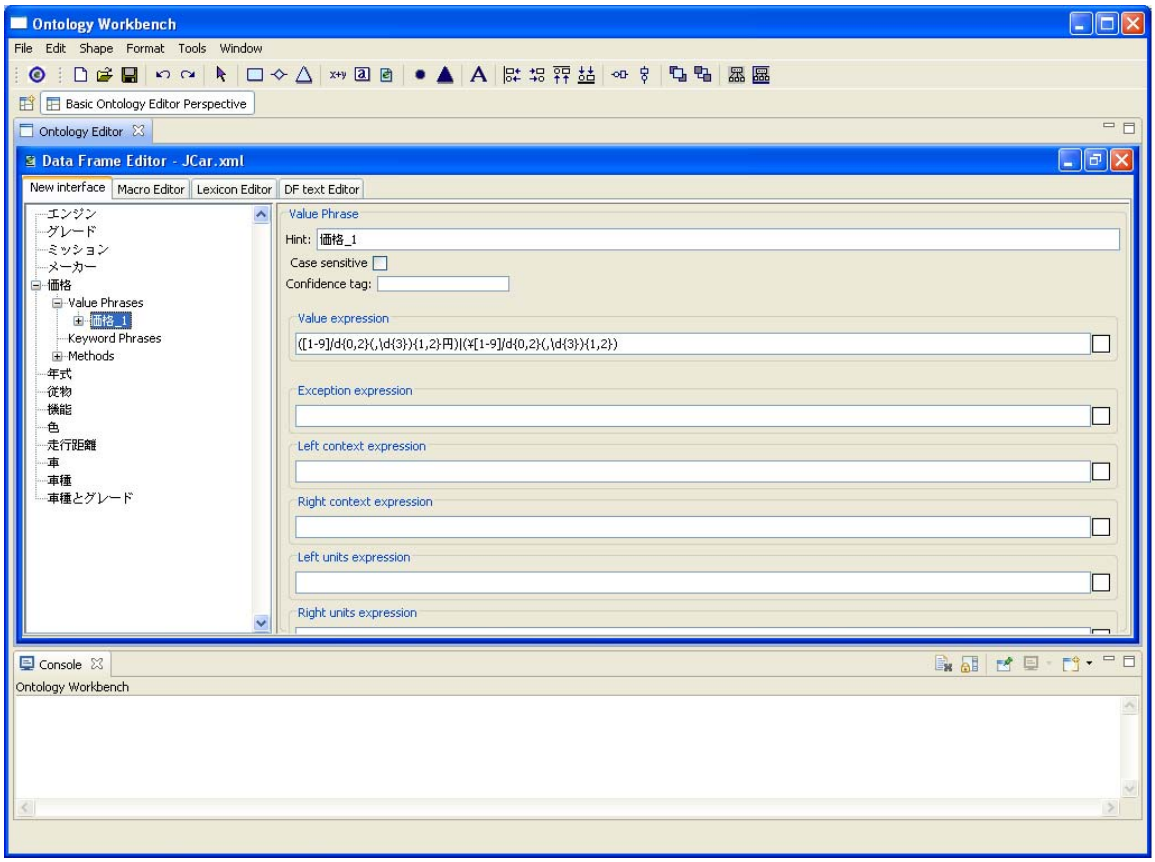

Figure 5: A Japanese Price Value Recognizer.

Figure 6: An Online Japanese Car Ad from auction.goo-net.com.

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); "*Transmission*, *Accessory*, and *Engine* features are distinct" (the + symbol in the white triangle makes the hyponym sets mutually exclusive); "*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 crosslinguistic ability to read in one language and then translate to and write in another language is the essence of our MEG proposal. For example, we want OntoES to "read" the price in yen from the car-ad in Figure 6 and "write" "Price: $24,124" and to "read" the Kanji symbols for the make and "write" "Make: Mitsubishi". To succeed, we need to encode unit or currency conversion routines for values like price and to encode crosslinguistic lexicons for named entities such as make. In principle, encoding this crosslinguistic mapping is currently possible, but represents a fair amount of manual effort. We believe, however, that a linguistics student can find ways to largely automate the construction of this mapping. (In the next section of this proposal, we give some ideas about possible research directions in pursuit of this goal.)

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

1. Compound Recognizers. We propose to augment OntoES to not only directly recognize ontological concepts but to also 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 [AME07], but much research remains to fully linguistically ground ontological relationships.

2. Patterns. We propose to augment OntoES to identify and extract from patterned text. The car ads in Figure 1 are in a table with *Price* in one column, *Year* in another column, and *Make* and *Model* in a third column. If we can recognize a patterns in documents, we can apply specialized extraction rules and likely improve extraction accuracy. DEG researchers have worked some with table patterns [ETL05], but much remains to be done to fully exploit patterns in text.

2.2 Multilingual Mappings

For this MEG project, we propose to extend in a principled way the crosslinguistic effectiveness of our OntoES system by adapting it for users of non-English languages. This process, called internationalization (i18n) is a well-defined area of software development that lies at the nexus of computer science, linguistics, business, and marketing.

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 the i18n 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, and the Ethnologue for language names). Our MEG work will in part involve developing analogous lexicons for other languages, and adapting OntoES as necessary to accommodate them in its processing. As was the case for English, this will involve 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 will be relatively straightforward: for example, WordNet is a sizable and important component for English OntoES, and similar and compatible resources exist for other languages. We will, though, also need to rely on linguistic knowledge and experience to find, convert, and implement appropriate crosslinguistic lexical resources.

We plan to consider techniques such as those reported in articles on semi-automatically constructing crosslinguistic lexical resources [LYY02, XGP$^+$09]. Further, some resources already exist, and hence we will be able to easily adapt them for our purposes. For example, OntoES uses WordNet [Fel98] as a source of lexical information, so with our prior experience, plugging in equivalent resources in other languages (EuroWordNet for European languages, HowNet for Chinese, Japanese WordNet for Japanese) will be straightforward. Additionally, though, more domain-specific vocabulary from online word lists [KT95], terminology banks [LDEM02], and parallel or comparable text corpora can be used to mine crosslinguistic lexical equivalents. We are in a position to exploit these current alignment techniques for building lexical equivalents, particularly those centered around ontology-based knowledge representations [NCF02].

In the realm of crosslinguistic 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. Our MEG project would substantiate this claim. Since an ontology represents a conceptualization of items and relationships of interest (e.g., interesting properties of a car and information needed to set up a doctors appointment), a given ontology should be appropriate crosslinguistically with perhaps rarely some slight cultural adaptation. 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 languages lexicon for anothers should provide OntoES with a true crosslinguistic capability. There is no need to do machine translation, which is the most currently used technique for crosslinguistic information retrieval and is at best only helpful for gisting webpage content.

2.3 Business Opportunities

Potential business opportunities for applying the technology of multilingual extraction ontologies are plentiful. Consider the case of a casual tourist who might travel with the type of information provided by a travel guide such as *Frommer's* or *Fodor's*. Usually such a traveler will have general background material on an area along with recommended highlights to include in the itinerary. A typical tourist will have made most hotel reservations ahead of time and will have a general plan

regarding what to see when. But along the way there will be many opportunities for eating meals, shopping, and discovering local treasures "off the beaten path." A traveler's assistant that is good at finding information about local services delivers significant business value, which implies that some tourists would be willing to pay for such a tool.

Non-tourist travelers have many of the same needs as tourist travelers. For example, everyone needs to eat on a regular basis, and travelers need to move from point to point along their journey. It is common for travelers to make reservations for air travel but to arrange ground transportation on the spot. Travelers who wish to move economically by train and bus may find it difficult to understand how to navigate the foreign mass transit system. A good multilingual traveler's assistant could resolve misunderstandings and enable more travelers to confidently use available mass transit. The cost savings and flexibility available to someone who can use public transportation creates business value that could be monetized.

The list of potential applications for travelers is endless, but to name a few: arranging transportation or navigating a mass transit system; choosing and visiting restaurants, grocery stores, and hotels; arranging longer-term accommodations such as an apartment for a month; assisting with the location and negotiation of souvenir purchases; and more generally, helping a traveler understand the linguistic and cultural context of a foreign country *in situ*.

In our vision, a "traveler's assistant" is a software application that has multilingual extraction ontologies at its core. A server program scours the web finding information sources related to particular locations. The system uses multilingual extraction ontologies to arrange useful information in an understandable way. Then a local application (say, an iPhone application that has location awareness built into it) contacts the server to search the pre-arranged information and provide helpful translations as needed. Information could be provided on a subscription basis or using an advertizing-supported model. Given suitable technology, it is not overly difficult to create a business plan to take advantage of this opportunity.

3 Research Plan and Mentoring Environment

3.1 Research Plan

During the Winter semester, after the announcement of MEG grant recipients, we plan to recruit students: one upper-class undergraduate from each discipline—computer science, linguistics, and entrepreneurial e-business. Beginning soon thereafter and throughout the summer, we plan to mentor each student in carrying out the following activities.

- Extraction-ontology enhancements (computer science student):

    1. Install our Eclipse software development environment, gain access to OntoES under the project-change control system, and secure rights to update and modify the code base.
    2. Understand the architecture of OntoES and determine how new code for compound recognizers and for patterns integrates with the current system.
    3. Design and implement algorithms for compound recognizers.
    4. Design and run experimental validation tests for compound recognizers.
    5. Design and implement algorithms for pattern recognition and extraction processing.
    6. Design and run experimental validation tests for patterns.

- Multilingual mapping development (linguistics student):

    1. Become familiar with existing OntoES lexical knowledge sources and content.
    2. Locate analogous sources of lexical knowledge in the languages to be developed.
    3. Adapt existing natural language tools to process and extract target lexical content.
    4. Convert extracted lexical knowledge for compatibility with the OntoES framework.
    5. Work with OntoES programmers to assure proper functionality of knowledge sources.
    6. Help advise OntoES programmers on necessary adaptations to ontologies for multilingual and multicultural processing.
    7. Help in evaluation of the system's performance and coverage

- Business plan development (tech entrepreneurship student):

    1. Research the needs of international travelers to understand their desires, challenges, and frustrations.
    2. Brainstorm business ideas that could use multilingual extraction ontologies as the core technology for addressing international travel needs.
    3. Perform market research on the various ideas and determine which would be the best two or three opportunities to develop.
    4. Recruit team members who are interested in competing in the various business plan, idea pitch, and other entrepreneurial competitions available to students through the Rollins Center for Entrepreneurship and Technology.
    5. Write business plans and supporting materials to present the chosen opportunities for business competitions.
    6. Participate in the various competitions.
    7. Work with BYU's Technology Transfer Office to generate a patent application for the relevant underlying technology.

During the Fall semester, we plan to co-author, with the students involved, academic papers reporting our experimental evaluation of extraction-ontology enhancements and reporting our performance evaluation of multilingual mapping development. We also plan, in conjunction with the students involved, to work with the tech-transfer office to generate a patent application. Furthermore, we will work with the business student to present the business plan in at least one student competition. The Rollins Center has a slate of supporting activities year-round, leading up to the main Business Plan Competition in the Winter semester. We will mentor the business team through this process.

3.2 Mentoring Environment

We believe we can create a wonderful cross-disciplinary mentoring environment that will be ideal for the students involved in carrying out this project. The cross-disciplinary nature of our proposal will naturally expose students to experts and exemplars in several disciplines, thus broadening the students' outlook, their understanding of the target disciplines and their professional networks. These characteristics will help the students accelerate their learning while simultaneously giving a boost to their career prospects.

Students involved in this project will be integrated into our lab that includes both undergraduate and graduate students. Through daily interactions with other students, we expect they will build strong peer relationships. The faculty will continue to hold hold regular weekly research and development meetings (1) with the entire research group and (2) at a separate time with each individual student, thus providing regular contact points for group-oriented and individual attention. In these meetings there will be many opportunities for professional and personal mentoring. We will also sponsor social gatherings that offer a more relaxed setting for interaction. Students will be partners with faculty members in a range of activities and thus will be able to observe how faculty handle day-to-day matters, high-pressure deadlines, and preparation for public presentation of research ideas. We begin our research meetings with prayer, highlighting the special nature of the environment we have at BYU.

A further benefit available to our group is through the founders organization that supports the Rollins Center for Entrepreneurship and Technology (CET), of which Liddle is the academic director. CET founders are successful entrepreneurs who are interested in returning to BYU and giving back. They are men and women of character and professional ability who love to build students and help the faculty deliver a more powerful and influential experience to the students. We will be able to leverage the resources of the CET to provide significant support to the business plan portion of this project. There are numerous events throughout the year where we will be able to put our students in contact with CET founders for an enhanced mentoring experience. Students will see successful individuals who model career success combined with lives of faith and integrity.

Given the multilingual nature of our project, there will also be opportunities for students to build their international skill sets and professional networks. We anticipate hosting the two Chinese students from Tijerino's university in Japan in a second visit to BYU near the end of Summer 2010. This could lead to the possibility of arranging a visit for our students to Japan as well.

We believe that students become significantly more capable as they blend their knowledge and experience across multiple dimensions. We want our students to excel at the core academic research that leads to top-quality publications while simultaneously learning from the challenges that arise when applying research findings in business. We want our students to learn how to balance the competing demands of work, family, church, and community. To accomplish this, our research plan includes elements of core academic work in an environment where students will see their ideas through to the consumer level, while working with mentors who model the attributes we value. Whether our students choose an academic or a professional career, this experience will be a powerful combination of activities that will broaden their cross-disciplinary horizons.

4 Expected Significance

For the intellectual merit of the proposed work, we expect to be able to make the following contributions.

- Enhance 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. (For mentoring, we plan to have a computer science student devise, investigate, design, code, and evaluate algorithms for compound recognizers and for pattern discovery and patterned information extraction.)

- Find ways to efficiently create crosslinguistic mappings for lexicons and language-value recognizers and demonstrate the viability of our ontologies as a core component for multilingual

data extraction. (For mentoring, we plan to have a linguistics student work on creating, adapting, and deploying language resources in the OntoES system and performing linguistic evaluations of the systems functionality.)

- Develop a patent application for multilingual extraction ontologies together with a business plan that leverages the patent-pending technology to deliver value to consumers. (For mentoring, we plan to have an information-systems or tech-entrepreneurship student work with BYU's Technology Transfer Office on the patent application process and on the business plan. The student would likely assemble an extended team to support presenting the business plan in the various related BYU competitions. If the plan does well, there will be opportunities to present the plan at non-BYU competitions as well.)

For the broader impact of the proposed work, we expect the work to contribute to our larger effort to create a Web of Knowledge [ELL+08, EZ10]. Our research centers around resolving some of tough technical issues involved in a community-wide effort to deploy the semantic web [W3C]. It also coincides 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 [DSC+07, HNP09, KPR+09]. 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.

References

[AME07] M. Al-Muhammed and D.W. 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.

[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.

[DSC+07] 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.

[ECJ+99] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, November 1999.

[ECLS98] D.W. Embley, D.M. Campbell, S.W. Liddle, and R.D. 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.

[ELL+08] D.W. Embley, S.W. Liddle, D. Lonsdale, G. Nagy, Y. Tijerino, R. Clawson, J. Crabtree, Y. Ding, P. Jha, Z. Lian, S. Lynn, R.K. 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*, pages 532–533, Barcelona, Spain, October 2008.

[ETL05] D.W. Embley, C. Tao, and S.W. Liddle. Automating the extraction of data from HTML tables with unknown structure. *Data & Knowledge Engineering*, 54(1):3–28, July 2005.

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

[Fel98] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachussets, 1998.

[HLF+08] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. 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.

[HNP09] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, March/April 2009.

[KPR+09] 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.

[KT95] J. Klavans and E. Tzoukermann. Combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation*, 10(2):185–218, September 1995.

[LDEM02] D.W. Lonsdale, Y. Ding, D.W. Embley, and A. Melby. Peppering knowledge sources with SALT: Boosting conceptual content for ontology generation. In *Proceedings of the AAAI Workshop: Semantic Web Meets Language Resources*, pages 30–36, Edmonton, Alberta, Canada, July 2002.

[LYY02] Y. Liu, S. Yu, and J. Yu. Building a bilingual wordnet-like lexicon: The new approach and algorithms. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*, pages 1–5, Taipei, Taiwan, August/September 2002.

[NCF02] G. Ngai, M. Carupat, and P. Fung. Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, August/September 2002.

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

[XGP+09] R. Xu, Z. Gao, Y. Pan, Y. Qu, and Z. Huang. An integrated approach for automatic construction of bilingual Chinese-English wordnet. In *Proceedings of the Third Asian Semantic Web Conference*, pages 302–314, Bangkok, Thailand, February 2009.