Building Ontologies from Collaborative Knowledge Bases to Search and Interpret Multilingual Corpora

Yegin Genc **Elizabeth A. Lennon** **Winter Mason** **Jeffrey V. Nickerson**
Stevens Institute of Technology
Center for Decision Technologies
Castle Point on Hudson, Hoboken, NJ USA
{ygenc, elennon, wmason, jnickerson}@stevens.edu

Abstract

Tools and techniques that automate the interpretation of multilingual corpora are useful on many fronts; scholars, as an example, could use such tools to more readily pinpoint relevant articles from journals in a wide variety of languages. This work describes techniques to build and characterize ontologies using collaborative knowledge bases, e.g., Wikipedia. These ontologies can then be used to search and classify texts. Originally developed for monolingual corpora, we extend the approach to multilingual texts and test the methods with Mandarin scientific abstracts. The presented techniques provide a novel and efficient mechanism to obtain contextually rich ontologies and measure document relevancy within multilingual corpora.

1 Introduction

The wealth of data available online in the form of unstructured text drives the development of tools that automatically extract meaning from cross-lingual corpora. Techniques that quantify the degree to which texts exhibit similar meaning improve a variety of search processes – for example, academic research. However, automating the interpretation of multilingual corpora requires detecting similarities in meaning, while ignoring irrelevant linguistic differences. For example, the understanding that emerges from the connections and associations among words, i.e. context, can manifest very differently in different languages (Goddard, 2011). Furthermore, the meanings of words used in natural language are often context dependent, and context itself both shapes and reveals meaning (Gennaro et al., 2007).

For the purposes of this work, an ontology is defined as a model that represents word entities as concepts and their interrelationships (Lanzenberger et al., 2010). In this sense, ontologies represent the relevant aspects of context. To effectively comprehend cross-lingual corpora, tools that can explore the dependencies between language and context are needed.

One way to do this is to make use of well-understood existing texts that have explicitly linked concept graphs. Examples of such texts are collaborative knowledge stores, databases built up through the contributions of many individuals.

The techniques described here use Wikipedia to build ontologies from journal article abstracts in different languages, which we test on text written in Mandarin. In order to compare alternative ways of deriving ontologies, a set of articles that have both Mandarin and English abstracts are used as the test corpus.

The rest of the paper is organized into four sections. The background section briefly summarizes prior research relevant to this work. Next, the methods section details the processing steps used to create and visualize the ontologies for three experimental conditions. Sample ontology visualizations for each of the experimental conditions are shown. A discussion comparing some of the emergent features in each of the three generated ontologies follows. Finally, we outline next steps for the extension of these techniques.

2 Background

Translation is used to convey the meaning represented in one language in another language. Automated text translation was a goal of early computing (Locke and Booth, 1955), and is still challenging today. Approaches taken include dictionary look-ups, cognate matching, and parallel corpora based methods (Kishida, 2005). Cognate matching uses untranslatable terms such as proper nouns or technical terminology as the bases of cross-lingual connections. For example, Freitas-Juniar et al. (2006) leveraged medical terms, commonly used across languages, to classify medical documents from multiple languages.

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 87–94,
Sofia, Bulgaria, August 8, 2013.

Landauer and Littman (1991) used parallel corpora based methods when they created a language independent indexing space via Singular Value Decomposition to generate a comparable corpus. This permitted texts to be represented in a language-independent space, solely using the terms of the presentation language.

One early machine translation system, DIONYSUS, used three static knowledge sources: a lexicon, an ontological domain model, and a text-meaning-representation language in an effort to automate translation. The DIONYSUS researchers noted the challenge of developing an ontology based on a detailed version of a "constructed reality" (Onyshkevych and Nirenburg 1992). In other words, an ontological model of concepts representing a worldview is only as good as its ability to capture the breadth and depth of the world it attempts to model. Creating ontologies for machine translation applications arguably require knowledge stores as rich, expansive, and comprehensive as human language itself (Hovy, 2005).

One challenge related to reliable ontology creation is the relevance of the produced ontology in the future (Hovy, 2005). That is, word meanings morph over time, and so the ontology needs to shift also. Moreover, shifts in word meanings happen differently in different languages. Nichols et al. (2006) explored multilingual ontology acquisition using robust minimal recursion semantics and machine-readable dictionaries. Though they demonstrated a language-agnostic tool for automated ontology generation, it was still limited to the static database of words contained in the dictionaries.

Attempting to overcome the limitations of dictionaries, Gabrilovich and Markovitch (2009) turned to Wikipedia to perform what they called *explicit semantic analysis* (ESA). They drew upon both the reference and contextual knowledge embedded throughout Wikipedia with the goal of outperforming statistical methods, like latent semantic analysis (LSA), in computing semantic relatedness of texts (Gabrilovich and Markovitch, 2009). However, in explicit semantic analysis, the semantic interpreter, which consists of weighted lists of concepts, i.e. Wikipedia articles, is built directly from Wikipedia's text, a time-consuming process. Sorg and Cimiano (20 - 12) developed an approach leveraging explicit semantic analysis for cross-lingual information retrieval using Wikipedia.

Building on the premise that collaborative knowledge stores, like Wikipedia, are superior for semantic-analysis related tasks, other researchers have mapped extracted word entities from Twitter tweets directly to the titles of Wikipedia pages. The reported technique outperformed statistically-based, semantic categorization methods, specifically LSA and string-edit-distance (Genc et al. 2011). In addition, the approach could categorize concepts in short text strings, a widely known challenge in semantics (Michelson and Macskassy, 2010). In addition, using the Wikipedia title pages instead of the actual article content enabled a faster semantic transform (Genc et al. 2012). Mapping extracted entities to online collaborative knowledge bases, like Wikipedia, also presents a path to accessing an ever-relevant contextual framework based upon the most current human knowledge base (Michelson and Macskassy 2010).

3 Methods

This study compares simplified Chinese Wikipedia and English Wikipedia in their resourcefulness to build ontologies. For the comparison, we used a sample abstract that is available in both Mandarin and English (Figure 1). We constructed ontologies from our sample using both Chinese and English Wikipedia according to the experimental conditions detailed in section 3.2.

3.1 Text Segmentation and Entity Extraction

To extract entities, atomic, meaningful elements of text, we first segmented the texts into phrases – single words, bi-grams, and tri-grams – that overlap in a sliding window fashion. To give an example: the first few words of the English abstract, 'In recent years, there have', yielded: {'in', 'in recent', 'in recent years', 'recent', 'recent years', 'recent years there', 'years'}. In Mandarin, word boundaries are not explicit. Thus, we segmented the Chinese version of the abstract into words first with the tools from (Youli, 2011), and then proceeded to phrase segmentation.

Journal Abstract in Mandarin

文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术,近年来得到了泛的关注和快速的发展.提出了基于机器学习的文本分类技术所面临的互联网内容信息处理等复杂应用的挑战,从模型、算法和评测等方面对其研究进展进行综述评论.认为非线性、数据集偏斜、标注瓶颈、多层分类、算法的扩展性及 Web 页分类等问题是目前文本分类研究的关键问题,并讨论了这些问题可能采取的方法.最后对研究的方向进行了展望.

Journal Abstract in English

In recent years, there have been extensive studies and rapid progresses in automatic text categorization, which is one of the hotspots and key techniques in the information retrieval and data mining field. Highlighting the state-of-art challenging issues and research trends for content information processing of Internet and other complex applications, this paper presents a survey on the up-to-date development in text categorization based on machine learning, including model, algorithm and evaluation. It is pointed out that problems such as nonlinearity, skewed data distribution, labeling bottleneck, hierarchical categorization, scalability of algorithms and categorization of Web pages are the key problems to the study of text categorization. Possible solutions to these problems are also discussed respectively. Finally, some future directions of research are given.

Figure 1: Sample journal abstract in Mandarin and English (from Su et al. 2006)

The words and word phrases resulting from text segmentation are potential entities. We then check which of these phrases match a title in Wikipedia. These titles are either a page name in Wikipedia domain or a redirection page to an entity with an alternate title. Redirections happen for alternative names, plurals, closely related words, adjectives/adverbs pointing to the corresponding noun, less or more specific forms of names, abbreviations, alternative spellings, or punctuation and likely misspellings. The potential entities that have matches to Wikipedia titles are then considered existing entities, and are used in the ontology generation.

3.2 Ontology Generation

Wikipedia offers a network of networks: each language domain provides concepts and their relationships. These language-specific networks connect through the language links given on a Wikipedia page for a particular concept, and point users to pages with the same conceptual meaning in the alternate, target language. It is important to note that language links in Wikipedia do not direct the reader to the translation of the original content but to another Wikipedia page created for the same concept in the designated language. To give an example, machine learning page in English is linked to 机器学习 (Machine learning) in the Chinese Wikipedia, but the contents of these two pages are different; the two pages are created and updated by different users at different times.

We build the ontology of a document using the entities extracted from the text (see 3.1) and the Wikipedia categories of those entities. More specifically, we captured the immediate first level categories of the entities with existing Wikipedia title pages via Wikipedia's API. During the process, hidden categories were excluded since they are used for administrative purposes. Ontologies were constructed according to the following experimental conditions. For experiment A, Mandarin entities were extracted from the Mandarin version of the abstract, and the Chinese Wikipedia (http://zh.wikipedia.org/wiki/Wikipedia:首页) was used to build an ontology. For experiment B, Mandarin entities were extracted from the Mandarin version of the abstract. Next, we identified the corresponding English Wikipedia pages for the Mandarin entities and used the English entities to build the ontology from English Wikipedia. Entities without a corresponding English page were ignored. For experiment C, English entities were extracted from the English version of the abstract, and English Wikipedia (http://en.wikipedia.org/wiki/Main_Page) was used to build the ontology. These experimental conditions are summarized in Table 1.

Experiment	Language of Entities	Wikipedia Language
A	Mandarin	Mandarin
B	Mandarin	English
C	English	English

Table 1: Summary of Experimental Conditions

3.3 Ontology Visualization

For the visualizations, the python library, pyprocessing, was used to apply Processing (www.processing.org), a platform that allows for the creation of interactive visualizations. Orange circles show extracted entities that landed on Wikipedia titles with existing pages in the respective language. The first-level categories associated with those pages were visualized as blue circles. A line shows the link back to the corresponding entity represented as a Wikipedia title. At this time a spring weighting function is used to automate the positioning of the items in the bipartite graphs constituting the ontology visualizations.

4 Results and Discussion

Figures 2 (below), 3, and 4 (following pages) display the ontologies resulting from experimental conditions A, B, and C respectively. Figure 2 shows several key concepts from the journal abstract about machine learning in NLP have been effectively captured as entities using the collective knowledge base of Chinese Wikipedia. In Figure 2 the English entity names are given in

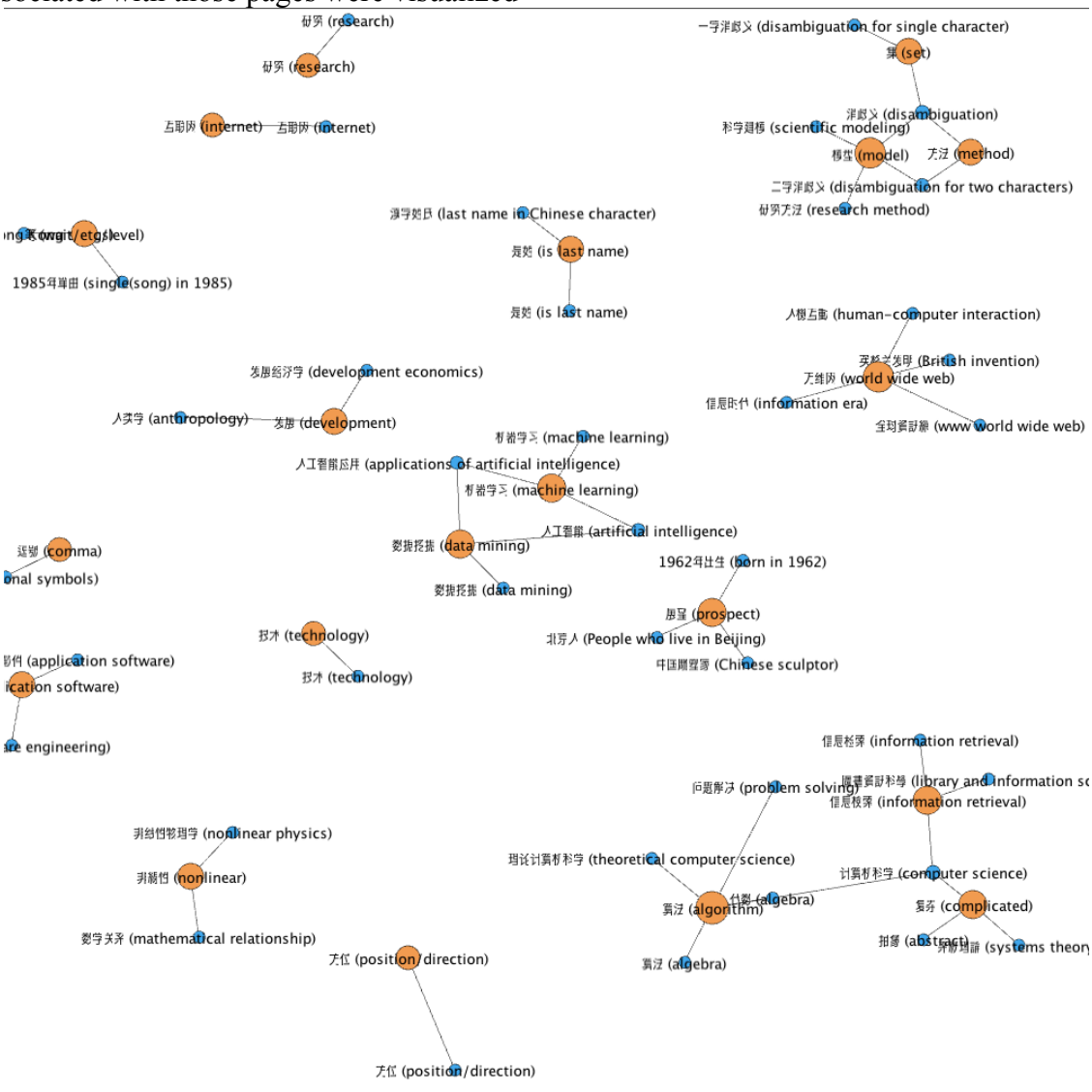

Figure 2: Ontology generated using experimental condition A, in which Chinese Wikipedia is used to build an ontology from Mandarin entities. The English translation of the entities are given for reference. Note that all nodes display, but the current algorithm uses the edge of the canvas as x=0, so some of the entities may not display as complete circles.

parentheses for reference. A native Mandarin speaker translated the Mandarin characters, which had been presented as a list of terms. Figure 3 contains many of the same concepts seen in Figure 2. Figure 4, created from the English abstract and English Wikipedia, displays approximately fifty percent more entities (excluding disambiguation). The entities associated with the disambiguation category currently in figure 4 can be filtered out as needed.

Table 2 summarizes the number of entities and maximum number of categories for each of the experiments. A total of eight entities were shared among all three experimental conditions.

Experiment	# of Entities	Max # of 1st level Categories
A	20	4
B	16	10
C	33 *	11

Table 2: Summary of Ontology Metrics, (* Excludes entities connected to disambiguation categories)

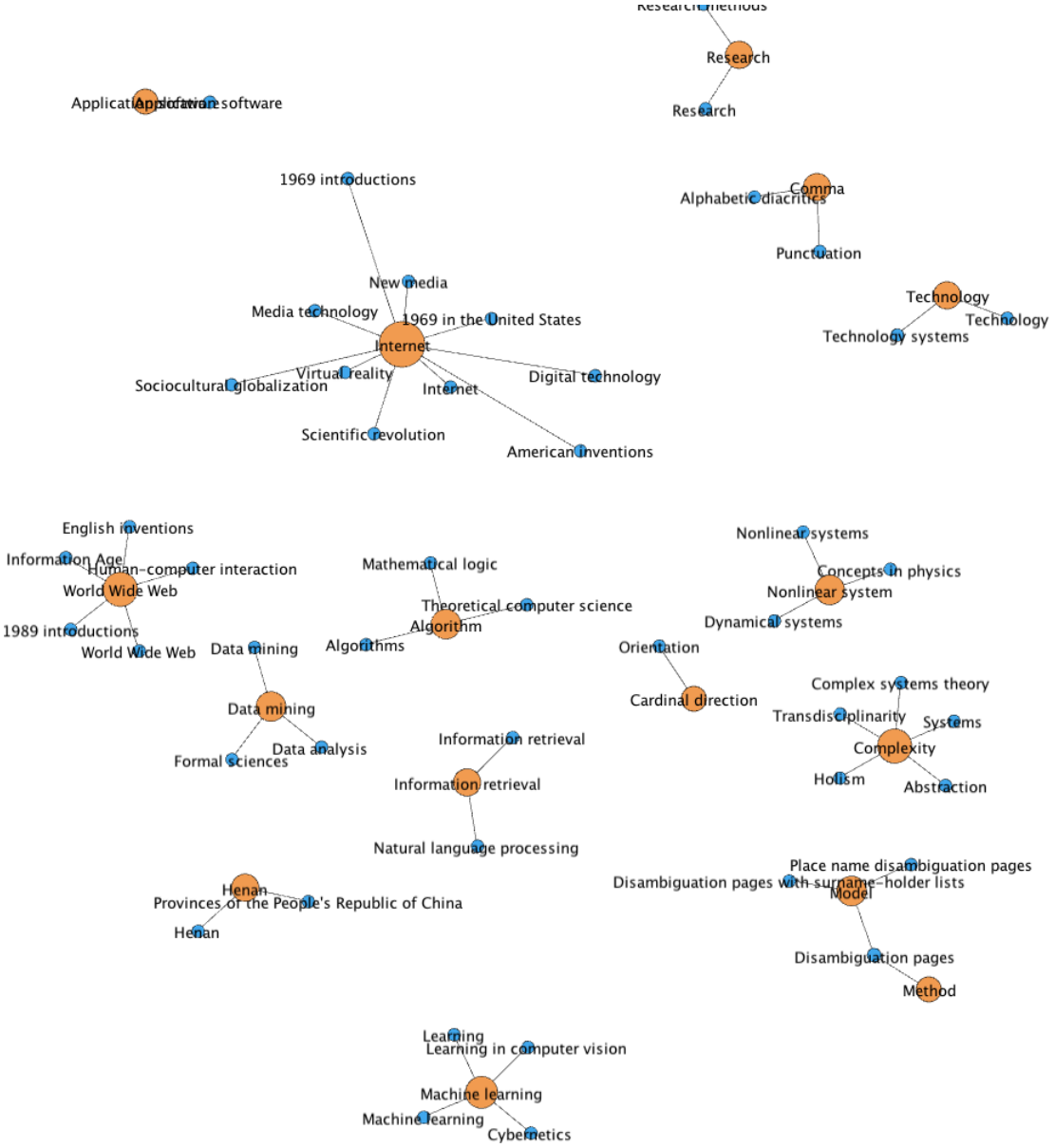

Figure 3: Ontology generated using experimental condition B, in which Chinese Wikipedia's links to the English Wikipedia in order to build an ontology from Mandarin entities.

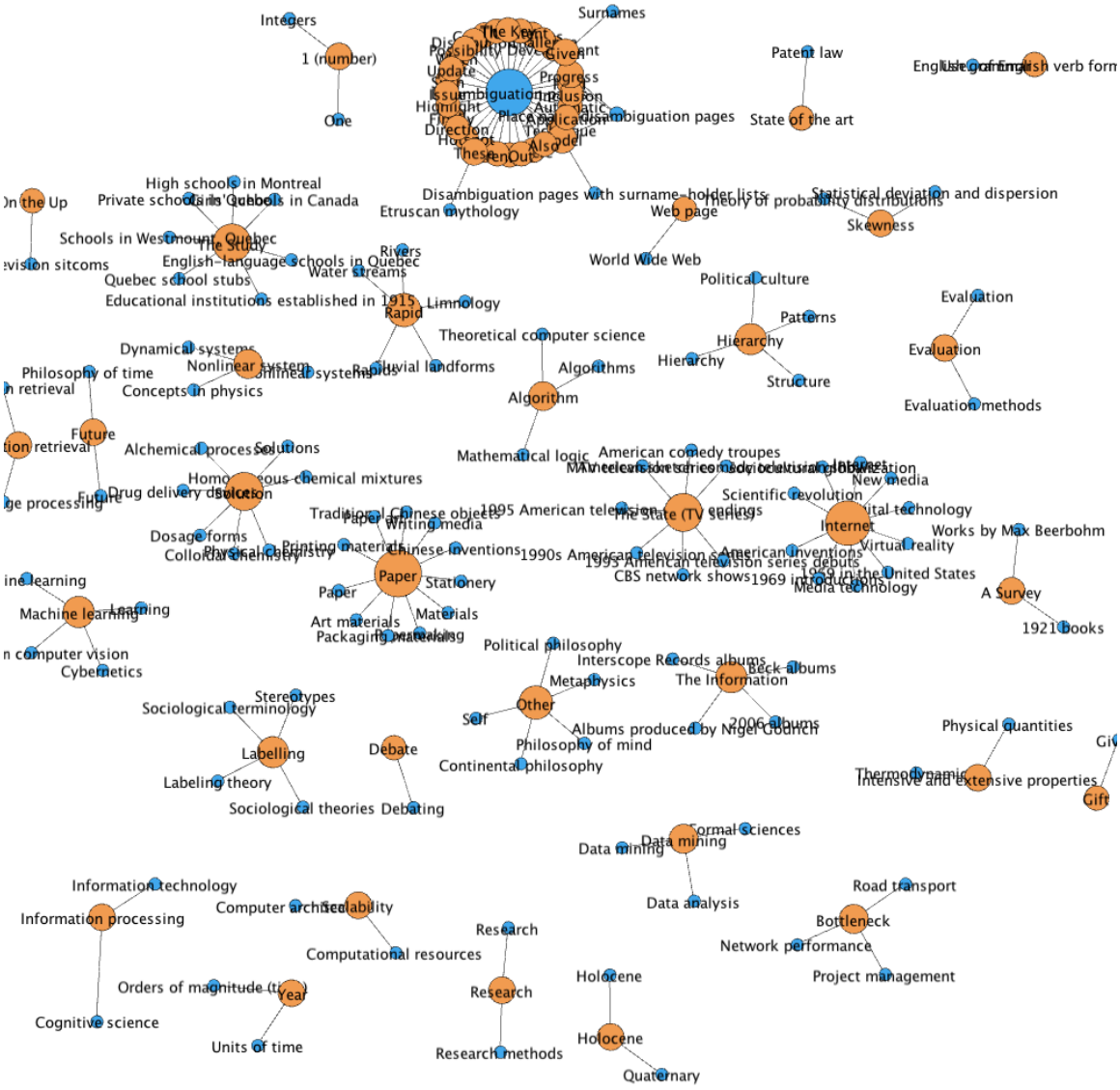

Figure 4: Ontology generated using experimental condition C, in which English Wikipedia is used to build an ontology from English entities.

These visualizations yield preliminary insights into the manner in which varying languages represent concepts in Wikipedia. Comparing the ontologies in Figures 2-4 reveals different languages in Wikipedia exhibit different breadth. English Wikipedia provided more concepts than the Chinese counterpart for this text sample. This is not surprising given the English Wikipedia is larger. However, the Mandarin entities shown in Figure 2 offer a satisfactory representation of the text. In addition, the extra concepts from English Wikipedia add little to the general understanding of the text, and may even distract from the abstract's key concepts.

Wikipedia pages from different languages generate different ontologies for seemingly similar concepts. For example, in Experiment A (Figure 2), *algorithm*, *information retrieval*, and *complexity* (which has the English label 'complicated') are connected through the *computer science* category. However, the corresponding English pages of these entities used in experiment B (Figure 3) are not connected through any shared first-level categories. This suggests English Wikipedia pages are categorized in greater detail, making it difficult to capture relationships among concepts through the immediate, first-level categories. In other words, the detailed ontology of English Wikipedia may not be as effective a reference as the simple ontology in Chinese Wikipedia. It could also be that the translation process introduces noise. Identifying and visualizing the second-level category connections might

provide further insight into the differences between the two methods.

5 Summary and Next Steps

As a context-rich, collaborative knowledge base, Wikipedia is ideal for building ontologies. This study presented varying approaches to constructing ontologies from simplified Chinese and English Wikipedias, as a first step in evaluating cross-lingual corpora. The methods employed in this study can be further adopted to extract ontologies across multiple languages provided the analogous collaborative knowledge stores exist in the target languages. The sample ontology visualizations generated in this work demonstrated there are multiple ways to pursue concept representation using the Chinese and English versions of Wikipedia.

Wikipedia offers networks of concepts in different languages. Networks of different languages in Wikipedia are mapped through language links within pages, but this is rarely a one-to-one mapping. Thus, we also need ways to align ontologies with different levels of explicitness and formalization.

Future research might build on the visualization techniques discussed here in order to explore mechanisms for ontology alignment. For example, the percentage of entity coexistence within a set of ontologies could be used as a metric for the alignment of ontologies. In addition, the techniques described here could be used to assess semantic similarity using ontologies coming from different collaborative data stores in different languages.

Finally, there are two approaches to extracting ontologies from cross-lingual corpora: work can be translated first and then ontologies extracted, or ontologies can be extracted, and then the ontologies translated. With more experiments, it may be possible to determine which is the best order to use, taking into account the corpus, the languages involved, and the collaborative data stores available.

Acknowledgments

The authors acknowledge and thank Dorian Zaccaria for helping with the ontology visualizations and Yue Han for help with verification of Chinese character translations.

References

Freitas-Junior, H. R., Ribeiro-Neto, B., Vale, R. F., Laender, A. H. F., & Lima, L. R. S. (2006). Categorization-driven cross-language retrieval of medical information. *Journal of the American Society for Information Science and Technology*, 57(4), 501–510. doi:10.1002/asi.20320

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*(2), 443-498.

Genc, Y., Mason, W., & Nickerson, J. (2012). Semantic transforms using collaborative knowledge bases. Paper presented at *Workshop on Information in Networks*, New York University, September 28-29, 2012. Available at SSRN 2154367.

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: Classifying tweets through a semantic transform based on Wikipedia, In *Proceedings on Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*. Orlando, FL. pp. 484-492.

Gennari, S. P., MacDonald, M. C., Postle, B. R., & Seidenberg, M. S. (2007). Context-dependent interpretation of words: Evidence for interactive neural processes. *Neuroimage*, *35*(3), 1278-1286. doi:10.1016/j.neuroimage.2007.01.015

Goddard, C. (2011). *Semantic Analysis: A Practical Introduction* (2nd Ed.). Oxford University Press, New York, NY.

Hovy, E. (2005). Methodologies for the reliable construction of ontological knowledge. In *Proceedings of the 13th international conference on Conceptual Structures: Common Semantics for Sharing Knowledge (ICCS '05)*. pp. 91–106. doi:10.1007/11524564_6

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information Processing & Management*, 41(3), 433–455.

Landauer, T. K., & Littman, M. L. (1991). A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Eleventh International Conference: Expert Systems and Their Applications*, 8, pp. 77-85.

Lanzenberger, M., Sampson, J., & Rester, M. (2010). Ontology visualization: Tools and techniques for visual representation of semi-structured meta-data. *Journal of Universal Computer Science,* 16(7), 1036-1054.

Locke, W. N., & Booth, A. D. (Eds.). (1955). *Machine translation of languages: fourteen essays*. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York, NY.

Michelson, M. & Macskassy, S. A. (2010). Discovering users' topics of interest on Twitter: A first

look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data (AND '10)* Toronto, Canada. doi:10.1145/1871840.1871852

Nichols, E., Bond, F., Tanaka, T., & Fujita, S. (2006). Multilingual ontology acquisition from multiple MRDS. In *Proceedings of the 2nd Workshop on Ontology Learning and Population,* Sydney, Australia pp. 10–17.

Onyshkevych, B. A., & Nirenburg, S. (1992). Lexicon, ontology, and text meaning, 289–303. doi:10.1007/3-540-55801-2_42

Sorg, P., & Cimiano, P. (2012). Exploiting Wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering, 74 (2012) 26–45*. doi:10.1016/j.datak.2012.02.003

Su J.S., Zhang B.F., & Xu X. (2006). Advances in machine learning based text categorization. *Journal of Software*, 2006,17(9), 1848-1859. doi: 10.1360/jos171848

Youli, D. (2011). Chinese Segmentation Analysis [Software]. Available from http://trac.xapian.org/wiki/GSoC2011/ChineseSegmentationAnalysis