Language Resources and Evaluation Vol. 42, No. 2 pp. 109-116, 2008
This is a preprint version.
Published version is available at http://dx.doi.org/10.1007/s10579-008-9071-y

Asian language resources: The state-of-the-art

Takenobu Tokunaga (take@cl.cs.titech.ac.jp)
Tokyo Institute of Technology

Chu-Ren Huang (churenhuang@gmail.com)
Academia Sinica

Sophia Yat Mei Lee (sophiaym@gmail.com)
Academia Sinica

1. Resource Development for Asian Languages

This special issue of Language Resources and Evaluation, entitled "New Frontiers in Asian Language Resources", complements the earlier special double issue on *Asian Language Processing: State of the Art Resources and Processing* (Huang et al., 2006) by presenting eight papers describing specific Asian language resources. As Bird and Simons (2003) explain, research on language resources must deal with how the resources can be acquired and documented as well as how the resources can be accessed and used. Among the eight papers in this issue, the first four papers focus on resources, while the latter four target specific application tasks and describe resource building in the contexts of these applications.

In the early days of corpus building, a "large scale" corpus might consist of one million words. Kilgarriff and Grenfenstette's (2003) survey of the historical developments in corpus construction shows that the size of English corpora has increased roughly tenfold every decade since the 1960's, when the one million word Brown Corpus was developed. In the 1980's, the COBUILD project built an eight million word corpus, and the British National Corpus (BNC), completed in 1994, includes 100 million words. This trend continues with LDC's Gigaword Corpus, published in 2003, which contains nearly two billion words. A central question for the development of resources for Asian languages, for which far less electronic data is in existence than for English, is whether the same amount of time will be required for Asian language resources to achieve a similar magnitude of scale. If the answer is yes, study of these languages relying on language resources will remain in its infancy for at least another decade. If not, it is yet to be determined how fast language resources for these languages can be developed.

The explosive growth of the Internet in the 1990s, and particularly the prevalence of Web technologies, drastically changed the potential to gather very large-scale language data; in addition to the vast reduction in computer costs, especially for storage, the Web enabled researchers to easily collect

enormous amounts of on-line text of various types and genres, such as news articles, novels, and blogs, and today, there exist terabyte scale data in some specific areas (Clarke et al., 2004; Brants and Franz, 2006) collected from the Internet. Given the impact of the Web on the size of language resources for English, we can imagine that access to Web data will significantly decrease the development time for language resources for Asian languages, and in particular, Asian languages for which few or no resources exist. The example of Chinese suggests that although it takes time to catch up, the four-decade process can be compressed: the Chinese Gigaword Corpus appeared in 2003 and was fully tagged by 2007, roughly 10 years after the two million word version of Sinica Corpus became available in 1995. We can anticipate that the development of language resources for less computerized languages in Asia will progress at an even faster pace.

Building language resources by treating the Web as the main source of data has attracted much attention in recent years, and the "Web as corpus" is now the topic of a series of workshops held in conjunction with major conferences on computational linguistics. Three papers in this volume deal explicitly with building language resources from Web data: Ekbal and Bandyopadhyay, Zhao and Liu, and Wong and Xia. Ekbal and Bandyopadhyay attempt to identify proper names by using specific tags in on-line newspaper articles. Zhao and Liu focus on product name extraction, and build a corpus from Web pages concerning product information, such as those concerned with product releases, market trends, product evaluation, etc. Both of these papers demonstrate that careful selection of the data source is indispensable for successful results when using the Web data.

A repeated criticism of using the Web data naively and indiscriminately is that Web data are fraught with orthographic and grammatical errors (Kilgarriff, 2007; Ringlstetter et al., 2006). Wong and Xia's paper addresses this crucial issue by tackling problems involving chat style texts. Chat room texts are typically more grammatical than spoken language and less grammatical than written text. Their proposed normalization procedure takes a crucial step towards more reliable Web data for language analysis.

2. Research Issues

This section looks at several research issues discussed by the papers in this volume.

2.1. Knowledge for building language resources

Building language resources is a labor-intensive and time-consuming task. In spite of the recent development of machine learning techniques, manually

constructed resources are still required to provide accurate data for training. The most difficult part of manual resource building is maintaining consistency. Even when corpus designers decide on a set of annotation criteria and provide substantial documentation as annotation guidelines, they cannot foresee every phenomenon the annotators may come across during the annotation process. To address problems that arise during the annotation process, a cycle of discussion among the corpus designers and the annotators followed by refinement and/or revision of the annotation guidelines is typically undertaken. In this way, knowledge about the processes and problems of annotation are accumulated as a byproduct of corpus building.

The work described by Hashimoto et al. in this issue is unique in that building the knowledge for annotation is the primary goal. They attempt to construct a lexical type database from existing language resources, with the aim of providing guidelines for keeping the consistency in constructing a Japanese treebank. A database entry consists of five types of information: type name & linguistic discussion, exemplification, implementation, links to confusing lexical types, and links to other dictionaries, which, in turn, help annotators to make decisions on problematic cases. The database was originally built for treebank construction, but as the authors point out, it could provide an *interlingual hub* connecting various kinds of language resources that have been developed independently at different research sites.

2.2. RESOURCE INTEGRATION

There have been many attempts to build a new resource by extracting information from existing (structured) resources rather than unstructured raw text. Based on an existing method proposed by Tanaka and Iwasaki (1996), Bond and Ogura construct a Japanese-Malay bilingual dictionary from Japanese-English and Malay-English dictionaries, using English as a pivot language. In addition to the surface string matching used in the Tanaka's original method, Bond and Ogura exploit syntactic and semantic information as well as translation preferences from the source dictionaries. Some entries of the Malay-English dictionary have Chinese translations as well; for these entries, Chinese was used as a second pivot language by using a Japanese-Chinese dictionary. The experimental results show that about 50,000 out of 350,000 Japanese words were linked to Malay counterparts. Bond and Ogura report that the POS and second language filtering are effective to improve the precision. In particular, the second language filtering improves the precision up to 0.97 at the cost of recall. This figure shows a remarkable improvement in comparison with the original method, which had around 0.85 precision.

2.3. ONTOLOGY BUILDING

There is a long history of research for extracting ontological knowledge from language resources, in particular, hypernym (is-a) relations. It is notable that this type of research started in Asia in incunabula of corpus-based NLP. Tsurumaru et al. (1986) extracted hypernym relations from the gloss of a Japanese dictionary using a pattern-based method, and Nakamura and Nagao (1988) extracted semantic information, including hierarchical relations, from the Longman Dictionary of Contemporary English (LDOCE). These approaches are still alive today and have been applied on a large scale of Web data (Pantel and Pennacchiotti, 2006).

Imsombut and Kawtrakul's paper in this volume also adopts the pattern-based approach to ontology building by automatic learning from plain Thai text corpora (i.e., data containing no HTML markup). They extract ontological concepts and taxonomic relations by using lexico-syntactic patterns and an item list. This approach, however, leads to three problems in identifying the relevant terms and relations: cue word ambiguity, item list identification, and candidate term selection. To overcome these problems, Imsombut and Kawtrakul exploit the lexicon and co-occurrence features of each candidate. They also weight each feature to measure its relevance with information gain. Though the work is still at the preliminary stage, the results are promising, with precision, recall, and F-measure of the system at 0.74, 0.78, and 0.76 respectively.

Classifier is a feature of many Asian languages such as Chinese, Japanese, Korean and Thai. There is a very small class of counterparts in English such as "a piece of furniture." Unlike English, classifiers of Asian languages are ubiquitous, i.e., they are used with almost every noun to denote its quantity. In addition, they demand semantic agreement with co-occurring nouns. The following are examples of classifiers in Chinese, Japanese, and Thai, where 'CLS' denotes a classifier.

Chinese:	*yi-ju*	*dian-hua*	··· a telephone		
	(CLS)	(telephone)			
Japanese:	*2*	*hiki*	*no*	*inu*	··· 2 dogs
		(CLS)	(of)	(dog)	
Thai:	*nakriian*	*3*	*khon*	··· 3 students	
	(student)		(CLS)		

In this example, Japanese *hiki* is a specific classifier used for counting animals. In applications such as machine translation involving a non-classifier language and a classifier language, it is important to select the proper classifier to express the number of objects (Bond and Paik, 2000). Kwon et al. describe efforts to build ontological knowledge of Korean numerative classifiers from various language sources, including a Korean dictionary, corpora, and a WordNet-like thesaurus. Their paper mentions two important

reasons for studying classifiers in addition to the development of NLP applications: language acquisition and classification of human's recognition of things, particularly, the semantic classification of nouns. There is seminal work on classifiers by Allan (1977) in which he conducted comparative study on classifiers of more than fifty languages from Africa, the Americas, Asia and Oceania. However, classifiers have been less studied from a corpus-based viewpoint (Shirai et al., 2008). According to Shirai et al, each Thai classifier tends to have tighter relation with a specific noun. It will be interesting to see if ontological structure is also possible for other Asian languages such as Thai.

2.4. NAMED ENTITY RECOGNITION

This volume includes two papers concerned with named entity recognition (NER), both of which start from building corpora for this specific task.

Ekbal and Bandyopadhyay build a corpus from on-line Bengali newspaper articles, claiming that the Web is a good source for less computerized languages to create language resources. Although the statistical approach is dominant in the NER task, they take a pattern-based bootstrapping approach. Assuming that a specific type of named entity appears in certain fields of newspaper articles, they manually build extraction patterns. For instance, person names tend to appear in the `<reporter>` field, location names in the `<location>` field, and organization names in the `<agency>` field. Frequent words appearing in these fields are put into a candidate list of NEs and used for annotation to build a training corpus. Four words around the target word in the list are then taken as an extraction pattern and applied to new text to extract new NEs, which are manually checked. This cycle continues until no new pattern is acquired. The overall performance of the system remains around 0.75 in F-measure, which leaves room for improvement compared to state-of-the-art NER systems for languages such as English.

Zhao and Liu explore NER task in Mandarin Chinese, focusing on product named entities. After building a corpus by crawling Web pages concerning products, the corpus (the *CASIA_PRO Corpus*) is manually annotated with three types of entities, namely Brand Name, Product Type, and Product Name. As this is the first product NE corpus for Chinese, annotation specifications are also defined. Using this corpus, Zhao and Liu integrate two hierarchical hidden Markov Models (HHMM) to recognize the product NEs, one based on word form features and semantic categories, and the other based on part of speech tag information. A series of experiments show that neither HHMM-1 nor HHMM-2 alone attains higher F-measures than those of the integrated model. The integrated model also outperforms the Maximum Entropy Model.

2.5. MACHINE TRANSLATION

Machine translation has been a typical and important application of NLP. Roxas et al. attempt to build an English-Filipino Machine Translation (MT) system using both rule-based and corpus-based approaches. Several language resources are exploited, including a bilingual English-Filipino lexicon, a Filipino grammar, translation rules, and annotated corpora. The project builds a 207,000-word Filipino corpus, half of which is manually annotated. Roxas et al. also address certain features of Filipino — free word order, complex verbal morphology, and the importance of focus — which make MT difficult. To improve the MT system, they develop relevant language tools such as a morphological analyzer and generator and an automatic part-of-speech tagger. Their work pioneers computational approaches to both language archives and language processing for the Filipino languages.

2.6. WEB SPECIFIC RESOURCES

Chat language poses a challenge to NLP because of its use of non-standard forms and neologisms. The work of Wong and Xia sheds some light on chat language processing by extending the existing Source Channel Model (SCM), a widely used statistical approach in speech recognition and machine translation.

Wong and Xia build and analyze a Chinese chat language corpus and determine that phonetic transcription between chat terms and their standard language counterparts might be an important means to improve chat language processing. They propose the eXtended Source Channel Model (XSCM) to convert the chat language to standard language by incorporating the phonetic mapping into SCM. With the integration of phonetic mapping between chat terms and their standard language counterparts, XSCM outperforms SCM in both chat term recognition and normalization accuracy.

3. Conclusion and Prospects

The publication of 24 papers dealing with 11 different Asian languages, including Bengali, Mandarin Chinese, Hindi, Japanese, Korean, Malay, Marathi, Thai, Filipino/Tagalog, Urdu, and Vietnamese, in the two special issues and one regular issue of this journal demonstrates both the language diversity and the vibrant emergence of human language technology in Asia. The appearance of papers spanning the range of language processing procedures, from the correction of spelling errors (Naseem and Hussain, 2007), to the automatic acquisition of grammatical information (Butt and King, 2007) underlines the challenges as well as the opportunities for Asian language technology. Basic linguistic issues must be solved to build the essential infras-

tructure for Asian language processing, but at the same time, state-of-the-art methodologies can be applied to solve sophisticated and pioneering language processing issues. In addition, the need to simultaneously study a wide range of issues for the same language offers a rare opportunity to examine how long-held presuppositions affect current research directions. The results of this research should provide a healthy and realistic model for the study of the development and use of language resources as well as the processing of less computerized and endangered languages. We look forward to future contributions that will enrich our knowledge of linguistic diversity and narrow the digital gap at the same time.

Acknowledgements

We would like to thank all the authors who submitted 74 papers on a wide range of research topics on Asian languages. We had the privilege of going through all these papers and wished that the full range of resources and topics could have been presented. We would also like to thank all the reviewers, whose prompt action helped us through all the submitted papers with helpful comments. We would like to thank AFNLP for its support of the initiative to promote Asian language processing. Various colleagues helped us processing all the papers, including Dr. Sara Goggi at CNR-Italy, and Liwu Chen at Academia Sinica. Finally, we could like to thank four people at LRE and Springer that made this special issue possible. Without the generous support of the chief editors Nancy Ide and Nicoletta Calzolari, this volume would not have been possible. In addition, without the diligent work of both Estella La Jappon and Jenna Cataluna at Springer, we would never have been able to negotiate all the steps of publication. For this introductory chapter, we would like to thank Kathleen Ahrens, Nicoletta Calzolari, and Nancy Ide for their detailed comments. Any remaining errors are, of course, ours.

References

Allan, K.: 1977, 'Classifiers'. *Language* **53**(2), 285–311.

Bird, S. and G. Simons: 2003, 'Seven dimensions of portability for language documentation and description'. *Language* **79**(4), 557–582.

Bond, F. and K. Paik: 2000, 'Reusing an ontology to generate numerical classifiers'. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. pp. 90–96.

Brants, T. and A. Franz: 2006, 'Web 1T 5-gram Version 1'. LCD Catalog No. LDC2006T13.

Butt, M. and T. H. King: 2007, 'Urdu in a parallel grammar development environment'. *Language Resources and Evaluation* **41**(2), 191–207.

Clarke, C., N. Craswell, and I. Soboroff: 2004, 'Overview of the TREC 2004 Terabyte Track'. In: *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*.

Huang, C.-R., T. Tokunaga, and S. Y. M. Lee: 2006, 'Special Issue on: Asian Language Processing: State-of-the Art Resources and Processing'. *Language Resources and Evaluation* **40**(3-4).

Kilgarriff, A.: 2007, 'Googleology is Bad Science'. *Comutational Linguistics* **33**(1), 147–151.

Kilgarriff, A. and G. Grenfenstette: 2003, 'Introduction to the Special Issue on the Web as Corpus'. *Computational Linguistics* **29**(3), 333–347.

Nakramura, J. and M. Nagao: 1988, 'Extraction of semantic information from an ordinary English dictionary and its evaluation'. In: *Proceedings of the 12th International Conference on Computational linguistics (COLING 1988)*. pp. 459–464.

Naseem, T. and S. Hussain: 2007, 'A novel approach for ranking spelling error corrections for Urdu'. *Language Resources and Evaluation* **41**(2), 117–128.

Pantel, P. and M. Pennacchiotti: 2006, 'Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations'. In: *Proceedings of the 21st International Conference on Computational Linguistics/the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*. pp. 113–120.

Ringlstetter, C., K. U. Schulz, and S. Mihov: 2006, 'Orthographic Errors in Web Pages: TowardCleaner Web Corpora'. *Computational Linguistics* **32**(3), 295–340.

Shirai, K., T. Tokunaga, C.-R. Huang, S.-K. Hsieh, T.-Y. Kuo, V. Sornlertlamvanich, and T. Charoenporn: 2008, 'Constructing Taxonomy of Numerative Classifiers for Asian Languages'. In: *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*. pp. 397–402.

Tanaka, K. and H. Iwasaki: 1996, 'Extraction of lexical translations from non-aligned corpora'. In: *Proceedings of the 16th International Conference on Computational linguistics (COLING 1996)*. pp. 580–585.

Tsurumaru, H., T. Hitaka, and S. Yoshida: 1986, 'An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary'. In: *Proceedings of the 11th International Coference on Computational linguistics (COLING 1986)*. pp. 445–447.

Resources

British National Corpus.	http://www.natcorp.ox.ac.uk/
Brown Corpus.	http://icame.uib.no/brown/bcm.html
Cobuild Project.	http://www.collins.co.uk/corpus/CorpusSearch.aspx
Sinica Corpus.	http://www.sinica.edu.tw/SinicaCorpus
Chinese Gigaword	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09
English Gigaword	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05
Tagged Chinese Gigaword	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T03