Language Resources and Evaluation (2006) Vol. 40, No. 3-4, pp. 208-218

# Asian language processing: current state-of-the-art

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

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

Sophia Y. M. Lee (sophiaym@gmail.com)
*Academia Sinica*

## 1. Background: The Challenge of Asian Language Processing

Asian language processing presents formidable challenges to achieving multilingualism and multiculturalism in our society. One of the first and most obvious challenges is the multitude and diversity of languages: more than 2,000 languages are listed as languages in Asia by Ethnologue (Gordon, 2005), representing four major language families: Austronesian, Trans-New Guinea, Indo-European, and Sino-Tibetan[1]. The challenge is made more formidable by the fact that as a whole, Asian languages range from the language with most speakers in the world (Mandarin Chinese, close to 900 million native speakers) to the more than 70 nearly extinct languages (e.g. Pazeh in Taiwan, one speaker). As a result, there are vast differences in the level of language processing capability and the number of sharable resources available for individual languages. Major Asian languages such as Mandarin Chinese, Hindi, Japanese, Korean, and Thai have benefited from several years of intense language processing research, and fast-developing languages (e.g., Filipino, Urdu, and Vietnamese) are gaining ground. However, for many near-extinct languages, research and resources are scarce, and computerization represents the last resort for preservation after extinction.

A comprehensive overview of the current state of Asian language processing must necessarily address the range of issues that arise due to the diversity of Asian languages and must reflect the vastly different state-of-the-art for specific languages. Therefore, we have divided the special issues on Asian language technology into two parts. The first is a double issue entitled *Asian Language Processing: State of the Art Resources and Processing*, which focuses on state-of-the-art research issues given the diversity of Asian languages. Although the majority of papers in this double issue deal with

[1] These four language families, plus the Niger-Congo family in Africa, each include more than 400 languages. Other larger language families in Asia include Austro-Asiatic (169), Tai-Kadai (76), Dravidian (73), and Altaic (66).

DOI: 10.1007/s10579-007-9041-9

major languages and familiar topics, such as spell-checking and tree-banking, they are distinguished by the innovations and adaptations motivated by the need to account for the linguistic characteristics of their target languages. For instance, Dasgupta and Ng's morphological processing of Bengali has an innovative way to deal with multiple stems while Ohno et al.'s parsing of monologues makes crucial use of *bunsetsu*[2] and utterance-final particles, two important characteristics of Japanese. A subsequent issue entitled *New Frontiers in Asian Language Resources* will focus on both under-computerized languages and new research issues, such as the processing of non-standard language found on the web. Overall, these special issues on Asian language processing assess the state-of-the-art for more than thirteen languages from six of the eight major Asian language families[3]. As such, they provide a snapshot of the state of Asian language processing as well as an indication of the research and development issues that pose a major challenge to the accommodation of Asian languages in the future.

## 2. Language Processing in Asia: A Brief Overview

Research on Asian language technology has thrived in the past few years. The Asian Language Resources Workshops, initiated in 2001, have had over sixty papers presented in five workshops so far (http://www.cl.cs.titech.ac.jp/alr/). Interest in Asian language processing among researchers throughout the world was made evident in a panel entitled *Challenges in NLP: Some New Perspectives from the East* at the COLING/ACL 2006 joint conference. At the same conference, fifteen papers were accepted in the Asian language track, while many other accepted papers also dealt with processing Asian languages. The growing literature on Asian language processing attests to the robustness of current paradigms. For instance, corpus-based stochastic models have been widely adopted in processing of various Asian languages with results comparable to that of European languages. Studies on less computerized languages in Asia, however, do not have the luxury of simple adaptation of accepted paradigms and benchmarks. They are burdened by the dual expectations of infrastructure building and language engineering applications. On one hand, early stages of computerization mean that many types of language resources must be built from scratch. On the other hand, the maturing field of computational linguistics expects attested and quantifiable results not tenable without

[2] *Bunsetsu*, often translated as base phrase, is the basic unit of Japanese text proposed by Hashimoto (1984). A *bunsetsu* is a written and prosodic unit which is typically composed of a root and particles and can be identified by phonological principles. The concept of *bunsetsu* is also adopted in Korean linguistics.

[3] Two of these languages, Filipino and Urdu, do not appear in the current issue and will be represented in the subsequent issue.

DOI: 10.1007/s10579-007-9041-9

substantial language resources. It is remarkable that this delicate balancing act has been performed successfully, as attested by many papers appearing in this and the subsequent issues that deal with Bengali, Filipina, Hindi, Marathi, Thai, Urdu, and Vietnamese, among others. A particularly striking example of how infrastructure building can go hand in hand with technological innovation is Collier et al.'s work on multilingual medical information extraction for Asian languages.

Japanese scholars were the pioneers in Asian language processing. The Information Processing Society of Japan (IPSJ) was formed in 1960 with a significant number of members interested in Machine Translation and related areas. Natural language processing (NLP) activities in Japan were extensive in the 1980's, starting with the first international conference on computational linguistics held in Asia: the 1980 Tokyo COLING. In 1982, the Fifth Generation Computer Project contained significant segments on NLP. One of the most visible products of this project was the EDR dictionary from the Electronic Dictionary Research Center founded in 1986. Lastly, the Association for Natural Language Processing was formally formed by the Japanese in 1994.

The development of NLP research in Japan is atypical of Asian languages, largely because Japan leads Asian countries in terms of technology development. In most other Asian countries, research on NLP is relatively new or in its infancy: interest in Chinese has increased dramatically over the past ten years due to China's emergence as a world power, but many other countries are only now initiating work on NLP for their languages. In general, the history of the development of language processing capabilities for Chinese is more similar to that of other Asian languages than to Japanese.

T'sou (2004) summarizes the developments of Chinese language processing. Even though the earliest efforts on Chinese language processing can be traced back to the 1960's, more concerted efforts started in the late 1980's, marked by the first computational linguistics conferences in both China and Taiwan in 1988 and followed by increased research activity in the 1990s (T'sou, 2004). Related research became more visible in the 1990's. Based on a chronology provided by Chu-Ren Huang, T'sou (2004) showed that the maturing of the field was marked by the arrival of sharable resources in the early 1990's, which were developed independently at the Academia Sinica and at Peking University. The quantity and quality of NLP research increased through the years, and finally reached the milestone of the formation of SigHAN, the special interest group on Chinese language processing, within the Association for Computational Linguistics in 2002. One may observe that in this chronology, the availability of language resources has served as both a foundation for research activity and a landmark of its maturity. This observation underlines the design feature of this special issue on Asian language processing. The dual foci on both language resources and language technology allow us to

DOI: 10.1007/s10579-007-9041-9

capture the dynamic, multi-dimensional state of Asian language processing, a research sub-field in its early development stage yet already producing exciting and challenging results.

## 3. Research and Technical Issues: An Integrated Perspective

We attempt to provide an integrated and unified perspective on the research issues and technological developments of Asian language processing in spite of the wide range of their linguistic diversity and lack of uniform level of computerization. We take as our point of departure the questions and answers presented by Joshi, Bhattacharyya, T'sou, and Tsujii at their COLING/ACL 2006 panel (Joshi, 2006)[4]. Two scientific questions are posed and answered by integrating their findings with new information derived from papers in this volume in the first half of this section. The second half of this section consists of the synopsis of the research issues and technological developments reported in the papers.

### 3.1. Whither Asian language processing? Two critical issues

Given the vast linguistic diversity and great computational disparity among Asian languages, it has been a challenge to characterize a set of linguistic and research topics common among all Asian languages. However, we do find a set of characteristics uniquely shared by most Asian languages: flourishing yet relatively early stages of language resource development, and the need for native language processing as an empowering socio-economical tool. We believe that this sense of shared purpose and the common challenge to balance energy spent on resource construction and technological development both play a crucial role in creating a strong sense of community among Asian language processing researchers. They also form the premise of two questions regarding the direction and significance of Asian language processing, which should have broader implications for the field of computational linguistics in general:

1. Are Asian Language Processing studies merely applications of old technology or innovative advances in the field?

2. Can linguistic knowledge help advance Asian language technology?

[4] The panel, entitled *Challenges in NLP: Some New Perspectives from the East*, covers three different issues: Jun'ichi Tsujii's *Diversity vs. Universality: Are Asian language special*, Benjamin T'sou's *Some Salient Linguistic Differences in Asia and Implications for NLP*, and Pushpak Bhattacharyya's *Can the availability of detailed linguistic information (for example, morphology) help in ameliorating the scarcity of large annotated corpora.*

DOI: 10.1007/s10579-007-9041-9

### 3.1.1. *Are Asian Language Processing studies merely applications of old technology or innovative advances in the field?*

The diversity vs. universality dichotomy discussed by both Tsujii, and T'sou in 2006 in the context of Asian language processing draws attention to the scientific merit of carrying out a specific language processing task in a new language. Indeed, if the same set of algorithms and procedures can be applied to all languages with the same expected results, any subsequent application after the methodology is first established will bear little scientific interest. Strict paradigmatic interpretation of scientific developments seems to encourage such monotonic and incremental views. It is not unusual to find the opinion that research topics such as POS-tagging and spell-checking are old and *solved*. This can be true for other phenomena as well: for example, Tsujii (2006) discusses discourse analysis motivated by *discourse-oriented* characteristics of many Asian languages, but it could be argued that this research follows well-established paradigms. If such arguments are valid, Asian language processing would simply be an application of existing technology and would have little to contribute to advancing computational linguistic research. However, there are many examples that contradict this view. For example, it is well-known that computational finite state morphology started with Koskenniemi's (1983) study of Finnish. The research on two-level morphology was greatly and immediately enhanced by the implementation of the KIMMO system and its application to English, Japanese, Romanian, and French (Karttunen and McCarthy, 1983). If extending established methods to a new language were indeed trivial, then structural mapping between any two languages would also be trivial given that they were obtained by the same set of algorithms. This is obviously not the case since the field of Machine Translation continues to tackle similar issues after more than 50 years of study. This *ad absurdum* deduction shows that extending NLP tasks to new languages is significant and non-trivial work.

The challenge of developing language processing capabilities for Asian languages may also contribute to the development of a more robust paradigm of computational linguistic theories and human language technologies in other ways. First and foremost, Asian language processing differs from previous work in that it requires a research paradigm where language resource development must be done in tandem with the development of language technology applications. This situation presents an opportunity to bootstrap resource development using state-of-the-art technologies that were unavailable to those creating resources ten or twenty years ago. In addition, issues involving processing of various hitherto unfamiliar linguistic facts are introduced by these new languages and brought to the attention of computational linguistics.

Indeed, shared regional features of Asian languages have had some implications for NLP theories and frameworks. T'sou (2006) underlined the role of non-western writing systems in language processing. He observed that the

DOI: 10.1007/s10579-007-9041-9

variety of phonemic, morphemic and mixed writing systems implied a different information load of the lexica, which can be measured by the differences in the entropies of these signs. In other words, basic atoms in NLP may be determined by conventions in language writing systems and have implications in NLP.

#### 3.1.2. *Can linguistic knowledge help advance Asian language technology?*

Bhattacharyya (2006) dealt with the issue of if and how linguistic information can help NLP when large annotated corpora are not available. His observation is based on the fact that Indian languages have a rich morphology and linguistic tradition. Faced with the daunting and time-consuming task of creating large annotated corpora for many languages simultaneously, he argued that morphological information can be an immediate source of knowledge needed for NLP applications. Morphological rules, instead of stochastic models, can be used in parsing and extraction of collocational information. It can also be used to semi-automatically tag large scale language resources. It is claimed that rule-based morphological knowledge can be easily and effectively adapted between different related languages. This approach ameliorates the problem caused by the lack of large scale grammatically annotated corpora and the time and resources usually required to build them. It also exemplifies one of the important roles that will be played by linguistic knowledge in Asian language processing.

Linguistic knowledge can also directly impact the construction of international standards. Asian colleagues' contribution to the drafting of language technologies related standards under ISO TC37 SC4 (Language Resources Management) offers a good example. As discussed in Tokunaga et al. (2006) and Huang et al. (2007), observations and generalizations of a set of salient linguistic characteristics based on several Asian languages, led to reconsiderations of the original proposal[5]. Issues specific to Asian languages that were not addressed in previous NLP studies include morpho-lexical reduplication, noun-classifier agreement, and the honorific system. Although consideration of such new linguistic issues does not alter the metamodel of LMF (Lexical Markup Framework, Francopoulo et al. 2006), it does require additions of new features in the Data Category Registry (DCR) and modifications in how some core concepts are defined. It is important to note that incorporation into a metamodel and proposed international standard is only possible as long as knowledge of these linguistic behaviors is available and felicitously accounted for. In sum, the availability of linguistic knowledge can certainly help to advance Asian language processing as long as such knowledge is formalized and situated in an appropriate model.

[5] Research on this issue was carried out in a project funded by NEDO and directed by Tokunaga Takenobu, as well as within the WG meetings of ISO TC37 SC4.

DOI: 10.1007/s10579-007-9041-9

## 3.2. Summary of Research Issues

In this section, we integrate and summarize four research foci raised by the papers in this special issue.

### 3.2.1. *Intelligent Document and Text Processing*

Document and text processing are rarely considered as core issues in NLP. However, for Asian languages, they are often deeply involved in the most fundamental issue of how to identify a linguistic unit in each language and are an essential first step in NLP. For instance, automatic spelling error correction is one of the most successful applications of NLP and is widely used in daily life. There has been a lot of work on spelling error correction. Readers can find a thorough survey by Kukich (1992) in which the work on spelling error correction is classified into three types: non-word error detection, isolated-word error correction, and context-dependent word correction. Unlike most European languages, however, many Asian languages do not put explicit delimiters between words. This feature makes segmentation indispensable in NLP. In addition, segmentation for Asian languages can be considered as a type of context-dependent word correction. Unfortunately, Kukich's survey does not take into consideration the processing of Asian languages. However, recently, there is a definite trend (e.g. Nagata, 1996; Nagata, 1998) to apply statistical approaches to morphological analysis of Asian languages. Two papers in this volume utilize text level distribution in their study: Chang and Yang use the distributional data of Chinese characters to recover the full form of Chinese abbreviations; and Bao et al. apply textual distribution to the higher level task of copy detection in Chinese documents.

### 3.2.2. *Resources and Standardization*

As mentioned earlier, one of the most urgent tasks in Asian language processing is the development of language resources. When multiple resources are being developed simultaneously with the expectation of facilitating cross-lingual knowledge-sharing, standards become a central issue. Resources-building is a labor intensive task requiring skilled annotators and it is therefore important to maximize the efforts of annotation. There are at least three ways to do so. The first is to coordinate manual labor and computer power to build corpora efficiently. The second is to integrate multiple layers of grammatical information in one annotated corpus. The third is to establish standards such that resources can be shared and reused. All these three ways present technical challenges. The coordination of human/machine work often involves design of efficient annotation tools to allow the annotator and the computer to share the burden in the annotation process. The challenge in integrating different layers of linguistic information lies in how to harmonize different linguistic modules. And the challenge in standardization is, of course, to have

DOI: 10.1007/s10579-007-9041-9

Language Resources and Evaluation (2006) Vol. 40, No. 3-4, pp. 208-218

a general framework which can anticipate all types of linguistic variations. Three papers in this volume deal with these three aspects respectively: Rim et al. concentrate on building tools such that the computer and human can share the burden when annotating, Bond et al. describe how different modes of linguistic information can be integrated into one annotated corpus, and Nguyen et al. implement the proposed ISO TC37 SC4 standard for lexicon representation in building a lexicon for Vietnamese.

### 3.2.3. *Syntactic Processing*

Syntactic processing requires probably the highest level of abstraction. Even though semantic processing is often considered post-syntactic in terms of procedure, it at least has real word meanings to be anchored on. The high level of abstraction is probably the reason why syntactic processing work is more popular among highly computerized languages, such as Japanese, and less popular among other Asian languages. It is also interesting to note that syntactic processing is often theory dependent. In addition to two papers dealing with syntactically annotated corpora (Bond et al. and Rim et al.), Ohno et al. describe Japanese parsing with a dependency grammar, while Butt and King's paper (to appear) adopts Lexical-Functional Grammar (LFG)'s Parallel Grammar (ParGram)[6] environment to implement a grammar for Urdu. Since Japanese is a head-final language, each element basically depends on a later element. Therefore, ambiguity of dependency increases combinatorially as the input becomes longer for compound and complex sentences. More than a decade ago, Kurohashi and Nagao (1994) introduced a technique into Japanese dependency parsing, dividing a long Japanese sentence into simple sentences based on their parallel structure, and succeeded in improving the performance of the parser. Their method has been implemented as KNP[7] and widely used until now. For Butt and King's work, the emphasis is on the sharable cross-lingual core of grammar building. The position of taking grammatical function as the basic level of representation allows LFG to propose a language-independent level of representation while at the same time specifying the idiosyncrasies by means of stipulating how each grammatical function is realized in the language in question.

### 3.2.4. *Semantic Processing*

Semantic processing is among the most popular topics in NLP recently because of its direct applicability to information and knowledge processing. It is important to note, however, that semantics itself has many complex aspects. Semantics ranges from atom-based sense or referent identification, to representation of complex concepts such as event and time. Three papers in this volume include two which focus on the construction of semantic lexica

[6] http://www2.parc.com/isl/groups/nltt/pargram/

[7] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

DOI: 10.1007/s10579-007-9041-9

(Bhattacharyya et al. and Xue), and one on establishing sense identity (Ji et al.). WordNet (Fellbaum, 1998) has become the *de facto* standard lexical database to represent sense and semantic relations, and is sometimes called a linguistic ontology. Bhattacharyya and colleagues apply WordNet to several Indian languages: Hindi and Marathi. Although the work is done in the context of India, it has great implication for future extensions when the Global WordNet initiative is undertaken to construct language WordNets as an infrastructure for knowledge exchange. In contrast, Xue adopts annotated data in Chinese PropBank to extract argument information for his construction of a semantic lexicon. Both approaches represent two of the state-of-the-art methods of constructing semantic lexical resources. Ji and colleagues, go one step further by adopting a novel approach toward learning word senses, by integrating feature selection and clustering.

### 3.2.5. *Task-Oriented Processing*

All three papers dealing with higher level task-oriented processing in this issue are on Japanese. This probably is a reflection of the fact that Japanese natural language processing developed first among Asian languages. The three tasks involved are idiom detection, information retrieval (IR), and machine translation (MT). NLP often assumes the compositionality principle. Idioms typically have a conventional structure with an unconventional meaning. The most effective way to deal with idioms is to list their meaning lexically. The fact that they are structurally identical with literal constructions, however, makes their discovery a challenge. Hashimoto et al. took on this challenge with success. The second paper, by Jones et al. deals with query expansion (QE) in IR. One of the crucial problems in IR is the mismatch between terms in user's queries and those in documents to be retrieved. For instance, a query containing "dog" does not typically yield the search result of "hound" in a document as the terms do not match. To alleviate this problem, every term in a query could be replaced with a set of the same or similar, or even related terms including itself. This technique is called query expansion (QE) (Baeza-Yates and Ribeiro-Neto, 1999). Jones et al.'s approach is unique since they use user's queries which are modified based on query logs of the search engine. In other words, their system modifies a query as the other users do. Query logs are usually proprietary, thus not accessible by ordinary researchers. In this respect, the research results using such data are quite informative. Lastly, the third paper, by Murata et al., takes on MT, the archetype of NLP task. The paradigm of MT has changed from a rule-based approach to an example-based approach and then to a statistical approach. Following the success of Brown et al.'s work of statistical machine translation (SMT) (Brown et al., 1993), SMT has been the main stream of the MT research. The unit of trans-

DOI: 10.1007/s10579-007-9041-9

lation grows from words to phrases or syntactic structures[8]. Murata et al. adopts this new method to deal with an old and persistent issue in MT, the representation of abstract temporal concepts such as tense and aspect. Since this issue involves cross-lingual idiosyncrasies, the example-based approach of SMT is well-suited for the solution of this issue.

#### 3.2.6. *Multilingual Processing*

Multilingual processing remains one of the last challenges in NLP. This challenge is especially acute in the context of Asian language processing. The most prototypical multilingual processing approach adopts an inter-lingual or a pivot language approach. The inter-lingual or pivot language is typically English. However, the Asian context poses a challenge since English is distant from most if not all Asian languages, and there is no clear alternative. An alternative to the tradition inter-lingual approach is the semantics-based ontology approach. Collier et al. (this volume) takes the approach of adopting an upper ontology as shared representation of cross-lingual information. The emergence of ontology based multilingual processing is one of the most promising recent trends in human language technology, as illustrated by Huang et al. (to appear). The fact that Collier and colleagues were able to successfully apply this approach to a new domain with several languages of varying degree of computerization is especially noteworthy.

## 4. Linguistic Features

We summarize the fourteen papers in this volume from a linguistic perspective, where individual summaries of these papers are grouped by the classification of the main language dealt with. This perspective presents a snapshot of the diversity of Asian languages as well as the rich linguistic and language technology issues covered. We hope that this perspective coupled with the summary of research topics in the previous section underscore the scientific contributions of Asian language processing to date and provide a basis for future research.

### 4.1. Summary of papers by language type

The current special double issue contains fourteen papers covering eight Asian languages: Bengali, Mandarin Chinese, Hindi, Japanese, Korean, Marathi, Thai, and Vietnamese. These languages belong to the following major language families: Altaic, Austro-Asiatic, Indo-European, Sino-Tibetan, and Tai-Kadai. The following summaries are grouped by language families and languages to underline how shared linguistic features advance language technol-

[8] http://www.cs.ust.hk/ dekai/ssst

DOI: 10.1007/s10579-007-9041-9

ogy and to present the diversity of linguistic issues involved in Asian language technology in a more systematic way.

### 4.2. Altaic Languages

The Altaic family includes 66 languages spoken mostly in Central and Northeast Asia. Japanese and Korean are usually categorized as members of the Altaic family[9], sometimes referred to as *Macro-Altaic*. Six papers in this volume deal with these two languages.

#### 4.2.1. *Japanese*

Jones and colleagues, in their paper "*Automatically generating related queries in Japanese*", describe the empirical study of applying query expansion originally developed for an English search engine to a Japanese one. Since Japanese uses four kinds of writing script, *Hiragana*, *Katakana*, *Kanzi* and the Roman alphabet, a direct application of the query expansion technique for English is not possible. The system needs to take into account the mismatches between scripts as well.

The paper "*Japanese-English translations of tense, aspect, and modality using machine learning methods and comparison of them and machine translation systems on market*" by Murata et al. deals with a structurally different language pair, English and Japanese. In particular, the paper focuses on translation of tense, aspect and modality which are notoriously difficult for translation because of the differences between English and Japanese. Their innovative approach adopts machine learning techniques. The proposed methods were evaluated in comparison to six machine translation products on the market. The paper reports that the proposed method with Support Vector Machine (SVM) outperformed its competitors. Although the evaluation was conducted on translation of tense, aspect and modality alone, this technique shows promise for improving translation systems in general.

In the paper "*Detecting Japanese idioms with a linguistically rich dictionary*", Hashimoto and colleagues propose a method to distinguish between idiomatic usages and literal usages of a given Japanese expression. Interestingly, they do not follow the current research trend involving machine learning, but rather adopt a rule-based approach using an idiom dictionary compiled especially for this purpose. Although the size of the evaluation experiments is small, the system achieved a good level of performance. It would be interesting to compare this approach with a machine learning approach in the future.

[9] There is no clear consensus on the language family of Japanese and Korean. We follow a position popular among theoretical linguists to classify both of them in the Altaic family. Note that Ethnologue lists Japanese as a separate family, while Korean is listed as an isolate and non-affiliated language.

DOI: 10.1007/s10579-007-9041-9

The paper "*The Hinoki syntactic and semantic treebank of Japanese*" by Bond et al. describes *the Hinoki Treebank*, a corpus annotated with syntactic and semantic information. There are three notable features of *the Hinoki corpus*. First, the annotation is based on well-established theories: Head Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005). The corpus is comprised of definition sentences from a print-based dictionary. In this respect, it is similar to MindNet[10] by Microsoft Research, but *the Hinoki corpus* adds more detailed information using these theories. Second, it is tightly integrated with other language resources they have created, including Lexeed (a dictionary defining word senses of basic Japanese words), JACY (HPSG-based grammar of Japanese), and a thesaurus derived from Lexeed. Third, the annotation includes statistical information such as frequency and familiarity of words derived from psychological experiments, as well as symbolic and structural information. These features reflect their ultimate goal to integrate lexical semantic and statistical information. Although readers must refer to the author's other papers for full details of each resource, this paper gives a good overview of the resource and the methods by which it was constructed.

The paper "*Dependency parsing of Japanese monologue using clause boundaries*" by Ohno et al. describes a modern version of KNP, which breaks a long sentence into shorter sentences using its parallel structure. Ohno's method utilizes additional clues to divide sentences. In addition, they incorporate statistical information in their dependency parser. They develop the system aiming at parsing transcribed texts of monologue speech and evaluated it by using a monologue corpus. It will be relevant to look at the improvement in performance when their method is applied to written texts such as newspaper articles, on which many parsing systems were already evaluated.

### 4.2.2. *Korean*

The paper "*A segment-based annotation tool for Korean treebank with minimal human intervention*" by Park et al. follows the most conventional approach for syntactic annotation of Korean texts, that is, manual correction of the parser's outputs. However, processing is broken down into two stages: the inside-structure analysis and the outside-structure analysis. Human annotators intervene after each of these processing stages. The paper claims that such a two-stage approach is more effective and efficient in corpus building. In fact, the same approach was taken in the paper by Ohno et al., but their goal was automatic parsing of Japanese texts rather than corpus building. It is important to note that Park et al.'s divide-and-conquer approach is effective in both corpus building and parsing in different languages.

[10] http://research.microsoft.com/nlp/Projects/MindNet.aspx

DOI: 10.1007/s10579-007-9041-9

### 4.3. Austro-Asiatic Languages

Ethnologue identifies 168 languages in the Austro-Asiatic family, with two branches: Mon-Khmer with 147 languages and Munda with 21 languages, which are spoken in South and Southeast Asia. Vietnamese is one of the few languages that have a long recorded history and has the highest population in the language family. One of the papers in this volume focuses on this representative language.

#### 4.3.1. *Vietnamese*

In the paper "*A lexicon for Vietnamese language processing*", Nguyen highlights the importance of the reusability of linguistic resources and their comparability in a multilingual model. The language on which Nguyen draws is Vietnamese, which has been rarely treated in the NLP literature. Her goal is to build a Vietnamese linguistic database that can be openly and fully used for NLP applications. In particular, she provides a detailed account of syntactic information in eight categories (noun, verb, adjective, pronoun, adjunct, conjunction, modal particle, and interjection) in Vietnamese. Such a description is considered valuable for tagset definition and morphosyntactic analysis. This paper therefore makes a strong contribution to the development of Vietnamese language processing.

### 4.4. Indo-European Languages

As the largest language family, the Indo-European language family includes more than 700 languages, spreading throughout Europe, South, Southwest, and Central Asia. Its largest branch is Indo-Iranian languages including Bengali, Hindi, and Marathi. Two papers in this volume contribute to these languages in this Indo-Aryan branch.

#### 4.4.1. *Bengali*

Bengali is a member of the group of highly inflectional languages which lacks automatic processing tools due to scarce resources. Dasgupta and Ng address the need for developing automatic tools for the language. In the paper "*Unsupervised morphological parsing of Bengali*", they propose a morphological parser using an unsupervised approach to tackle the well-known word segmentation problem in Bengali. Unlike previous knowledge-based algorithms, the unsupervised parser requires less time and linguistic expertise. Its high level of performance is attributed to the use of relative frequency information and composite suffix detection technique. This work makes a significant contribution to the development of language processing in Bengali and other Indo-Aryan languages.

DOI: 10.1007/s10579-007-9041-9

Language Resources and Evaluation (2006) Vol. 40, No. 3-4, pp. 208-218

### 4.4.2. *Hindi and Marathi*

Bhattacharyya and colleagues, in their paper "*Complex predicates in Indian language WordNets*" discuss their observations in the process of building Hindi and Marathi WordNets. Their focus lies in the treatment of complex predicates, a common linguistic phenomenon found in all Indian languages, which, they argue, is not accounted for in Princeton WordNet. They address this deficiency by offering a formal and systematic way of deciding whether a particular complex predicate should be included in lexical knowledge base as well as identifying complex predicates. A potentially important contribution when these analyses are implemented computationally is the automatic augmentation of different language WordNets.

## 4.5. Sino-Tibetan Languages

The Sino-Tibetan languages form the second largest language family, which are mainly spoken in East Asia. There are two main branches: Chinese and Tibetan-Burman languages. Chinese and its dialects have the largest number of speakers among all branches. Four papers in this volume address the language processing issues in Chinese.

### 4.5.1. *Chinese*

In their paper "*Copy detection in Chinese documents using the Ferret*", Bao and colleagues apply the well-known Ferret copy detector, which works effectively in detecting plagiarized material on English texts, to Chinese. They reveal that an adapted version of the Ferret achieves consistently good performance on Chinese texts. Although further modification is needed, this system will serve as the pioneer in Chinese copy detectors, while its investigation of the Ferret will be of great importance to developing copy detectors in other languages.

Word abbreviations have always been a problem to Chinese language processing. In the paper titled "*Mining atomic Chinese abbreviation pairs with a probabilistic single character word recovery model*", Chang and Teng attempt to solve the problem by designing a model for finding the root forms of the finite Chinese character set. By adapting the unified word segmentation model, they develop an HMM-based Single Character Recovery (SCR) Model extracting a large set of abbreviation-root pairs from a text corpus. The model achieves promising results in which the precisions are 50% and 62% for the test set and training set respectively.

The paper "*Word sense learning based on feature selection and MDL principle*" by Ji and colleagues recognizes the importance of automated learning of word senses in the field of information retrieval and machine translation. They argue that the two approaches to the analysis of word senses, namely Committee-based Method and Context-Group Discrimination, are in-

DOI: 10.1007/s10579-007-9041-9

sufficient. Instead, they design a word sense learning algorithm based on feature selection and cluster number identification. Such an algorithm is shown to be reliable in automatically retrieving important features and estimating the cluster numbers.

Xue, in the paper titled "*A Chinese semantic lexicon of senses and role*", proposes a Chinese semantic lexicon for the purpose of supporting the predicate-argument annotation of Chinese verbs and their nominalizations. He demonstrates how essential coarse-grained sense distinctions may specify semantic roles and how the semantic roles are realized. In addition to describing Xue's ongoing project, Chinese PropBank, this lexical semantic lexicon should raise interesting discussions for high-level semantic generalizations in the future.

### 4.6. Tai-Kadai Languages

There are 76 languages in the Tai-Kadai family, which is distributed in mainland south-east Asia and in Southern China. These languages are typically tonal languages. All languages in this family, except for Thai, are minority languages in the country where they are spoken.

#### 4.6.1. *Thai*

The work of Collier and colleagues titled "*A multilingual ontology for infectious disease surveillance: rationale, design and challenges*" exposes the need for developing a new surveillance system for monitoring early developments of spreading diseases in Asia-Pacific countries. The authors regard the availability of multilingual terminological resources as one of the crucial factors in significantly improving the disease surveillance system. As the first step of their project (BCO), they concentrate on the discussion of building a multilingual ontology including English, Chinese, Korean, Japanese, Thai, and Vietnamese in the paper. The ontology is expected not only to support the surveillance system as a whole, but also to bootstrap the development of monolingual biomedical text mining systems for Asia-Pacific languages.

## 5. Conclusion

As summarized above, the fourteen papers collected here draw a vibrant and fast-developing picture of research on Asian language processing, regardless of whether the target language is well-computerized or not. The collective diversity offers both a challenge and an opportunity to descriptive, theoretical, and computational linguists. Most crucially, these studies underline that the synergy of succinct formulation of felicitous linguistic description and optimal application of processing models is the key to successful Asian language processing. We hope that the work presented here will presage a new era of

DOI: 10.1007/s10579-007-9041-9

Language Resources and Evaluation (2006) Vol. 40, No. 3-4, pp. 208-218

human language technology where all languages as well as the knowledge they carry can be processed and accessed equally and openly.

## 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, Dain Kaplan at Tokyo Institute of Technology, 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. We would also like to thank Aravind Joshi, Pushpak Bhattacharyya, Benjamin T'sou, and Jun'ichi Tsujii for making their panel materials accessible to us. Any remaining errors are, of course, ours.

## References

Baeza-Yates, R. and B. Ribeiro-Neto: 1999, *Modern Information Retrieval*. Addison Wesley.

Bhattacharyya, P.: 2006, 'Can the availability of detailed linguistic information, say morphology, help in ameliorating the scarcity of large annotated corpora?'. In: *COLING/ACL 2006*. Sydney. Panel Presentation at the Panel: *Challenges in NLP: Some New Perspectives from the East*.

Brown, P. E., V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer: 1993, 'The Mathematics of Statistical Machine Translation: Parameter Estimation'. *Computational Linguistics* **19**(2), 263–311.

Butt, M. and T. King: (to appear), 'Urdu in a parallel grammar development environment'. *Language Resources and Evaluation*. A spesial issue: New Frontiers in Asian Language Resources.

Copestake, A., D. Flickinger, I. A. Sag, and C. Pollard: 2005, 'Minimal Recursion Semantics: An introduction'. *Journal of Research on Language and Computation* **3**(2-3), 281–332.

Fellbaum, C.: 1998, *WordNet: An Electronic Lexical Database*. The MIT Press.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, and C. Pet, M. Soria: 2006, 'Lexical Markup Framework (LMF)'. In: *Proceedings of LREC 2006: 5th International Conference on Language Resources and Evaluation*. pp. 233–236.

DOI: 10.1007/s10579-007-9041-9

Gordon, R. G. J. (ed.): 2005, *Ethnologue: Languages of the World*. SIL International, 15 edition.

Hashimoto, S.: 1984, *Kokugohô Yôsetu (Elements of Japanese Grammar)*, Vol. II of *The Complete Works of Dr. Shinkichi Hashimoto*. Iwanami Syoten.

Huang, C., N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prévot (eds.): (to appear), *Ontologies and the Lexicon*, Cambridge Studies in Natural Language Processing. Cambridge University Press.

Huang, C., T. Tokunaga, N. Calzolari, L. Prévot, S. Chung, T. Jiang, K. Hasan, S. Lee, and I. Su: 2007, 'Extending an international lexical framework for Asian languages, the case of Mandarin, Taiwanese, Cantonese, Bangla and Malay'. In: *Proceedings of the First International Workshop on Intercultural Collaboration (IWIC)*. pp. 24–26.

Joshi, A.: 2006, 'Panel: *Challenges in NLP: Some New Perspectives from the East*'. In: *COLING/ACL 2006*. Sydney.

Karttunen, L. and J. McCarthy: 1983, 'A special issue on Two-level morphology introducing the KIMMO system'. *Texas Linguistic Forum* **22**.

Koskenniemi, K.: 1983, 'Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production'. Ph.D. thesis, University of Helsinki.

Kukich, K.: 1992, 'Techniques for Automatically Correcting Words in Text'. *ACM Computing Surveys* **24**(4), 377–439.

Kurohashi, S. and M. Nagao: 1994, 'A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures'. *Computational Linguistics* **20**(4), 507–534.

Nagata, M.: 1996, 'Context-based Spelling Correction for Japanese OCR'. In: *Proceedings of the 16th International Conference on Computational Linguistics*. pp. 806–811.

Nagata, M.: 1998, 'Japanese OCR Error Correction using Character Shape Similarity and Statistical Language Model'. In: *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*. pp. 922–928.

Pollard, C. and I. A. Sag: 1994, *Head-Driven Phrase Structure Grammar*. CLSI.

Tokunaga, T., V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Sonia, C. Huang, Y. Xia, H. Yu, L. Prevot, and K. Shirai: 2006, 'Infrastructure for Standardization of Asian Language Resources'. In: *COLING/ACL 2006*. pp. 827–834.

T'sou, B.: 2004, 'Chinese Language Processing at the Dawn of the 21st Century'. In: C.-R. Huang and W. Lenders (eds.): *Computational Linguistics and Beyond*. pp. 189–206, Language and Linguistics.

T'sou, B.: 2006, 'Some Salient Linguistic Differences in Asia and Implications for NLP'. In: *COLING/ACL 2006*. Sydney. Panel Presentation at the Panel: *Challenges in NLP: Some New Perspectives from the East*.

Tsujii, J.: 2006, 'Diversity vs. Universality'. In: *COLING/ACL 2006*. Sydney. Panel Presentation at the Panel: *Challenges in NLP: Some New Perspectives from the East*.

DOI: 10.1007/s10579-007-9041-9

Language Resources and Evaluation (2006) Vol. 40, No. 3-4, pp. 208-218

DOI: 10.1007/s10579-007-9041-9