Putting NLP Tools into Action: Incorporating NLP Tools in Web-based Language Learning Environment

Md Maruf Hasan, Kazuhiro Takeuchi, Virach Sornlertlamvanich, Hitoshi Isahara
Thai Computational Linguistics Laboratory
Communications Research Laboratory
112 Pahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120
mmhasan@crl-asia.org, kazuh@crl.go.jp,
virach@crl-asia.org, isahara@crl.go.jp

Abstract

Sophisticated Natural Language Processing (NLP) Tools have been developed over the last few decades for many languages. For example, the *ChaSen* Morphological Analyzer and the *KNP* Parser for Japanese; the *SWATH* Segmenter for Thai; and *CRL-New Mexico's* Segmenter for Chinese are being widely used by NLP researchers. Several copyright-free machine readable dictionaries (MRD) are also available now, and some of these MRDs also include pronunciations (sound files) and example sentences, etc. By putting a few NLP Tools together, it is possible to build smart applications that can enhance foreign language teaching and learning experiences. In this paper, we will investigate such applications in a Web-based foreign language learning environment.

1 Introduction

Over the last few decades, researchers in Natural Language Processing (NLP) have been actively developing tools to facilitate research in NLP. A plethora of sophisticated NLP tools are now available for many languages. Examples of such tools include the *ChaSen* Morphological Analyzer [1] and the *KNP* Parser [2] for Japanese; the SWATH Segmenter for Thai [3]; the *CRL-New Mexico's* Segmenter for Chinese [4], and so on. At the same time, copyright-free machine readable dictionaries (MRD) are also growing in number, size and quality. Examples of such dictionaries are EDICT (Japaense-English) [5], LEXiTRON (Thai/English) [3], CEDICT (Chinese-English) [6] and UniDict (multilingual) [7], etc.

Most of the NLP tools and resources are originally developed *by* the researchers *for* the (use of) researchers. Because of such design philosophies, installation, configuration and use (e.g., the user interface) of these tools remain to be tricky. In the last decade, with the growing popularity of the WWW, some researchers took further initiatives to develop Web-based interfaces for some of these tools and resources [3]. Such initiatives are widely welcomed by the researchers of non- computer science background (such as, the Linguists), because the availability of Web-based services freed the users (client-side) from the burden of installation, configuration and maintenance tasks.

In the Web-based environment, installation and configuration of a tool are done on the server-side by computer professionals or the developers themselves. Therefore, researchers such as Linguists, found it greatly useful in their research (e.g., in analyzing their own experimental data) since they only need to use a Web-browser to access the service. Although with the Web-based implementation NLP tools reached more researchers, general users are still left behind to take any direct advan-

tage from these NLP tools or their Web-based services because these tools and services are developed with a specific (often, a single) research problem in mind: for example, word-segmentation, or parsing etc. It is difficult to use such tools in real-life applications. For instance, a foreign language learner seeking helps in reading a document (preferably with some dictionary look-up based annotations) can't take any direct advantage of a standalone *segmenter* (for tokenization) and a standalone *dictionary lookup* service unless these two services are integrated and customized in such a way that it can offer the intended reading help.

For languages such as, Chinese, Japanese and Thai where word-boundaries are not explicit, a segmentation tool can help in identifying words or phrases (also known as tokenization) and subsequently, a dictionary look-up tool can annotate each word or phrase with their dictionary entries which may as well include pronunciation sound files and other usage notes. The combined result is to offer reading helps to non-native speakers. Study shows that, such an integrated service of a segmenter and a dictionary lookup tool offer useful help to a beginner and intermediate level learner of Chinese, Japanese and Thai. In contrast, it is unlikely (and also inconvenient) for a student to think of or use two different services separately to obtain the same level of reading aid. See *Appendix* for more detail.

In this paper, we will investigate the prospects and issues of integrating sophisticated NLP tools and resources together, and demonstrate that by putting the bits and pieces together, it is possible to build smart applications to expand the reach of NLP tools and technologies to a broader range of real-life applications. Our discussions will be limited to Web-based foreign language learning.

2 Motivation

In the recent past, when we introduced NLP research and development to groups of foreign language teachers and learners, they naturally asked how our research outcomes could enhance their teaching and learning experiences in the long run. Students often asked whether there are tools they can use to help them reading and understanding a piece of text which includes some unknown vocabularies and syntactic structures they are not yet familiar, or tools which can check potential grammatical mistakes (and better yet, suggest corrections) in their compositions before they hand in the compositions to the teacher, etc. It should be noted that proprietary software including Microsoft Word, offer some helps for only certain languages. Teachers often asked whether it is possible to correct students' compositions with online annotations and leave selected compositions (or part thereof) online to allow other students to participate in online discussions on particular aspects of language usage/error, or whether it is possible to use NLP tools to draw the syntactic tree and generate usage annotations of a particular sentence pattern for pedagogical purpose. Some teachers even asked whether it is possible to automatically analyze a student's essay or speech for placement decision making. That is, whether or not it is possible to analyze and score a student's composition or speech based on its language usages and errors for the placement of a student to a particular level of proficiency. Most NLP researchers will agree that with the current state-of-the-art NLP technologies, we can fulfill most of the above expectations to some extent.

NLP researchers generally admit that the NLP tools we develop and use are sophisticated but they are not originally developed for real-life applications in mind. They are rather developed as so-called research tools. With the time-constraint and overwhelming theoretical problems tosolve, we often have little time to pay attention to user-related issues, such as ease of installation and configuration, and friendly user interfaces. However, most researchers are aware that with careful revision and integration most of the tools can be used in practical applications.

In this paper, we will discuss a number of interesting Web-based applications which can be built by adapting NLP tools.

3 Advantages of Web-based Application

In Web-based applications, the problems with installation, configuration and maintenance remain transparent to the end-users because they are all

done at the server-side. Moreover, a collection of tools can be integrated transparently on the server-side to build practical applications. For instance, the Reading System explained in the Appendix of this paper uses a segmenter and a dictionary lookup tool together with a dictionary; and all the integration-related complexities also remain hidden to the users of the system.

There are also other potential advantages of implementing a server-based application, such as collecting *Learners' Corpora* as explained below.

Researchers in foreign language acquisition are in increasing need of *Learners' Corpora* from people of different language groups. *Learners' Corpora* ideally consists of text (or speech) written (or spoken) by people from a particular language group, and in which errors (and error-patterns) are also identified and annotated in specific formats. In classroom environment, with a centralized server, it becomes easier to gather *raw* Learners' Corpora by logging student's submissions. Teachers can subsequently (and hopefully, collaboratively) annotate such a raw corpus using an easy-to-use annotation tool (also deployed on the server, and accessed through browser-based interface). By means of active collaborations of foreign language learners, teachers and researchers, it becomes easier to capture and annotate *Learners' Corpora* using a classroom based server.

4 Potential Web-based Applications in Foreign Language Learning

The availability of free and open-source tools is a crucial in building integrated services and applications on top of those tools. We are analyzing and validating such tools for Chinese, Japanese and Thai. Similar initiatives for learning English are being investigated in ALLES: *Advanced Long-distance Language Education System*, project [10]. At present, we are investigating the following applications for the above-mentioned Asian languages.

1. **Reading Systems** where users are able to submit a piece of text they want to read for tokenization and dictionary-based augmentation: A submission typically involves cutting and pasting of a piece of text into a *HTML Form* of a Web browser, and a click of a button after choosing some processing and annotation options. The system accepts such submissions and subsequently performs tokenization, dictionary lookup and annotation. A foreign language learner may choose a specific annotation language (which is most likely his or her mother tongue) and the system displays annotated output by looking up the appropriate dictionary provided that such a dictionary is already made available to the system. Also, based on availability, the system may include links to the pronunciations (sound files) or other usage information for each token if the user so desire.
2. **Writing Validation Systems** where end-users (typically, foreign language learners) submit their own compositions and essays for checking spelling, grammar etc.; and receive feedback with highlights on potential errors, and hints on correct usage: The submission process is similar to that of a reading system. However, the processing also involves parsing of each sentence to validate it or to make suggestions. The system also logs submissions of users for the purpose of building Learners' Corpora.
3. **Authoring Systems** where teachers analyze their existing teaching materials and decompose them into *Learning Objects* (LOs): The Learning Objects approach [8] is inspired by the *object-oriented* technology. Learning Objects are independent and reusable units of instructional materials with specific learning goals. LOs are typically designed through analysis and decomposition of traditional instructional materials. It is envisaged that NLP Tools will play vital roles in analyzing traditional instructional materials in the context of language learning. Such analysis subsequently helps teachers in decomposition decision-making. As a simple example, let us consider that a teacher wants to develop LOs for beginner's level learners of Japanese. A set of NLP tools including a *Kanji*-filtering

tool and a parser may list all the less-frequent *Kanji* and undesirable syntactic structures (respectively) accidentally appeared in the original materials, and alert the teacher. In personalized e-learning scenario, students or teachers assemble LOs into customized course-materials which best suit their needs. Therefore, extensive care must be taken to maintain independency and reusability in authoring LOs.

4. **Discussion and Annotation Systems** where end-users choose a particular sentence or an entire composition to discuss different issues in language usages and errors: Discussions and annotations could be done in natural language. However, to build Learners' Corpora, annotation in specific formats is required. Careful readers might have already noticed that such systems make extensive use of online collaboration tools (e.g., D3E [9] or similar other collaborative discussion systems). NLP tools only play secondary roles in this type of applications.

As of writing this paper, we have successfully deployed a combination of NLP tools and resources in developing prototypes of a Japanese Reading System and a Thai Reading System (c.f., Appendix of this paper). A similar Chinese Reading System is currently under development.

Writing Validation Systems are more demanding in terms of tools. The unavailability of an open-source parser for Thai and Chinese makes it difficult to implement such systems in these languages. However, we are looking forward to implementing a Writing System for Japanese in the near future.

In the third type of applications (Authoring Applications) NLP researchers, LO specialists and foreign language teachers need to work together closely. It is most likely the case that new NLP tools based on the state-of-the-art NLP technologies need to be developed for decomposition and authoring of LOs from existing instructional materials that suit the teachers' needs.

The fourth type of application (Discussion and Annotation Systems) is important in the context of e-learning and annotating Learners' corpora. However, at the moment we are not making it a priority.

5 Design Goals

Throughout the design and implementation of web-based applications for foreign language learning using NLP and other tools, we have given high priority to the following design goals:

1. Putting sophisticated NLP tools in practical application, such as in reading and writing systems
2. Freeing the users from installation and configuration related complexities by offering Web-based services. Such servers are configured with all the required tools and made available to the WWW or a particular group of students.
3. Presenting users with only human-understandable information: Outputs from NLP tools and language resources are often meant for machine to understand. For example, the Part-of-Speech tagset of a parser or the entries in a machine readable dictionary includes many details necessary for computer to make sense of the information. Although we internally take advantage of such machine-understandable information, when generating outputs for human, we exclude the details or map/convert them into human-understandable form.
4. Modular system architecture to facilitate multilingual extension: That is, when a Tokenizer, Dictionary and other resources and tolls are available for other languages, they can be easily integrated into the server.
5. Use of *Learning Objects* technologies in developing reusable instructional materials for foreign language for personalized and ubiquitous e-learning. We hope to maintain a (collaboratively built and maintained) Learning Object Repository (LOR) with LOs in several Asian languages on a central Web server.

One point to note here that humans are well-known as goal directed agents who actively seek knowl-

edge. They come to formal education [and training] with a range of prior knowledge, skills, beliefs and concepts that significantly influence what they notice about the environment, how they organize and interpret it, and how they make a learning plan. The *Learning Objects* approach facilitates personalized learning experience by making use of the goal-directed nature of humans and reusable independent learning objects.

If students or teacher are presented with the hierarchical views of the learning objects available in LO Repository, they may be better able to select and assemble their own instructional materials easily. And finally, in the process of learning, learners may make use of several other services, such as reading system, writing validations systems, etc. which will add values in their overall e-learning experience.

6 Conclusions

In this paper, we have analyzed a series of applications in the context of foreign language teaching and learning, where NLP tools play a significant role. With our experience in developing prototype Thai and Japanese reading systems (explained in the Appendix), we also pointed out key design issues in similar applications.

Due to the globalization and the development of technologies, the foreign language learning scenario is changing rapidly in the recent years. Nowadays, majority of foreign language learners are adult professionals, and they prefer to learn a foreign language with particular objectives in mind. They prefer to have (or sometimes, insists on having) personalized teaching materials and supporting tools in achieving their unique learning objectives. Majority of learners prefer (or force) to learn a language at a distance with minimum or just-in-time instructor's interventions. Applications such as, reading systems, writing validations systems are the preferred tools for many foreign language learners in such a learning scenario [11].

Foreign language teachers are also in need of course authoring tools, such as tools for analyzing traditional instructional materials and converting them into reusable LOs or other tools to develop customized teaching materials for individuals or groups.

Online collaborative teaching and learning of foreign language is another area where foreign language teachers and computer scientists including LO specialists and NLP researchers may find plenty of new opportunities.

In our recent collaborations with foreign language learners and teachers, we gained deeper insights and ideas about what we, NLP researchers and foreign language teachers/learners have in common, and what we can complement. Based on those insights, we identified and analyzed a set of new applications in this paper.

References

[1] Y. Matsumoto, H. Kitauchi and T. Yamashita. 1997. User's *Manual of Japanese Morphological Analyzer*, ChaSen version 1.0. IS-TR97007, Nara Institute of Science and Technology, Japan, (in Japanese)

[2] S. Kurohashi and M. Nagao. 1994. *A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures.* Computational Linguistics, Vol 20, No. 4.

[3] NECTEC: *NECTEC Free Web Services*, includes Thai Word-Break and LEXiTRON Dictionary Services. http://www.nectec.or.th/services/

[4] L. Chen and W. Jin. 1996. *A Chinese Text Display Supported by the Chinese Segmentation Algorithm.* In: Elisa M. del Galdo and Jakob Nielsen (eds.) International User Interfaces. John Wiley, pp. 151-177

[5] *The Edict Project Home Page*, http://www.csse.monash.edu.au/~jwb/edict.html

[6] *CEDICT: Chinese-English Dictionary* http://www.mandarintools.com/cedict.html

[7] *The UniDict Project Home Page* http://www.mandarintools.com/cedict.html

[8] David A. Wiley Eds. *The Instructional Use of Learning Objects*, Online Book, available at, http://www.reusability.org/read/

[9] D3E: *Digital Document Discourse Environment* http://d3e.sourceforge.net/

[10] Advanced Long-Distance Language Education System, http://alles.sema.es/

[11] H. Mochizuki and A. Tera, Constructing Web-based Japanese Text Reading Support System and Its Evaluation, International Conference on Computers in Education ICCE2002, pp. 1478-1479

APPENDIX

Examples of Web-based Language Learning Applications using Popular NLP Tools:

Thai and Japanese Reading Systems:

For languages such as, Thai and Japanese where word-boundaries are not explicit, a segmentation tool helps in identifying words or phrases (tokens); and subsequently, a dictionary lookup tool further augments each token with relevant dictionary entries. In our Web-based Reading Systems, users cut and paste a piece of text into a HTML form (and choose some processing options). The user input is submitted to the server (configured with the above-mentioned NLP tools and dictionaries). The server then performs tokenization and dictionary-based augmentation of the input text and sends the output back to user. Following Figures (Figure 1 and 2) are showing the snapshots of a Thai and a Japanese Reading System, respectively.

Notice that placing mouse over a token displays annotation. Tokens are also listed separately in a separate frame. If available in the dictionary, users may also play sound files by clicking on the speaker icon (c.f., Figure 1). Also, note that in our systems only Thai-English and Japanese English dictionaries are used and therefore, the annotations are made in English. By using a Thai-Japanese and Japanese-Thai dictionary, the system can generate annotation in Japanese for Thai learners, and vice versa.

Study shows that, such an integrated service of a segmenter and a dictionary lookup tool offer useful helps to beginner and intermediate level learners of Japanese and Thai language. In contrast, it is unlikely (and also inconvenient) for a student to think of or use two different services separately to obtain the same level of reading aid.

Figure 1: Thai Reading System

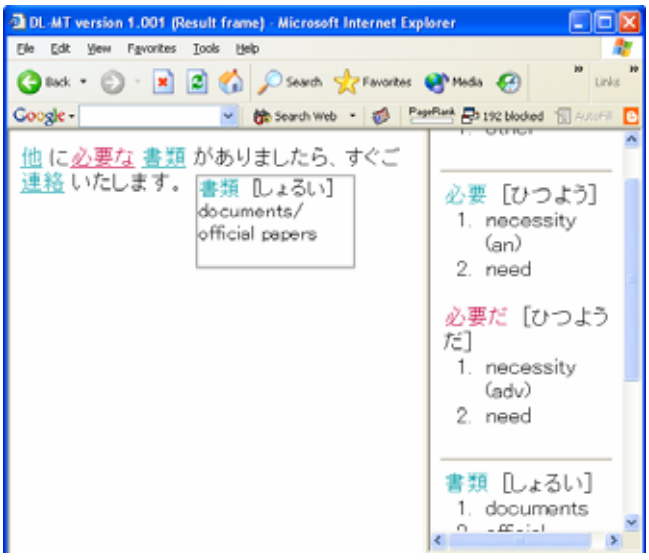

Figure 2: Japanese Reading System