Research Report on
Local Language Computing:
Development of Indonesian Language Resources and Translation System

Prepared by :

Mirna Adriani
University of Indonesia

Hammam Riza
IPTEKNET - BPPT

Contents

I. Background

The Republic of Indonesia is an archipelago of 13,579 islands that spread over an area of 1,900,000 square kilometers. It is the world's fourth largest nation with a population of 245,452,739 (July. 2006 estimated). Prior to the monetary crisis in 1997, Indonesia experienced a remarkable economic growth, while a rate of 7% growth of the Gross Domestic Produce (GDP) was recorded on per year average. Since the aftermath of the 1997 crisis, the current indicators have shown that Indonesian economy and political conditions are gradually stabilizing. Although some sectors still require special attention and empowerment, the overall condition show that Indonesia is back on the track to become an industrialized nation.

Bahasa Indonesia is the national language of Indonesia which is spoken by more than 222 millions people (Etnologue, 2005). It is widely used in Indonesia to communicate in school, government offices, etc. Bahasa Indonesia became the formal language of the country, uniting its citizens who speak different languages. There are more than 742 distinct languages in the country (Etnologue, 2005) which are still used widely in various regions of the country. Bahasa Indonesia has become the language that bridges the language barrier among Indonesians who have different mother-tongues. The vocabulary of bahasa Indonesia has been extensively influenced by outside languages, especially Sanskrit, Arabic, Chinese, Dutch, and English, as well as local languages such as Javanese and Batavian. European influence on syntax has also been considerable. A national language commission has existed since shortly after independence with the dual task of developing Indonesian as a language able to cope with a full range of technical and philosophical topics and of protecting the language against unwanted change, including outside influence. Despite these efforts, however, vocabulary in Indonesian changes rapidly, and urban dialects, incomprehensible to standard speakers, develop and disappear rapidly.

There is no universally accepted theory that explains the origin of the Indonesian people, nor the classification of Bahasa Indonesia, although it is generally agreed that bahasa Indonesia belongs to the Austronesian family that extends from Madagascar to

Polynesia. We can classify Bahasa Indonesia into Western Malayo-Polynesian group where there are 175 languages in the group (Riza, 2003). The written and spoken form is based on the Malay trade dialect that was used throughout the region in the past. However in the last few years, the influence of English has enormous impact on the language with the growth of western entertainment industry in the country. Many English words have been adapted into Indonesian words and used widely. This situation occurred not only in conversations, but also in the media. The increase in the number of English words used in the Bahasa Indonesia has raised many concerns about the future of the language. There are still many aspects of the language that have not been studied much yet, partly because there has not been much attention on the importance of the language.

In the era of globalization, communication among languages becomes much more important. Computer as a tool to help facilitating communication between different languages has been used more actively. People has been hoping that natural language processing and speech processing, which are branches of artificial intelligence with ICT (Information and Communication Technology), can assist in smoothening the communication among people with different languages. However, especially for Indonesian language, there were only few researches on natural language processing and speech processing in the past.

II. Project Summary

The project is to build Indonesian language resources and Indonesia-English translation system, using a statistical machine translation approach. In summary, scope of work (SOW) for this project consists of the following modules:

1. Corpus Collection and Translation (1000K words)
2. Manual Tagging (100K words)
3. Development of Part of Speech Tagger and Tagging of 1000K words
4. Lexicon (10K words)
5. English-Indonesia Statistical Machine Translation system

Statistical method or combination of statistical and symbolic method (using the parallel corpus and part of speech tagger) will be employed to achieve the above scope of works.

III. Project Plan and Research Goals

The research goals of the Indonesian project are:

a. Build text corpora

As a language is dynamic and constantly evolving, it is essential that the constructed linguistic resources are based on empirical evidence. To support this, the first phase of the work involves the compilation of corpora, i.e. bilingual collection of parallel English and Indonesian texts. The choice of documents is affected by the intended coverage of the linguistic resources. In this project, 1 Million words parallel corpus (Bahasa Indonesia and English) will be constructed from 500K words of Penn Tree-Bank corpus, licensed from PAN Localization project and 500K words of Indonesian news articles. The parallel corpus will contain approximately 100,000 sentences in each language.

b. Build a Part of Speech Tagger (POSTAG) for Bahasa Indonesia

POSTAG tagger: a set of morphological rules especially for derivations, described in a well-understood computational model. Implementation of part of speech tagger using available statistical and symbolic toolkit. Using this POS tagger, we will produce 1000 K Words Automatic Tagged Corpus

c. Build a machine translation system for translating English into Bahasa Indonesia

A machine translation system I/ETS, a web based system that translates English text into Indonesian using statistical method

The research plan is based on the project life cycle that spans across a 18-months period. Project management and administration by the Project Executing Agency (PEA) will be conducted throughout the project life cycle in the forms of coordination, supervision, monitoring and reporting. The project will be deployed in six (6) phases, each of 3 months length. After 1 month work on detailed requirement analysis, actual corpus collection will be started. In second phase, the POS tagger and machine translation including language model and translation model will start to cover the parallel corpus and domain specific requirement (e.g. WSJ, business domain).

Table 4-1 Deliverable Work for 18 months (Januari 2008-July 2009)

Timeline	University of Indonesia	IPTEKNET (BPPT)
1st Quarter	▪ Initial Research Report on Corpus Design and Collection and Cleaning Tools	▪ Initial Research Report on Corpus Design and Collection and Cleaning Tools
2nd Quarter	▪ Research Report on Translation, Alignment, and Other Issues Related to Parallel Corpus Development from English to Bahasa Indonesia ▪ POS Tagset of Bahasa Indonesia (Summary and Detailed Document) ▪ 100,000 Word Bahasa Indonesia Translated Text Corpus from Penn-Tree Bank	▪ Research Report on Translation, Alignment, and Other Issues Related to Parallel Corpus Development from English to Bahasa Indonesia ▪ 100,000 word Indonesian Corpus (with Creative Commons IPR) ▪ 100,000 word Parallel Translated English Corpus ▪ Initial Design Report on Statistical Machine Translation Framework
3rd Quarter	▪ Additional 150,000 words translated from PennTree Bank into Bahasa Indonesia ▪ 50,000 words of Bahasa Indonesia corpus manually tagged ▪ Research report on Tagset and Corpus tagging process	▪ 150,000 word Indonesian Corpus (with Creative Commons IPR) ▪ 150,000 word Parallel Translated English Corpus ▪ Final Design Report on Statistical Machine Translation Framework

5th Quarter	▪ Additional 100,000 words translated from PennTree Bank into Bahasa Indonesia ▪ POS Tagger for Bahasa Indonesia ▪ Research Report on POS Tagset, POS Tagging and POS Tagger for Bahasa Indonesia	▪ Additional 100,000 word Indonesian Corpus (with Creative Commons IPR) ▪ Initial Prototype of Statistical Machine Translation for Bahasa Indonesia
6th Quarter	▪ 1 Million POS Tagged Corpus of Bahasa Indonesia	▪ SMT from English to Bahasa Indonesia ▪ Research Report on SMT for English to Bahasa Indonesia

IV. Research Findings

a. Building Text Corpora

During this project, we specifically study **the policy for collecting the Indonesian corpus from various available sources, such as news agency, online media publisher, internet blog, websites and others.** Our team have look into the legal issues of creating Indonesian corpus. Based on the information in Wikipedia (http://en.wikipedia.org/wiki/Creative_Commons_licenses), and also observe the way Wikipedia Commons (refer to http://commons.wikimedia.org/wiki/Main_Page) which shape up the licensing of Wikimedia Projects. It is said that all text on all Wikimedia projects, except for Wikinews, is owned by the original writer and licensed under the GNU Free Documentation License (GFDL) and Wikinews is licensed under the Creative Commons Attribution 2.5. The Indonesian corpus is then translated into English manually.

Beside the Indonesian corpus, we also translate the English corpus into Indonesian. The English corpus which is Penn Treebank Corpus is translated into Bahasa Indonesia manually by several translators from the Faculty of Arts, University of Indonesia. Then the translated documents are evaluated by a senior linguist who reviews the result of the translation work. Some of the Indonesian texts are then annotated manually by some linguists. The parallel corpora are used in developing the part-of-speech tagger for Bahasa Indonseia and the machine translation tool.

b. Building Part of Speech Tagger (POSTAG)

Our work develops Part of Speech Tagging for Bahasa Indonesia using statistical approaches. We use Conditional Random Fields (CRF) and Maximum Entropy methods in assigning the tag to a word. We use a tagset containing 37 part-of-speech tags for Bahasa Indonesia. In this work we compared both methods using using two different corpus. The results of the experiments show that the Maximum Entropy method gives the best result.

c. Building a machine translation system

In this PANL project, we do an experiment with Open Source system for English-Indonesian translation system. Hence, the main finding is the ability to build an Open Source Software for Statistical machine translation (SMT) to achieve and perform at least similar to a proprietary commercial MT system, such as Kataku, SMART Translator and TransTools. In fact, in some cases, we perform better than Google Translator for English to Indonesia. We computed all system on the 500.000 word into 9 block of word Indonesia-to-English and English-to-Indonesia, and computed the Bleu Scored on these blocks individually. The first observation is that was very significantly increase the Bleu score.

V. Project Output

There are several outcomes of the research projects:

a. Two corpora are ready to be used for further researches. The two corpora are parallel text corpus (English and Indonesian) and annotated corpora (using Penn Tree-Bank tagset).

b. An Indonesian part-of-speech tagger.

The POS tagger is ready to be used

c. A machine translation system for translating English into Bahasa Indonesia. Statistical machine translation is the art technique based on sentence-level aligned parallel corpora. We summarize the current strategies fetching parallel corpora to the Web and classify them into two classes English-to-Indonesia and Indonesia-to-English. There is available on www.translator.iptek.net.id/PANL.

d. Publication based on the topic of the research project:

Adriani, Mirna, Ruli Manurung, and Femphy Pisceldo. **Statistical Based Part Of Speech Tagger for Bahasa Indonesia.** Third MALINDO International MALINDO Workshop, Co-located Event ACL-IJCNLP 2009. Singapore, August 1, 2009.

Budiono, Hammam Riza, Chairil Hakim. **Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System**, 7th Workshop on Asian Language Resource, ACL-IJCNLP 2009, Singapore, August 2009.

VI. Sustainability of work and Future Prospect

a. Human Resource Capacity Building

In this PANL project, our objective is not only to build a translation system for Bahasa Indonesia to English, but we also construct and develop a team of various competence: computer science, computational linguistic, corpus linguistic, translation and project management. We need to build a good quality of people in order to perform all necessary tasks related to collecting and processing Parallel Corpus Indonesia-English. Several people have contributed to collecting Indonesian corpus and perform raw corpus cleaning, manual translation, alignment and XML tagging.

Human resource development was also seen in designing and developing statistical machine translation component based on Open Source Software. The capacity that have acquired by these software engineers have shown our

effort for sustaining the research and development in language technology, especially in machine translation system.

b. Adoption of Local Language Computing by End Users

The outcomes of the project have several benefits:

1. Helps Indonesian users to understand English sentences fast and correctly by using the machine translation system.
2. Makes the language resources such as corpus (monolingual corpus, parallel corpus) and lexicon to be used for further researches.
3. Provides ready-to-use modules for specific applications such as Part of Speech (POS) tagger and machine translation that can be utilized as language processing toolkit for linguists, further research collaboration, etc.

c. Influence on Language Computing Policy

The availability of the Indonesian language resources and tools has motivated some computer scientists and linguists to do further work not only on Indonesian resources but also on local languages. We have work together with the government (Ministry of Communication and Informatics) to promote the research project by organizing some workshops in four different locations in the country. This project has opened many possibilities for collaboration between many institutions in Indonesia.

d. Beyond PAN Phase II

Currently there are many interesting topics needed to be explored not only for Indonesian but also for our local languages. We have about 742 languages in the country so the methods that we have used for developing language resources and tools can also be used in developing local language resources and tools. Furthermore this project ha made us possible to have many colleagues from Asia who work on the same topics. Hence we have some

kinds of network to share similar problems and experiences in working on the project. We also benefit from knowledge sharing by attending the PAN workshops. In the near future, we hope that the project is still able to be funded so that the research networks can still continue and open many collaboration between the PAN members.