Toward Asian Speech Translation System: Developing Speech Recognition and Machine Translation for Indonesian Language

Hammam Riza
IPTEKNET
Agency for the Assessment and
Application of Technology
Jakarta, Indonesia
hammam@iptek.net.id

Oskar Riandi
ICT Center
Agency for the Assessment and
Application of Technology
Jakarta, Indonesia
oskar@inn.bppt.go.id

Abstract

In this paper, we present a report on the research and development of speech to speech translation system for Asian languages, primarily on the design and implementation of speech recognition and machine translation systems for Indonesia language. As part of the A-STAR project, each participating country will need to develop each component of the full system for the corresponding language. We will specifically discuss our method on building speech recognition and stochastic language model for statistically translating Indonesian into other Asian languages. The system is equipped with a capability to handle variation of speech input, a more natural mode of communication between the system and the users.

1 Introduction

Indonesia is one of the ten most populous nations in the world with the population of about 235 million people as of 2004 and is located strategically within the Asia region. The exchange of people, goods and services as well as information increases and should not be hindered by language barrier. Even though, English language may be used as the main global communication language, the more direct and more natural way of communication is preferred by local and native people to ensure the smooth exchange of information among people of different languages.

It would be beneficial for Indonesia people, if there were a system that is able, to some extent in a certain domain, to capture either a speech or digital text based on Indonesian language and process it in order to output into meaningful text into other languages such as English, Japanese and other world languages. In addition to above mentioned benefit, large numbers of Indonesian people, statistically, have problem in using and comprehending any information presented in English language, The language barrier problem is compounded by the problem of the explosion of digital information whose majority uses English language via either Internet or any digital / printed form which may overwhelms potential users and pose a threat of inequality of access of information due to the language barrier (digital divide) especially for the common Indonesian people. We are now part of a multi national project to develop speech to speech translation system for Asian languages facilitated by ATR-Japan.

Our most recent work is focusing on developing Indonesian speech recognition engine and a statistical language model for machine translation. Our approach to MT is based on the integration of stochastic and symbolic approaches to be used for analyzing Indonesian. For creating the stochastic language model, it is worthwhile to utilize annotated data when it is available and use supervised learning mechanism to estimate model's parameter. In this case, an annotated corpus is created for multiple genres of documents. Of course, the costs of annotation are prohibitively labor intensive, and the resulting corpora sometimes are susceptible to a particular genre. Due to this limitation of annotated corpora, it is necessary that we use unsupervised and weakly supervised learning techniques, which do not require large, annotated data sets.

Unsupervised learning utilizes raw, unannotated corpora to discover underlying language structure such as lexical and contextual relationships. This gives rise to emergent patterns and principles found in symbolic systems. In this system, the language model is trained using weakly supervised learning on small annotated corpus to seed unsupervised learning using much larger, unannotated corpora. Unsupervised and weakly supervised methods have been used successfully in several areas of NLP, including acquiring verb sub-categorization frames, part-of-speech tagging, word-sense disambiguation and prepositional phrase attachment.

The significant contribution of this preliminary research is the development of ASR using speaker adaptation technique and a statistical language model for translating from/to Indonesian language as well as Indo-Malay language in some extent. The major language found in Indonesia, Malaysia, Brunei, Singapore, Southern Thailand and Philippines can be categorized into a single root Indo-Malay language spoken in different dialects. Creating an ideal language model for Indo-Malay language is expected to be used by more than 260 million people in the region.

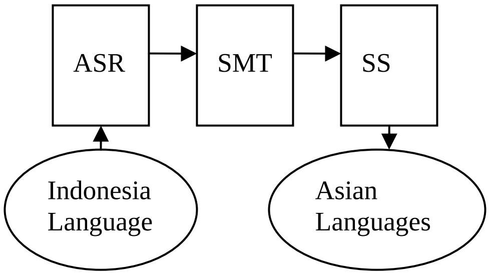

Figure 1. Scope of Indonesian-Asian Languages

2 Recognizing Indonesian Speech

Achievement of a high performance is often the most dominating design criterion when implementing speech recognition system. The current state of the art speech recognition technology is able to produce speaker independent recognizers which have extremely high recognition rates for small/medium vocabularies.

Although the average recognition rates are high, some speakers have recognition rates considerably worse than others. It is generally agreed that speaker dependent system will give the best performance in applications involving a specific speaker. This requires, however, that enough training data is available for training the system from scratch. An often used solution is to train speaker independent system using data from many speakers. But other experiments have shown that using such systems, in general, involves obtaining a lower performance than what is achievable with a speaker dependent system. This problem can be overcome, at least partially, by using speaker adaptation techniques, the aim of which is to take an initial model system which is already trained, and use a sample of a new speaker data to attempt to improve the modeling of the speaker with the current set of the model.

By collecting data from a speaker and training a model set on this speaker's data alone, the speaker's characteristics can be modeled more accurately. Such systems are commonly known as *speaker dependent* systems, and on a typical word recognition task, may have half the errors of a speaker independent system. The drawback of speaker dependent systems is that a large amount of data (typically hours) must be collected in order to obtain sufficient model accuracy. Rather than training speaker dependent models, *adaptation* techniques can be applied. In this case, by using only a small amount of data from a new speaker, a good speaker independent system model set can be adapted to better fit the characteristics of this new speaker.

Speaker adaptation techniques can be used in various different modes. If the true transcription of the adaptation data is known then it is termed *supervised adaptation*, whereas if the adaptation data is unlabelled then it is termed *unsupervised adaptation*. In the case where all the adaptation data is available in one block, e.g. from a speaker enrollment session, then this termed *static adaptation*. Alternatively adaptation can proceed incrementally as adaptation data becomes available, and this is termed *incremental adaptation*.

One of the researches on speaker adaptation techniques based on HMM is **Maximum Likelihood Linear Regression (MLLR)**. This method transforms the mean of continuous HMM. MLLR will generate a global adaptation transform when a small amount of data is available. While more adaptation data becomes available, improved adaptation is possible by increasing the number of transformation using the regression class. The problem then occurred when the number of regression class increased while the adaptation data is static. The transformation matrices are difficult to estimate well enough when the amount of adaptation data is reduced too much due to a fine regression class division.

To overcome this problem the use of **Vector Field Smoothing (VFS)** incorporated with MLLR is a one technique. VFS is used to deal with the problem of retraining with insufficient training data. The transformation matrices produced by MLLR is then be used to calculate the transform vector of VFS continued by smoothing process.

2.1 Maximum Likelihood Linear Regression

MLLR uses a set of regression based transform to tune the HMM mean parameter to new speaker. The aim of MLL is to estimate an appropriate transformation for the mean vectors of each mixture component so that original system is tuned to the new speaker. For mixture component s with mean μ_s, the adapted mean estimate $\hat{\mu}_s$ is given by the following equation.

$$\hat{\mu}_s = W_s \bullet \xi_s$$

where W_s is an $n \times (n+1)$ transformation matrix and ξ_s is the extended mean vector,

$$\xi_s = [\omega, \mu_s, \cdots, \mu_{sn}]'$$

where the value of ω indicated whether an offset term is to be included: $\omega = 1$ for an offset, $\omega = 0$ for no offset. The transformation matrix is determined with a re-estimation algorithm based upon the principle of maximum likelihood estimation. In this way, the re-estimated transformation matrix is the one that maximizes the probability of having generated the observed adaptation data using the model.

2.2 Vector Field Smoothing

The vector field smoothing technique assumes that the correspondence between feature vectors from different speaker is viewed as a smooth vector field. Based on this assumption, the correspondence obtained from adaptation data is considered to be an incomplete set of observation from the continuous vector filed, containing observation errors. To achieve both better correspondence and reduction errors, both interpolation and smoothing are introduce into adaptation process.

VFS has three steps, as follows:

- **Concatenation training:** In this step, the mean vector of the Gaussian distribution is trained by concatenation training.
- **Interpolation:** In this step, the untrained mean vector is transferred to the new speaker's voice space by using an interpolated transfer vector.
- **Smoothing of transfer vector:** In this step, each transfer vector is modified in accordance with the other transfer vector.

2.3 MLLR-VFS

The technique of MLLR-VFS can be separately performed in three steps. The first step is an extension of the MLLR to multiple regression matrixes. The second step is calculating the transfer vector of VFS using the regression matrix produced by MLLR. The third step is the smoothing of transfer vector as VFS usual manner.

- Extension to multiple regression class

If R states $\{s_1, s_2, \cdots, s_R\}$ are shared in a given regression class, then the regression matrix $\hat{W}_s$ can be written:

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{sr}(t)\sum_{s_r}^{-1} o_t \xi'_{sr} = \sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{sr}(t)\sum_{sr_r}^{-1}\hat{W}_s \xi_{sr} \xi'_{sr}$$

- Calculation of transfer vector

The transfer vector $\Delta\vec{\mu}_i$ is calculated from the difference between the mean vector of the initial continuous density HMM and the initial continuous density HMM multiplied by the regression matrix.

$$\Delta\vec{\mu}_i = \vec{\mu}_i - \hat{W}_s \vec{\mu}_i$$

- Smoothing of transfer vector

In this step, each transfer vector is modified in accordance with the other transfer vector as an usual VFS manner.

We conduct these steps to develop the Indonesia speech recognition system with favorable result. Using the speech data provided by ATR-Japan, we obtain a promising result with accuracy rate around 90%. The signal processing model takes 12 kHz sampled data and transform it into 39-dimensional MFCC vectors every 10 ms (see Table 1, A: speaker independent, B: speaker dependent, Data is number of words for adaptation). This experiment also used Left-to-Right HMM model with single Gaussian Mixture.

Table 1. Result of Indonesian ASR

A	Data	MAP	VFS	MLLR	MLLR -VFS	B
79.7	10	85.19	81.96	85.34	86.51	92.7
	20	86.50	84.28	87.91	89.22	
	40	87.75	86.50	89.80	89.34	
	80	90.23	90.26	90.57	91.39	
	100	90.11	90.76	90.29	91.97	

Based on this result, we are now in collaboration with Telkom RDC to develop speech data to enhance the accuracy. We will also improve the speed of the system.

3 Machine Translation for Indonesian Language

A large number of Indonesian people, statistically, have problem in using and comprehending any information presented in other cross-border languages. The language barrier problem is compounded by the problem of the explosion of digital information whose majority uses English language via either Internet or any digital printed form which may overwhelms potential users and pose a threat of inequality of access of information due to the language barrier (digital divide) especially for the common Indonesian people. This is one of the motivations for us to propose a collaborative project to develop speech to Asian speech translation system, between BPPT-Indonesia, ATR-Japan, ETRI-Korea, NECTEC-Thailand, CCNOIDA-India, NTU-Taiwan and CAS-China.

In line with the research objectives, our most recent experiment is focusing on developing Indonesian statistical language model - based on the integration of stochastic and symbolic approaches - to be used for analysis stage in the machine translation engine. For creating the stochastic language model, it is worthwhile to utilize annotated data when it is available and use supervised learning mechanism to estimate model’s parameter. In this case, an annotated corpus is created for multiple genres of documents. Of course, the costs of annotation are prohibitively labor intensive, and the resulting corpora sometimes are susceptible to a particular genre.

Due to this limitation of annotated corpora, it is necessary that we use unsupervised and weakly supervised learning techniques, which do not require large, annotated data sets. Unsupervised learning utilizes raw, un-annotated corpora to discover underlying language structure such as lexical and contextual relationships. This gives rise to emergent patterns and principles found in symbolic systems. In this system, the language model is trained using weakly supervised learning on small annotated corpus to seed unsupervised learning using much larger, un-annotated corpora. Unsupervised and weakly supervised methods have been used successfully in several areas of NLP, including acquiring verb sub-categorization frames, part-of-speech tagging, word-sense disambiguation and prepositional phrase attachment.

The Internet has proven to be a huge stimulus for statistical MT, with hundreds of millions of pages of text being used as corpus resources. Over the last few years, there has been an increasing awareness of the importance of corpus resources in MT research. As researchers begin to consider the implications of developing their systems beyond the level of proof-of-concept research prototypes with very restricted coverage, considerable attention is being paid to the role that existing bilingual and monolingual corpus and lexical resources can play. Such collections are a rich repository of information about actual language usage.

In developing monolingual corpus, we checked existing Indonesian news articles available on web (Purwarianti, 2007). We found that there are three candidates for the article collection. But in the article downloading, we were only able to retrieve one article collection, sourced from Tempointerakif. We downloaded about 56,471 articles which are noisy with many incorrect characters and some of them are English. We cleaned the articles semi-automatically by deleting articles with certain words as sub title. We joined our downloaded articles with the available Kompas corpus (Tala, 2003) at http://ilps.science.uva.nl/Resources/BI/ and resulted 71,109 articles.

In Indonesia, many research groups have been developing a large-scale annotated corpus to further the NLP and Speech research in trainable system. It should be clear that in statistical approach, there is no role whatsoever for the explicit encoding of linguistic information, and thus the knowledge acquisition problem is solved. On the other hand, the general applicability of the method might be doubted; it is heavily dependent on the availability of good quality of data in very large proportions, something that is currently lacking for Indonesian languages.

In order to experiment the feasibility of statistical MT for Indonesian, we build a prototype Indonesian-English MT. For that purpose, we need parallel corpus of Indonesian-English sentences, and there are none publicly available. Therefore, we have develop a collection of training and test sentences collected from a number of information sources mainly from Indonesia national news agency ANTARA, totaling 250.000 parallel sentences. We then use SRILM to build the n-gram language model and translation model, subsequently use PHARAOH (Koehn 2006) as a beam search decoder.

4 Discussion and Future Work

We are working forward to improve the quality of speech recognition and MT. Our collaboration with Telkom RDC and ATR-Japan will provide us with new speakers' data (40 speakers, 1000 words) which is expected to improve the accuracy of ASR to a better 90% level.

In other speech processing work, University of Indonesia (UI) and Bandung Institute of Technology (ITB) are also developing ASR and speech synthesis (SS) which will be integrated in the final speech translation system.

We are also building a new corpus in broadcasting news, to train the translation system, so as to enable automatic "tagline" in bilingual TV program. The experts in translation have two differing approaches toward the translation concept: universalism and monadic. We understood there is a possibility of "un-translation" which is "translation fails – or un-translability occurs when it is impossible to build functionally relevant features of the situation into contextual meaning of target language (TL) text. Broadly speaking, the cases where this happens fail into two categories. Those where the difficulty is linguistic, and those where it is cultural.

We examine further the translability concept by taking into account that most Asian language share very similar "culture" but different in language structure. We can not enforce the system and structure to target language without "knowing" the language itself. In this case, a rule-based system should be used as a preprocessing to enable the structure of source language to approximate the structure of target language. For example, in translating Indonesian-English, we need a rule-based system to transform the DM-MD rule. This rule approximates the order of noun and adjective phrase of Indonesian according to English noun or adjective phrase. For example:

```
      MD                    DM
sebuah rumah besar -> a big house
  (a)   (house) (big)
gunung  biru itu -> the blue mountain
(mountain) (blue) (the)
```

In our future work, by implementing several symbolic modules as pre-processor, it is expected that statistical MT will perform better in translating by having a "similar" language structure.

5 Conclusion

An updated report on speech to speech translation system is given together with a brief overview of some of the issues and techniques in speech recognition and statistical machine translation (SMT), which are being actively researched today in Indonesia.

It is particularly important for Indonesian language to have research on speech-to-speech translation systems, which is an ideal solution to the field of language technology. Such work is clearly important but difficult because it certainly will bring up many interesting differences of emphasis, for example in speech-to-speech work, there is an emphasis on speed, and on dealing with sentence fragments, since we would like to be able to translate each utterance as it is spoken, without waiting for the end. This gives importance to bottom up methods of language analysis, and severe restrictions on the input in terms of the type of text.

References

Ayu Purwarianti, Masatoshi Tsuchiya and Seiichi Nakagawa. 2007. Developing a Question Answering System for Limited Resource Language - Indonesian QA, submitted to *Journal of Language Resources and Evaluation*.

C.H. Lee, J.L. Gauvain. 1993. “Speaker Adaptation Based on MAP Estimation of HMM Parameters”, Proc.ICASSP, Minneapolis, USA, pp.II-558-561.

C.J. Leggetter, P.C. Woodland. 1995. “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models”, *Computer Speech and Language*, 9(2):171-185.

F.Z. Tala. 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, M.Sc. Thesis, University of Amsterdam.

H. Riza. 1999. The Indonesia National Corpus and Information Extraction Project (INC-IX), Technical Report, BPP Teknologi, Jakarta, Indonesia.

H. Riza. 2001. BIAS-II: Bahasa Indonesia Analyser System Using Stochastic-Symbolic Techniques, International Conference on Multimedia Annotation (MMA), Tokyo, Japan.

Heidi Christensen. 1996. “Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression”, Project Report, Aalborg University, Denmark.

J.C. Junqua, J.P Haton. 1996. Robustness in Automatic Speech Recognition – Fundamental and Application, Kluwer Academic Publiser, Netherland.

Kazumi Ohkura, Masahide Sugiyama, Shigeki Sagayama. 1992. “Speaker Adaptation Based on Transfer Vertor Field Smoothing with Continuous Mixture Density HMMS, Proc of ICSLP 92, pp. 369-372.

M.J.F. Gales. 1997. ”Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition”, TR 291, Tech. Report, Cambridge University Engineering Department.

Oskar Riandi. 2001. “A Study on the Combination of Maximum Likelihood Linear Regression and Vector Field Smoothing for Speaker adaptation”, M.Sc Thesis, Japan Advanced Institute of Science and Technology (JAIST), Japan.

S.Young, G. Evermann, M.J.F. Gales, T. Hain, Dan Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P.C. Woodland. 2005. “The HTK Book (for HTK Version 3.3)”, Revised for HTK Version 3.3 April 2005, Cambridge University Engineering Department

Philipp Koehn. 2006. Statistical Machine Translation: the Basic, the Novel and the Speculative, SMT Tutorial, University of Edinburgh.

Sakriani Sakti, Konstantin Markov, Satoshi Nakamura. 2005. “Rapid Development of initial Indonesian Phoneme-Based Speech Recognition Using The Cross-Language Approach”, Proceeding of O-COCOSDA, Jakarta.