PAN Localization Project

RESEARCH REPORT

PHASE 1.2

Initial Design Report on Statistical Machine Translation Framework

Agency for the Assessment and Application of Technology
Badan Pengkajian dan Penerapan Teknologi (BPPT)
December 2008

Initial Design Report on Statistical Machine Translation Framework

1. Introduction

Machine translation (MT) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. At its basic level, MT performs simple substitution of words in one natural language for words in another. Using corpus techniques, more complex translations may be attempted, allowing for better handling of differences in linguistic typology, phrase recognition, and translation of idioms, as well as the isolation of anomalies.

Current machine translation software often allows for customization by domain or profession (such as news) - improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows then that machine translation of government and legal documents more readily produces usable output than conversation or less standardized text.

Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators, and in some cases can even produce output that can be used "as is". However, current systems are unable to produce output of the same quality as a human translator, particularly where the text to be translated uses casual language.

In this PANL project, BPPT is determined to design and develop a statistical machine translation system of English to Indonesia, which can translate text from different domains. This research is necessary in order to exercise the applicability of customization and experiment with scarce language resource (parallel corpus).

2. Statistical Machine Translation in Theory

Statistical machine translation tries to generate translations using statistical methods based on bilingual text (parallel) corpora. The term parallel corpora are typically used in linguistic circles to refer to texts that are translations of each other. And the term comparable corpora refer to texts in two languages that are similar in content, but are not translations. In order to exploit a parallel text, some kind of text alignment, which identifies equivalent text segments (approximately sentences), is a prerequisite for analysis.

To yield good translation we needed a good bilingual Indonesian-English text corpus. There are many sources of parallel texts in order to create bilingual corpus, however they require to be selected properly. These written articles have many sentence structure and coming from different domain of texts, such as national news, sport, economy, etc. The designer of an SMT system constructs a general model of the translation relation, and then lets the system acquire specific rules automatically from bilingual and monolingual text corpora. These rules are usually coarse and probabilistic, for example, in a certain corpus 'bat' translates as 'kelelawar' (bat, animal) and as 'pemukul' (hit with a stick, as in batter) in the other sentences. The most established SMT system is based on word-for-word substitution (Berger et al), although some experimental SMT systems employ syntactic processing (Wu, Alshawi et al, Su et al).

An advantage of the SMT approach is that designers can improve translation accuracy by modestly changing the underlying model rather than overhauling large handcrafted Resources.

The benefits of statistical machine translation over traditional paradigms that are most often cited are the following:

- Better use of resources

There is a great deal of natural language in machine-readable format. Generally, SMT systems are not tailored to any specific pair of languages. Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.

- More natural translations

The ideas behind statistical machine translation come out of information theory. Essentially, the document is translated on the probability $p(e|f)$ that a string e in native

language (for example, English) is the translation of a string *f* in foreign language (such as Bahasa Indonesia). Generally, these probabilities are estimated using techniques of parameter estimation.

The Bayes Theorem is applied to *p(e|f),* the probability that the foreign string produces the native string to get,

$$p(e|f) \propto p(f|e)p(e)$$

Where the translation model *p(f|e)* is the probability that the foreign string is the translation of the native string, and the language model *p(e)* is the probability of seeing that native string. This basically turns the translation direction and splits the problem into two sub-problems. Mathematically speaking, finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability:

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

For a rigorous implementation of this, one would have to perform an exhaustive search by going through all strings *e** in the native language. Performing the search efficiently is the work of a machine translation decoder that uses the foreign string, heuristics and other methods to limit the search space and at the same time keeping acceptable quality. This trade-off between quality and time usage can also be found in speech recognition.

As the translation systems are not able to store all native strings and their translations, a document is typically translated sentence by sentence, but even this is not enough. Language models are typically approximated by smoothed n-gram models, and similar approaches have been applied to translation models, but there is additional complexity due to different sentence lengths and word orders in the languages.

The statistical translation models were initially word-based in the Models 1-5 from IBM Hidden Markov Model (S.Vogel, 1996) and Model 6 from (Koehn 2003), but significant advances were made with the introduction of phrase based models. Recent work has incorporated syntax or quasi-syntactic structures. In the following, we describe these two approaches in brief. .

Word-based translation

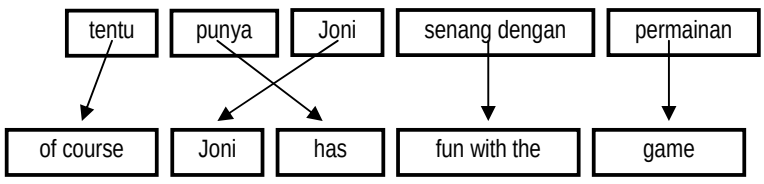
In word-based translation, translated elements are words. Typically, the number of words in translated sentences is different due to compound words, morphology and idioms. The ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces. Simple word-based translation is not able to translate language pairs with fertility rates different from one. To make word-based translation systems manage, for instance, high fertility rates, and the system could be able to map a single word to multiple words, but not vice versa. For instance, if we are translating from French to English, each word in English could produce zero or more French words. But there's no way to group two English words producing a single French word.

An example of a word-based translation system is the freely available GIZA++ package, which includes the training program for IBM models and HMM model and Model 6 (Koehn et.al 2007).

The word-based translation is not widely used today comparing to phrase-based systems, whereas, most phrase based system are still using GIZA++ to align the corpus. The alignments are then used to extract phrase or induce syntactical rules, and the word alignment problem is still actively discussed in the community. There are now several distributed implementations of GIZA++ available online.

Phrase-based translation

In phrase-based translation, the restrictions produced by word-based translation have been tried to reduce by translating sequences of words to sequences of words, where the lengths can differ. The sequences of words are called, for instance, blocks or phrases, but typically are not linguistic phrases but phrases found using statistical methods from the corpus. Restricting the phrases to linguistic phrases has been shown to decrease translation quality.



3. Development Framework

The development objective on our work in PANL Project is to build a statistical machine translation toolkit and make it available to researchers in our PANL community. This SMT toolkit would include corpus preparation software, bilingual text training software, and run time decoding software for performing actual translation. All these software components are based on Open Source Software (OSS).

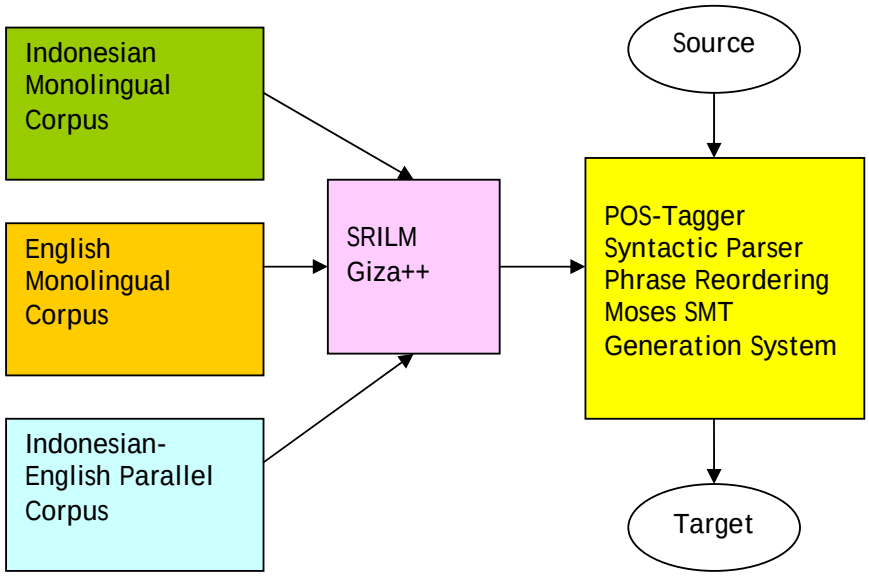


Figure 1. English-Indonesian SMT Development

Based on our initial analysis of Indonesia-English translation model, it is our goal to use a statistical method or combination of statistical and symbolic method in our MT experiment. The proposed system will consist of the following symbolic modules:

Morphological analyzer and Syntactic parser

The derivational morphology of Bahasa Indonesia, particularly that of verbal morphology, is quite complex. Therefore, to process Indonesian words require careful analysis. Existing Indonesian morphological literature (Alwi, et. al. 2003) is surveyed and compared with the empirical evidence as found in the corpus. The morphological analysis

will be built using both n-gram or HMM and symbolic rule-based system. Completion of the morphological analysis will enable the development of a hybrid phrase parser. This parser will decide noun phrases, verb phrases, adjective phrases, etc. in a sentence. Furthermore, a robust syntactic parser can be build using output of the shallow parser to decide syntactical structure of a sentence, i.e., which part of a sentence is the subject, predicate, object.

Phrase Reordering System

This module will perform transformation of phrase structure from Indonesian to English, especially to enable correct translation of noun phrase (NP) and adjective phrase (AdjP). The word order in Indonesia NP <N,Adj> is reordered to match the word order of English NP <AdjP,N> This module will prepare the input sentence or files prior of translation process, to enable word reorder and aligned to achieve better translation of NP and AdjP.

Generation system

This module will produce target sentences (Indonesian or English) based on an intermediate representation created by the statistical MT.

4. Issues on Statistical Machine Translation

The performance of a statistical machine translation system depends on the *size* of the available task-specific *bilingual training corpus*. On the other hand, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and, for some language pairs, is not even possible. Besides, small corpora have certain advantages like low memory and time requirements for the training of a translation system, the possibility of manual corrections and even manual creation.

Therefore, investigation of statistical machine translation with small amounts of bilingual training data is receiving more and more attention. The goal of statistical machine translation is to translate a source language sequence into a target language sequence by maximizing the posterior probability of the target sequence given the source sequence. In state-of-the-art translation systems, this posterior probability usually is modeled as a combination of several different models, such as: phrase-based models for both translation directions, lexicon models for translation directions, target language model, phrase and word penalties, etc. Probabilities that describe correspondences between the

words in the source language and the words in the target language are learned from a bilingual parallel text corpus and language model probabilities are learned from a monolingual text in the target language. It is suggested that the larger the availability of training corpus, the better the performance of a translation system. Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort and for some language pairs is not even possible. In addition, small corpora have certain advantages: the possibility of manual creation of the corpus, possible manual corrections of automatically collected corpus, low memory and time requirements for the training of a translation system, etc. Therefore, the strategies for exploiting limited amounts of bilingual data are receiving more and more attention.

In the last five years various publications have dealt with the issue of sparse bilingual corpora. (Al-Onaizan et al., 2000) report an experiment of Tetun- English translation with a small parallel corpus, although this work was not focused on the statistical approach. The translation experiment has been done by different groups including one using statistical machine translation. They found that the human mind is very capable of deriving dependencies such as morphology, cognates, proper names, etc. and that this capability is the crucial reason for the better results produced by humans compared to corpus based machine translation. If a program sees a particular word or phrase one thousand times during the training, it is more likely to learn a correct translation than if it sees it ten times, or never. Because of this, statistical translation techniques are less likely to work well when only a small amount of data is given.

Based on the above, we will also perform baseline evaluations. These evaluations would consist of both objective measures on statistical model perplexity and subjective measures on human judgments of quality, as well as attempts to correlate the two. We would also produce learning curves that show how system performance changes when we vary the amount of bilingual training text.

In our planned experiments, we would like to address the challenges with statistical machine translation which have been addressed by many researchers. In addition to domain and size of corpus, some of the problems that statistical machine translation has to deal with include:

- Compound words
- Idioms
- Morphology
- Different word orders

  Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are located. In Speech Recognition, the speech signal and the corresponding textual representation can be mapped to each other in blocks in order. This is not always the case with the same text in two languages. For SMT, the translation model is only able to translate small sequences of words and word order has to be taken into account somehow. Typical solution has been re-ordering models, where a distribution of location changes for each item of translation is approximated from aligned bi-text. Different location changes can be ranked with the help of the language model and the best can be selected.
- Syntax
- Out of vocabulary (OOV) words

  SMT systems store different word forms as separate symbols without any relation to each other and word forms or phrases that were not in the training data cannot be translated. Main reasons for out of vocabulary words are the limitation of training data, domain changes and morphology.

Following this initial report, we will present the Final Design SMT Framework during the next phase.

5. Reference

1. Brown et al, 1993 “The Mathematics of Statisti-cal Machine Translation: Parameter Estimation”, P. Brown, S. Della Pietra, V. Della Pietra, R. Mer-cer. Computational Linguistics, 19(2).
2. Knight, 1997 “Automating Knowledge Acquisi-tion for Machine Translation”, K. Knight, AI Magazine, 18(4).
3. Al-Onaizan et al, 2000 “Translating with Scarce Resources”, Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, K. Yamada, AAAI-2000.

4. Pang et al, 2003 “Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences,” B. Pang, K. Knight, and D. Marcu. NAACL-HLT-2003.
5. Papineni et al, 2002, “Corpus-based Comprehen-sive and Disagnostic MT Evaluation: Initial Ara-bic, Chinese, French, and Spanish Results”, K. Papineni, S. Roukos, T. Ward, J. Henderson, F. Reeder. NAACL-HLT-2002.
6. Wikipedia Creative Commons, http://en.wikipedia.org/wiki/, retrieved August 08
7. Commons Wikipedia, http://commons.wikimedia.org/wiki/Main_Page, Sept 2008
8. Garside, R., Leech, G., and McEnery, T. Corpus annotation: linguistic information from computer text corpora, Harlow: Addison-Wesley Longman, 1997.
9. S. Vogel, H. Ney and C. Tillmann. 1996. HMM-based Word Alignment in Statistical Translation. In COLING ’96: The 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen, Denmark
10. P. Koehn, F.J. Och, and D. Marcu (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.
11. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. ACL 2007, Demonstration Session, Prague, Czech Republic
12. W. J. Hutchins and H. Somers. (1992). An Introduction to Machine Translation, 18.3:322. ISBN 0-12-36280-X