Producing a Cross-Language Dictionary using Statistical Machine Translation: A First Experiment with English and Indonesian

Joseph Cathcart and Robert Dale

March 2, 2001

Abstract

Well-developed Statistical Machine Translation techniques now exist for carrying out word alignment in parallel corpora. A by-product of training data for this task is a set of translation probabilities for the correspondences between target and source tokens. In the literature, these techniques have relied on the use of bitexts of significant size; however, for many languages no such corpora exist. In this paper, we report on an experiment where a relatively small corpus was used to generate word correspondences, which were then evaluated against a hand-constructed bilingual lexicon. We present the suprisingly good results achieved, and discuss some possible improvements to the technique.

1 Introduction

In this paper, we report on a first experiment in using Statistical Machine Translation (SMT) results in the context of English–Indonesian translation. Our long-term goal is the automatic development of resources for machine translation using parallel text corpora in conjunction with SMT techniques. In our initial experiments, we have been exploring how a relatively small corpus can be used to generate Indonesian translates of English words.

The paper is structured as follows. First, Section 2 gives a brief overview of the SMT techniques used. Section 3 then describes the corpus we used for the current experiments, and Section 4 outlines the results of the experiment described here. Finally, Section 5 draws some conclusions and outlines some future directions that follow from the work presented.

2 Statistical Machine Translation

The statistical translation model pioneered by IBM (Brown et al., 1990) casts translation as a channel process. Given an Indonesian string $\mathbf{i}$ to be translated into an English string $\mathbf{e}$, the model considers $\mathbf{i}$ as the target of a communication channel, and its translation $\mathbf{e}$ as the source of the channel. Viewing every English string as a possible source for each Indonesian target string, the machine translation task is to recover the source from the target by assigning a probability $\Pr(\mathbf{e} \mid \mathbf{i})$ to each pair of sentences $(\mathbf{e}, \mathbf{i})$, and seeking the particular $\mathbf{e}$ which

maximizes $\Pr(\mathbf{e} \mid \mathbf{i})$. According to Bayes' rule

$$\Pr(\mathbf{e} \mid \mathbf{i}) = \frac{\Pr(\mathbf{e}) \Pr(\mathbf{i} \mid \mathbf{e})}{\Pr(\mathbf{i})}.$$

Clearly then

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{i} \mid \mathbf{e}),$$

where $\hat{\mathbf{e}}$ denotes the English string which maximizes the conditional probability $\Pr(\mathbf{e} \mid \mathbf{i})$. We shall focus here on the conditional probability $\Pr(\mathbf{i} \mid \mathbf{e})$, the so-called translation model. For this we adopt IBM Model 1 (Brown et al., 1993). Setting $\mathbf{e} = e_1 e_2 \ldots e_l$ and $\mathbf{i} = i_1 i_2 \ldots i_m$, we obtain

$$\Pr(\mathbf{i} \mid \mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{s=1}^{m} \sum_{r=0}^{l} t(i_s \mid e_r).$$

Seeking to maximize $\Pr(\mathbf{i} \mid \mathbf{e})$ subject to the constraints that for each e,

$$\sum_{i} t(i \mid e) = 1$$

allows us to employ the EM algorithm (Dempster et al. 1977) to estimate translation probabilities $t(i \mid e)$ from an Indonesian–English bitext. It is these probabilities which are used to locate our potential Indonesian translates.

3 The Corpus

Most SMT work that uses parallel corpora has focussed on well-known and widely available corpora, with the Canadian Hansard being the most popular. This corpus consists of Canadian parliamentary proceedings in both French and English, collected over a period of several years. Unfortunately, finding parallel corpora of this size for other language pairs is somewhat difficult. In exploring how these techniques might be used in the context of English–Indonesian translation, the largest bitext sources we could find were the Koran and the King James Bible. Comparative sizes are shown in Table 1; as can be seen, Hansard is more than 50 times the size of the King James Bible corpus. An interesting question, then, is how useful such a comparatively small corpus can be. The work reported in the present paper uses the King James Bible, aligned at the verse level, although we have also used the techniques on the smaller Koran corpus.

The King James version of the Bible presents several problems as a raw text, both in English and Indonesian, because of non-standard variations in capitalisation and punctuation usage; given these idiosyncracies, we decided to remove all punctuation and to convert the text entirely to lower-case. As an exception, hyphenation was retained in the Indonesian version since to remove it would risk conflating orthographically distinct words. Due to processing

*As used by Brown et al., 1993.
†Aligned at sura level.
‡Aligned at verse level and not including James 3.
§After hapax were replaced by a symbol indicating the presence of an unknown word.

	Canadian Hansard*	Koran†	Bible‡
Alignments	1,778,620	6,236	31,084
Source tokens	unreported	133,193	666,895
Target tokens	unreported	133,036	657,714
Source types	42,005§	6,140	13,450
Target types	58,016§	7853	20,932

Table 1: Comparison of corpora

$e =$ *beast*		$e =$ *darkness*	
i	$Pr(i\|e)$	i	$Pr(i\|e)$
binatang	0.73432	kegelapan	0.54073
hewan	0.16532	gelap	0.36118
ternak	0.03337	kekelaman	0.05771
makhluk	0.03033	gulita	0.01502
gantinya	0.01188	kelam	0.01152
meterai	0.00455	terpendam	0.00342
nasar	0.00366	kegelapanku	0.00239
binatangpun	0.00351	pelitaku	0.00239
teji	0.00200	gelaplah	0.00195
hewannya	0.00172	kegelapanlah	0.00132

Table 2: Probability tables

constraints, verse pairs for which either the English or Indonesian verse exceeded 40 tokens were removed from the corpus. The resulting adjusted corpus contains $19,244$ Indonesian word types and $13,294$ English word types.¶

4 Results: An English–Indonesian Dictionary

The EGYPT Statistical Machine Translation Toolkit was used to implement the EM algorithm. Translation probabilities below a threshold of 1.0×10^{-7} were removed at each iteration. After 10 iterations $751,584$ translation probabilities remained. The 10 $highest$ probabilities were collated for each of the 500 most frequent English word types from within the adjusted corpus. As an example, Table 2 presents the results for the words *beast* and *darkness* respectively; in each case the translate with the significantly highest probability is indeed the correct translate.

To test the performance of this technique, we independently constructed a benchmark dictionary against which to compare the results. Relying on the authors' knowledge and some (paper) dictionary lookup, each of the 500 most frequent words in the English corpus were assigned their Indonesian translates.‖ These Indonesian translates were then compared with

¶ We suspect that the difference in type counts is due principally to the agglutinative nature of Indonesian, but this remains to be verified.

¶ Of course, many words have more than one translate; accordingly, several Indonesian translates were typically assigned for each English word to allow for word sense ambiguity.

Rank	Freq.	Relative Freq. (%)	Cumulative Freq. (%)
1	383	76.6	76.6
2	20	4.0	80.6
3	20	4.0	84.6
4	12	2.4	87.0
5	5	1.0	88.0
6	—	—	88.0
7	5	1.0	89.0
8	—	—	89.0
9	—	—	89.0
10	3	0.6	89.6
> 10	52	10.4	100.0

Table 3: Ranking frequencies

the results of the experiment, and the English word assigned a rank from 1 to 10 according to the highest match between the automated translate and the benchmark translates. As we carried out no prior morphological processing, we extended this comparison to the root or base word of the Indonesian translates. Thus, in the above example, had the benchmark dictionary included *binatangku* and *kebinatangan* but neither *binatang* nor *binatangpun*, the English type *beast* would be assigned a rank of one.

The overall results are shown in Table 3; this demonstrates that the correct translation is ranked first by the method 76.6% of the time.

5 Conclusions and Future Work

As demonstrated in Table 3, using the EM algorithm even on a relatively small corpus provides suprisingly good results. In the automatic construction of a dictionary resource, an n-best processing regime that could take account of other information could conceivably access the translates ranked second, third, and so on. However, it is not immediately apparent how such a post-processing step would be built, and so we are interested to see if the base probabilities can be increased by other means.

In the short term, we intend to carry out four further experiments:

- We intend to compare the results presented here with the same process applied to the Koran corpus, and to a blend of the two corpora; this will give some indication of the stability of the numbers with respect to corpus size and content.
- We intend to explore the use of morphological processing to reduce word instances to their morphological roots, so as to reduce the number of distinct word types and increase the number of instances of each. This will require the development of a morphological component for Indonesian.
- We intend to explore the performance of the techniques on words of other frequencies, since the most frequent words may have other characteristics that impact on the process; our intention here would be to test on words chosen from a range of different

frequency bands, rather than just selecting the most frequent words.

- Although not presented here, we have noted that the algorithm is particularly effective at determining correct translations of proper nouns; we intend to explore how this characteristic can be exploited in the translation of technical terms.

The resuls of this first experiment are promising: they suggest that relatively small corpora may be used to automatically derive translation lexica. The next step is to see just how reliable the technique can be made, and what factors impact on this reliability.

References

P F Brown, J Cocke, S A Della Pietra, V J Della Pietra, F Jelinek, J D Lafferty, R L Mercer, and P S Roosin, P.S. (1990) A statistical approach to machine translation. *Computational Linguistics*, 16(2):29–85.

P F Brown, J Cocke, S A Della Pietra, V J Della Pietra and R L Mercer (1993) The mathematics of statistical machine translation: Parameter estimation *Computational Linguistics*, 19(2):263–311.

J M Echols and H Shadily (1989) *Kamus Indonesia—Inggriss*, PUB.

J M Echols and Shadily (1989) *Kamus Inggris—Indonesia*, PUB.