Malay-English Bitext Mapping and Alignment Using SIMR/GSA Algorithms

Mosleh H. Al-Adhaileh and **Tang Enya Kong**
Computer Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, MALAYSIA
{mosleh, enyakong}@cs.usm.my

I. Dan Melamed
Department of Computer Science
Courant Institute
New York University
New York, NY, USA
10003

Abstract

Parallel texts or Bitexts – where the same content is available in several languages, due to document translation, are becoming plentiful and available, both in private data warehouses and on publicly accessible sites on the WWW. Bitexts can be used as knowledge resources in many domains such as for machine translation, bilingual lexicography, word sense disambiguation, or multilingual information retrieval. Text alignment can also be a useful practical tool for assisting translators. The first step in extracting information from a bitext is to describe the correspondence between the two halves of the bitext (bitext mapping and alignment). In this paper we described our experiment in porting SIMR and GSA algorithms for bitext mapping and alignment to Malay-English Bitexts. This will help to compile these bitexts into a useful format for research and development on Malay language. The output results and evaluation show that SIMR/GSA algorithms perform on the Malay-English Bitexts with high accuracy, as they perform on the other variety of language pairs (i.e. French-English, Korean-English, Chinese-English, etc.). The Bitexts, the maps and the alignments are available for research use from UTMK[1].

Keywords: SIMR, GSA, Malay-English Bitext Mapping, Malay-English Text Alignment.

1. Introduction

A text in one language and its translation constitute a bitext [3]. Between different language pairs, bitexts are becoming plentiful and available, both in private data warehouses and on publicly accessible sites on the WWW. The first step in extracting information from a bitext is to describe the correspondence between the two halves of the bitext (**bitext mapping and alignment**). For a given bitext, bitext mapping is to find the corresponding points (i.e. words, text units, or segments boundaries) between its two halves. In contrast to a correspondence relation, "an alignment is a segmentation of the two texts such that the n^{th} segment of one text is the translation of the n^{th} segment of the other" [11:68], (i.e. which segments, in one language correspond to which segments in the other language).

By doing this, bitexts will be compiled into a useful sources of knowledge for many machine translation strategies, since they depend on aligned sentences or other aligned text segments. Text alignment can be used not only for the task of machine translation, but it is also a first step in using multilingual corpora as knowledge resources in other domains such as for bilingual lexicography, word sense disambiguation, or multilingual information retrieval. Text alignment can also be a useful practical tool for assisting translators. As we plan to do here in this experiment, we will invest in the Malay-English bitexts to collect information, and compile these bitexts into a form of bitext maps and alignments, which can be used in NLP research.

This paper is organized as following: Section 2 gives a brief overview about the idea behind SIMR/GSA algorithms, and an orientation to the main sources of information and details about them, Section 3 is the main contribution of this paper, it describes the steps of porting SIMR and GSA to Malay-English Bitexts, results and evaluation for both SIMR and

[1] See **http://utmk-ultra.cs.usm.my/** , the URL for Unit Terjemahan Melalui Komputer (UTMK) – USM.

GSA is detailed in section 4, and Section 5 ends the paper with a brief conclusion.

2. SIMR and GSA algorithms

A bitext can form the axes of a rectangular **bitext space**, as in Figure 1. The height and width of the rectangle correspond to the lengths of the two texts, in character. The lower left corner and the upper right corner of the rectangle represent the text beginning (origin) and end (terminus) respectively. The other corresponding character positions between the two texts, the **true points of correspondence** (TPCs), other than the origin and the terminus, can be plotted as points in the bitext space. TPCs exist both at the coordinates of matching text units and at the coordinates of matching text units boundaries. If a token at position x on the x-axis and a token at position y on the y-axis are translation of each other, then the coordinate **(x, y)** in the bitext space is a TPC. A **bitext map** is the real-valued function obtained by interpolating successive points in a given bitext space. A complete set of TPCs for a particular bitext is the **true bitext map** (TBM). The purpose of the Smooth Injective Map Recognizer (SIMR) algorithm is to produce bitext maps that are the best possible approximations of each bitext TBM as illustrated in Figure 2. For more details on SIMR see [7], and [8].

The Geometric Segment Alignment (GSA) algorithm reduces sets of correspondence points in SIMR's output to segment alignments. Given a set of correspondence points, supplemented with segment boundary information, expresses segment correspondence; segment boundaries form a grid over the bitext space. Figure 3 illustrates how segment boundaries form a grid over the bitext space. Each cell in the grid represents the intersection of two segments, one from each half of the bitext. A point of correspondence inside cell (X,y) indicates that some token in segment X corresponds with some token in segment y; i.e., segments X and y correspond. For example, in Figure 3, the segment **e** corresponds with segments **G** and **H**, also the segment f corresponds with the segment **H**, so the segment <**G, H**> should be aligned with segment <**e, f**>. In Figure 3 the aligned blocks are outlined with solid lines. The GSA algorithm can be applied equally well to sentences, paragraphs, lists of items, or any other text units for which boundary information is available. For more details on GSA see [8].

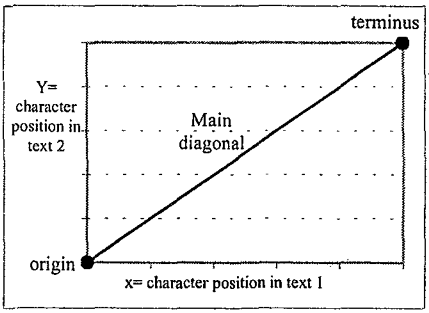
terminus
Y= character position in text 2
Main diagonal
origin
x= character position in text 1

Figure 1: A bitext space

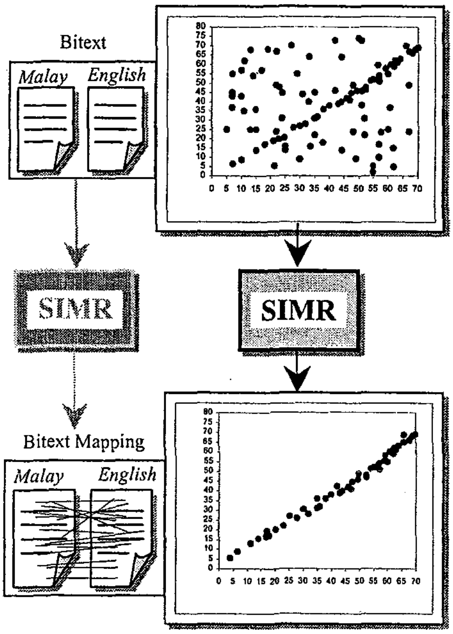

Figure 2: Bitext mapping using SIMR

3. Porting SIMR/GSA to Malay-English Bitext

The report on porting SIMR to a new language pairs [6] describes the steps that should be done in order to adapt the SIMR into a new language pairs. These steps are: *i*- choosing a match predicate, *ii*- axis generator and *iii*- SIMR's parameters re-optimization. In this Experiment, we describe the process on porting the SIMR on Malay-English bitext.

We describe the way we collected our data, the matching predicate heuristic and the construction of SIMR's parameter for Malay-English bitext. Results and evaluation of the SIMR/GSA on the tested data are given in Section 4. Figure 4 illustrates the process of porting the SIMR/GSA to Malay-English bitext.

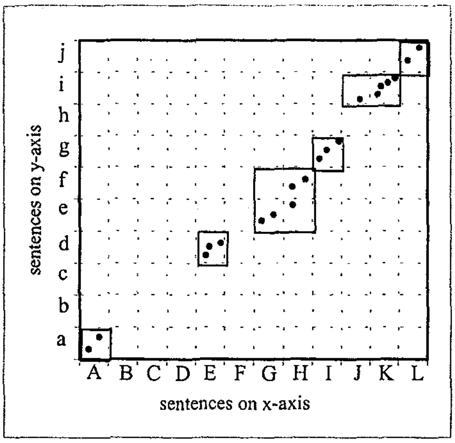

Figure 3: Segment boundaries form a grid over the bitext space. Each cell in the grid represents the intersection of two segments, one from each half of the bitext. A point of correspondence inside cell (X, y) indicates that some token in segment X corresponds with some token in segment y; i.e., segments X and y correspond. For example, segment E corresponds with segment d. The aligned blocks are outlined with solid lines.

3.1 Data Collection

The linguistic department at Unit Terjemahan Melalui Komputer (UTMK) collected some Malay-English bitexts in different genres (novels, user guides, literature books, etc.). The problem is: these bitexts were processed using a scanner to scan the text first, and then a tool was used to change them into texts. By doing this, the bitexts need to be edited, and most of them are not edited yet, or in process of editing. Fortunately, we find the two books "The 7 Habits of Highly Effective People" [1] and "Semantics" [10] are almost edited. The "7 Habits" book consisted of 101,790 words in the English version and 107,161 words in the Malay version. It is divided into 13 chapters. From the "Semantics" book we collected 8 chapters, all together about 50,170 words in the English version, and 51,802 words in the Malay version. Both the "7 Habits" and the "Semantics" books are hand-aligned at the level of sentences. Also we find that The "Microsoft Word For Windows version 2.0: User's Guide" [9] can be used in our test, but it is raw data, so we took only the first 20 pages and we hand-aligned them for the purpose of testing SIMR/GSA. These bitexts were suitable for our testing at the beginning stage; the data is specified in Table 2.

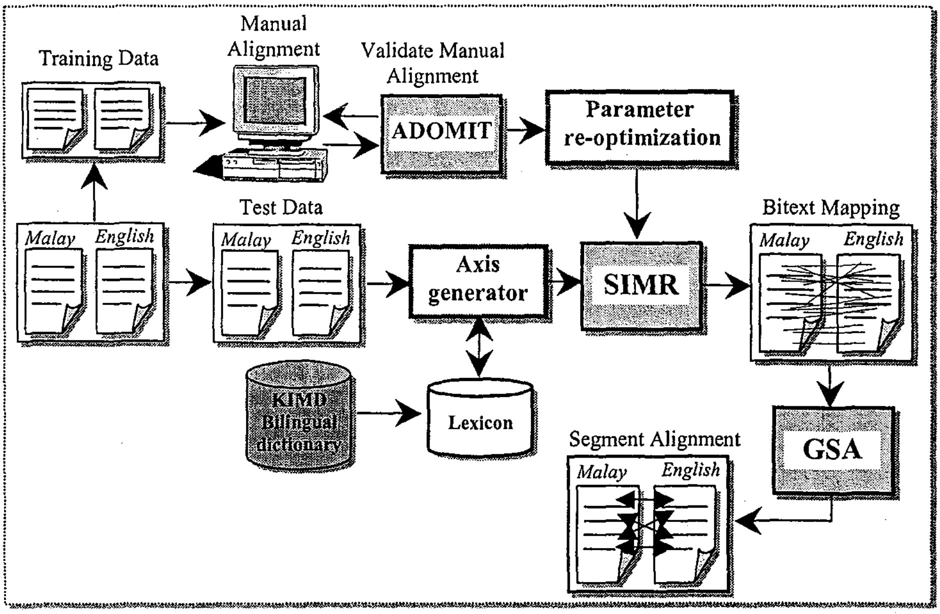

Figure 4: Malay-English bitext mapping and alignment using SIMR/GSA

3.2 Matching Predicate

A matching predicate is a heuristic used to decide whether two given tokens in the bitext might be mutual translations. The matching predicate is one of the SIMR's parameters, and should be defined before using the SIMR. The Malay and English languages share the same alphabet, and orthographic cognates (i.e. two tokens in the bitext with the same meaning and similar spellings) exist between them, but the correspondence points generated based on cognates are not enough signal for SIMR to achieve an accurate mapping. In this case another matching predicate is used to strengthen the signal by generating more correspondence points, which is a translation lexicon. The translation lexicon is a list of word pairs that are believed to be mutual translations. This lexicon can be extracted from a machine-readable bilingual dictionary. In our case, we have an English-Malay machine-readable bilingual dictionary "Kamus Inggeris Melayu Dewan (KIMD)" [2]. It consisted of 42,000 entries. We find that the KIMD, that we have, has more than 20000 entries, which are single-word to single-word, also we can use the phrasal equivalences, which give us a lexicon of 38,343 word pairs. We think this is probably enough to get good results. The lexicon is lemmatized in the same way of lemmatizing the collected bitexts, as we will explain in Section 3.3. The matching predicates were fine-tuned with a stop-list words for both English and Malay languages. The translation lexicon pairs with the cognates and the punctuation marks between the two sides of the bitext should give very good signals to the SIMR algorithm.

3.3 Axis Generator

The next step in porting SIMR/GSA to a new language pair is axis generating. As mentioned previously, SIMR takes two bitext space axis as parameters. The mapping from tokens to axis position is performed by using a language-specific axis generator program, one for each half of the bitext. Tokens should correspond to the smallest semantic unit of the language in hand. Usually, such units are words. Like English tokens, Malay tokens are separated by space. So a simple approximation is to assign a position to every space-delimited token, taking in consideration punctuation marks and numbers. The position of a token (in character) is the position of its median character. For example, the axis for the sentence **"tujuh tabiat gambaran seluruh ."** looks like this:

```
0 <EOS>
3 tujuh
9.5 tabiat
17.5 gambaran
26 seluruh
31 .
31.5 <EOS>
```

Although SIMR doesn't care about the segment boundaries, they are central to the process that converts SIMR's bitext maps to segment alignments, i.e. GSA process. Therefore, the input axes must carry segment boundary information in the form of markers, such as <EOS> in the above example. These markers must have a text position just like any other text token.

Data Lemmatization

Before using the bitext, we need to lemmatize both the English and the Malay versions. For lemmatizing the English we use Brill's POS tagger and the XTAG lexicon, it contains 90000 roots yield over 317000 inflected forms [4]. The English text is tokenized, and then it is tagged with a simplified version of the Penn Tree Bank tag set using Brill's POS tagger. A program is used to compute the stem for each word using the POS tag and the XTAG lexicon.

For the Malay side, a root construction program is used to construct the root for every word in the Malay text (if possible), it based on rules and a lexicon for the root words in Malay. The lexicon contains the roots for 10000 popular words in the Malay language.

3.4 Parameter Optimization

The report on porting SIMR to new language pairs [6] recommends re-optimization of the SIMR's parameters. To optimize the parameters, a training data should be prepared and mapped (i.e. Normally, the creation of the bitext maps is done manually). The training bitext maps should at least consist of 500 points. For this purpose, since the "7 Habits" book is manually aligned at the level of sentences, we took Chapter 3, 7 and 11 as a training data. This data consists of 1245 segments.

Validate the Manual Alignment Before starting parameters re-optimization; we make sure that this data is not noisy by checking the manual alignment using an omission detection tool ADOMIT [5], because we can't do parameter re-optimization on a noisy data. ADOMIT is an algorithm for Automatic Detection of OMIssions in Translations. It relies solely on geometric analysis of bitext maps and uses no linguistic information. The basic method of ADOMIT is: Given a noise-free bitext map, firstly, a bitext space is constructed by placing the original text on the x-axis, and the translation on the y-axis, secondly, the known point of correspondence are plotted in the bitext space. Each adjacent pair of points delimits a segment of the bitext map. Any segment whose slope is unusually low is a likely omission, as illustrated in Figure 5. For more details about ADOMIT see[5].

ADOMIT is not used only for omission detection, but also to detect the errors in hand-aligned bitexts. It is not surprising to find errors in hand-aligned bitexts. So before re-optimizing the SIMR's parameters on the hand-aligned training data, we validate it using ADOMIT.

After validating the hand-aligned training data, and fixing the detected error, re-optimization is done using simulated annealing [12]. The construct set of the optimized parameters for the Malay-English testing bitexts is shown in Table 1.

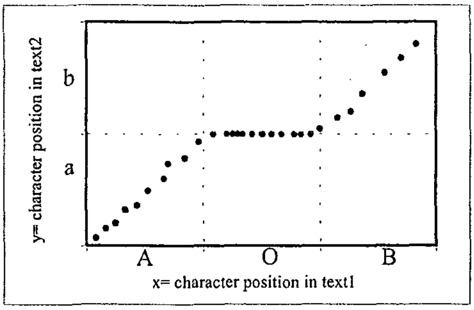

Figure 5: An omission in bitext space. Regions A and B correspond to regions a and b, respectively. Region O has no corresponding region on the vertical axis.

Parameter	Value
Chain size	7
Max. point ambiguity	6
Max. angle deviation	0.14
Max. linear regression error	5
Min. Cognate length ratio	0.80

Table 1: The optimized set of SIMR's parameters based on the Malay-English training bitext.

4. Results and Evaluation

The evaluation metric, for SIMR results, is the root mean square distance (RMS), in character, between each TPC and the interpolated bitext map produced by SIMR, where the distance was measured perpendicular to the main diagonal. For each test data, the number of test TPCs were derived from segment hand-alignments by pairing the character position at the ends of aligned pairs. In Table 2, for each test data, the number of test TPCs is: ***#of segments – 1***. For the GSA results, the error rate *(%)* is the percentage of the wrong segment alignments (comparing to the hand-aligned test data) from the total segments in the test data.

Based on the results we noticed:

-Most of the GSA errors happened in case of omissions in one side, or there should be a combination of segments in one side. (i.e. 1-omitted, 1-n alignments).

- The results of GSA on the "Semantics" book and the "MS Word User Guide" are very good. Most of the test data are aligned correctly as compared to the hand-aligned data. These surprisingly good results may be a result of: ***i-*** A more literal translation between the two halves of the bitext. The stronger signal can improve SIMR's/GSA's performance. ***ii-*** The "7 Habits" data might still be noisy. Table 2 reports SIMR's and GSA errors on the test data. The output results and evaluation have shown that SIMR/GSA algorithms can map/align Malay-English bitexts with high accuracy as they performed on the other variety of language pairs and text genres. These results encourage us, as a future work, to think of extending the text alignment to word alignment aiming at the identification of correspondence between linguistic units below the sentence level within a bitext.

5. Conclusion

In this paper we described our experiment in porting SIMR and GSA algorithms to Malay-English bitexts. The output results and evaluation have shown that SIMR/GSA algorithms can map/align Malay-English bitexts with high accuracy as they performed on the other variety of language pairs and text genres. SIMR is robust in the face of texts that are different in genre and structure (i.e. missing segments or omission, inversion, and crossing dependencies). These features make SIMR/GSA algorithms one of the most widely applicable bitext mapping and alignment published to date. Also, a word for researchers who need to do Malay-English text alignment, SIMR/GSA can be ported to Malay-English bitexts with minimal efforts. This experiment will form the base for researchers to work on bitexts where Malay language is involved, and more importantly to consider Malay-English bitext, which are available, to test and evaluate their algorithms.

Bitext	# of Segments	# of SIMR points	RMS error in chartacter for SIMR	Error rate (%) for GSA
The "7 Habits" book				
Chapter 1	27	184	12.3	0.0
Chapter 2	620	3094	60.7	2.7
Chapter 4	581	2391	8.0	6.2
Chapter 5	970	4826	9.0	2.6
Chapter 6	714	3489	8.3	2.4
Chapter 8	670	3403	8.1	1.8
Chapter 9	752	2670	8.6	6.9
Chapter 10	501	2241	9.2	4.6
Chapter 12	212	1007	9.3	1.9
Chapter 13	119	840	11.6	4.2
The "Semantics" book				
Chapter 1	136	969	10.7	0.0
Chapter 2	194	1856	10.6	0.0
Chapter 3	346	2768	11.0	0.0
Chapter 4	194	1856	10.7	0.0
Chapter 5	217	1438	10.2	0.0
Chapter 6	237	1897	10.0	0.0
Chapter 7	318	2218	12.3	0.0
Chapter 8	132	1039	18.6	1.5
Microsoft word User Guide				
The first 20 pages	400	3233	7.6	0.5

Table 2: RMS errors in character s for SIMR results, and the Error rate % for the GSA results on the tested bitexts.

References

[1] Covey S. R. 1990. *The 7 habits of highly effective people*, Published by Simon & Schuster, 1st edt.

[2] Dewan Bahasa dan Pustaka. 1992, *Kamus Inggeris Melayu Dewan*, Kuala Lumpur, 1st edt.

[3] Harris B. 1988. *Bi-text, a new concept in translation theory*. Language Monthly, 54: pp 8-10.

[4] Karp D., Schabes Y., Zaidel M. and Egedi D. 1992. *A Freely Available Wide Coverage Morphological Analyzer for English*, In Proc. of COLING '92. pp 950- 954, Nantes, France.

[5] Melamed I.D. 1996. *Automatic detection of omissions in translations*. In Proc. of the 16th International Conference on Computational Linguistic, pp 764-769, Copenhagen, Denmark.

[6] Melamed I.D. 1996. *Porting SIMR to new language pairs*. Institute for research in Cognitive Science Technical Report 96-26, University of Pennsylvania, PA.

[7] Melamed I.D. 1997. *A portable algorithm for mapping bitext correspondence*. In Proc. of the 35th Annual Meeting ACL, pp 305-312, Madrid, Spain.

[8] Melamed I.D. 1999. *Bitext maps and Alignment via Pattern Recognition*. Computational Linguistic, 25, No 1, pp 107-130.

[9] Microsoft Corporation. 1991. *Microsoft Word for Windows Version 2.0: User's Guide*, Document Number OB-22376-1091, USA.

[10] Palmer F.R. 1981. *Semantics*, Cambridge University Press.

[11] Simard M., Foster G. and Isabelle P. 1992. *Using Cognates to align sentences in bilingual corpora*. In Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, pp 67-81, Montreal, Canada.

[12] Vidal R.V. editor. 1993. *Applied Simulated Annealing*. Springer-Verlag, Heidelberg, Germany.