Clause Alignment for Hong Kong Legal Texts: A Lexical-based Approach*

Chunyu Kit Jonathan J. Webster King Kui Sin Haihua Pan Heng Li

Department of Chinese, Translation and Linguistics
City University of Hong Kong
{ctckit, ctjjw, ctsinkk, cthpan, ctliheng }@cityu.edu.hk

Abstract

In this paper we report on our recent work in clause alignment for English-Chinese legal texts using available lexical resources including a bilingual legal glossary and a bilingual dictionary, for the purpose of acquiring examples at various linguistic levels for example-based machine translation. We present our formulation of an appropriate measure for the similarity of a candidate pair of clauses with respect to matched lexical items and the corresponding implementation of an effective algorithm for clause alignment based on this similarity measure. Experimental results show that the similarity measure and the lexical-based clause alignment algorithm, though very simple, are very effective, with a performance of 94.6% alignment accuracy. It confirms our intuition that lexical information gives a reliable indication of correct alignment. The significance of this lexical-based approach lies in both its simplicity and effectiveness.

Keywords Clause alignment, text alignment, example-based machine translation

1 Introduction

Text alignment at various linguistic levels is a critical task in current MT technology. It serves different purposes in different researchers' work. For example, it is utilised to construct statistical translation models [3, 5] and to acquire examples for example-based machine translation (EBMT). EBMT is one of the most prominent modern MT paradigms whose basic ideas originally evolved from Nagao [19]. However, EBMT shares a fundamental philosophy with translation memory (TM), which was brought to light in Kay's "proper place" paper [15] on the real potential of MT technology. Namely, the use of existing translations to facilitate translating new texts, be it by machine or human translators, enhances both translation productivity and quality.

Basically, approaches to text alignment can be classified into two types: *statistical-* or *probabilistic-based* and *lexical-based*. The statistical-based approaches rely on non-lexical information (such as sentence length, sentence position, co-occurrence frequency, sentence length ratio in two languages, etc.) to achieve alignment tasks, as illustrated in previous research [13, 8, 10, 16]. We may refer to such approaches as *resource-poor* approaches. The attraction of these approaches arises from the sharp contrast between their poor resources and their rich outcomes.

The sentence alignment approach proposed in [16] produces a bilingual lexicon in addition to sentence alignment outcomes, because of its adoption of a strategy of integrating sentence alignment and word alignment for iterative refinement of both. It is also known as "lexical sentence alignment" in the text alignment literature. However, this should not be confused with "lexical-based" sentence alignment

*The work presented here is part of the CERG project "EBMT for HK Legal Texts" funded by HK UGC under the grant #9040482, with Jonathan J. Webster as the principal investigator and Chunyu Kit, Caesar S. Lun, Haihua Pan, King Kuai Sin and Vincent Wong as co-investigators. The authors wish to thank all team members who have contributed to the research work that enables this paper, in particular, Yan Wu who worked for the project as research associate. Correspondence concerning this work should be addressed to Dr. Jonathan J. Webster, CTL, CityU of HK, Tat Chee Ave., Kowloon, HK.

mentioned above, which refers to the *resource-rich* alignment techniques that heavily rely on existing lexical resources such as large-scale bilingual dictionaries and glossaries. As more and more bilingual lexical resources become available, it is worth investing more research effort to investigate the effectiveness of lexical-based approaches.

In this paper we report on our current work in clause alignment for English-Chinese bitext of Hong Kong legislation using available lexical resources including a glossary of bilingual legal terms and a large-scale bilingual dictionary which is part of the example acquisition phase of an ongoing EBMT project. We are interested in acquiring examples at various linguistic levels, including the clause, phrase, and word. By the term "example" we refer to a pair of texts in two languages that translate each other. One reason why we conduct text alignment at the so-called clause level is that the legal bitext in use contains many long sentences that are broken into segments by various types of numbering and punctuation. We refer to such naturally-occurring segments in a sentence as clauses, although many of them are in fact long phrases and even single words instead of well-formed clauses in a strict linguistic sense. Clause alignment serves the purpose of acquiring examples at the clause level, including complete sentences (most of which are very long in legal texts), clauses, and such fragments. Then we can move on to finer-grained alignment at the phrase and word levels.

The paper is organised as follows: In Section 2, we introduce (1) the English-Chinese parallel corpus of Hong Kong laws, namely, the BLIS Corpus, on which we have based our research, and (2) the resources, namely, a glossary of English-Chinese legal terminology and a large-scale bilingual dictionary, which is exploited to facilitate clause alignment. In Section 3, we present a simple but effective method for clause identification in the corpus with the aid of punctuation marks. In Section 4, we formulate a number of similarity measures for paired candidate clause pairs. A number of factors can be utilised, including matched glossary and dictionary items, clause length and clause position. After careful testing, we select the most effective similarity measure for our clause alignment task. In Section 5, we formulate the alignment algorithm based on the similarity measure chosen. The algorithm is designed to implement a simple strategy. It selects a minimal optimal set of scores in the similarity matrix that covers all clauses in both languages. In Section 6, we present the alignment experiments and evaluation results. Our clause alignment approach achieves a performance of 94.60% alignment accuracy; correspondingly, 88.64% of words and 88.49% of characters from the input corpus are in the properly aligned clause pairs. Finally, Section 7 concludes the paper, and briefly discusses our future work on example acquisition for EBMT.

2 Corpus and Resources

The bilingual corpus we use for our research is extracted from the Bilingual Laws Information System (BLIS). It contains the complete text collection of the statutory laws of Hong Kong, originally encoded in the Lotus Notes format by the technical unit of the Hong Kong S.A.R. Department of Justice[1]. We have converted the entire corpus into pure texts.

2.1 Corpus structure and size

The statutory laws of Hong Kong are divided into three main categories, namely,

- public ordinances (i.e., laws which concern the general public),
- private ordinances (i.e., laws which concern individual bodies, whether statutory or otherwise), and

[1] The general public may access the BLIS database at `http://www.justice.gov.hk` via the Internet. Given the enormous size of the database, corpus building would have been impossible without the full support of the Hong Kong S.A.R. Department of Justice. In particular, a CD ROM of the BLIS system was provided to us for the purpose of research.

- miscellaneous ordinances (i.e., laws which do not belong to either of the preceding categories).

To date, there are 564 public ordinances, 166 private ordinances and 12 miscellaneous ordinances, adding up to a total of 742 ordinances. The entire bilingual corpus of BLIS legal texts contains approximately 9 million English words and 5 million Chinese characters.

Hong Kong ordinances are arranged by chapters, each of which is identified by an *assigned number* and a *short title*, e.g., `Cap 1, The Interpretation and General Clauses Ordinance (第一章 釋義及通則條例)`. Chapters 1 - 564 are public ordinances and Chapters 1001 - 1166 private ordinances. The numbers assigned to these two categories of ordinance are official numbers, i.e., they are the numbers appearing in the Loose-Leaf Edition of the Laws of Hong Kong, which is the most authoritative version of the Laws of Hong Kong. The miscellaneous ordinances are also assigned numbers (from 2401, e.g., Cap 2401 The National Flag and National Emblem Ordinance) in the BLIS, but these are unofficial numbers, i.e., such numbers do not appear in the Loose-Leaf Edition of the Laws of Hong Kong, but are assigned to the miscellaneous ordinances purely for the sake of compiling the BLIS database.

The content of an ordinance, exclusive of its *long title*, is divided and identified according to a very rigid numbering system.

1. Parts, identified by uppercase Roman numerals, I, II, III, IV, ... etc.
2. Sections, identified by Arabic numerals, 1, 2, 3, 4, ... etc.
3. Subsections, identified by Arabic numerals in brackets (1), (2), (3), (4), ... etc.
4. Paragraphs, identified by lowercase letters in brackets (a), (b), (c), (d), ... etc.
5. Subparagraphs, identified by lowercase Roman numerals (i), (ii), (iii), (iv), ... etc.

Parts and sections are also given headings, which are, however, not an operative part of an ordinance. Many ordinances contain schedules and/or forms identifiable by Arabic numerals.

The Chinese version of an ordinance follows the same numbering system as its English counterpart. Chinese numerals are not used. Accordingly, the Chinese texts of Hong Kong laws are perfectly aligned with the English texts in terms of chapters (章), parts (部), sections (條), subsections (款), paragraphs (段) and subparagraphs (節). It is this feature that makes the bilingual texts of Hong Kong laws particularly suitable for our project, because a well-aligned text of this size is seldom readily available. Excerpts from the corpus are illustrated in Table 1.

2.2 Glossary and dictionary

The resources used in our clause alignment include a bilingual glossary of HK law and a large-scale English-Chinese bilingual dictionary. Their size is presented in Table 2, together with a number of example entries of the glossary and dictionary. An *entry* (or *item*, interchangeably) is an English-Chinese pair of words or terms that are translation candidates for each other. It is different from the conventional entry in a bilingual dictionary that consists of one lemma in the source language and one or more translations in the target language, although we opted for this conventional structure for our dictionary implementation. That is, a conventional entry may accommodate more than one word pair. Such a pair is referred to as a dictionary or glossary item in our terminology throughout the paper.

The dictionary we used actually subsumes a Chinese-English and an English-Chinese dictionary, for the sake of simplicity of implementation via using existing facilities. The English-Chinese dictionary has 109.3K entries (i.e., lemmas), and the Chinese-English one has 119.6K entries, residing together in a memory space of 5.4M.

The glossary consists of legal terminology compiled by the Department of Justice of the HK S.A.R. government. However, the glossary and the dictionary do overlap. Many word pairs, e.g., `solicitor/律師` and `abuse/濫用` appear as an entry in both the glossary and the dictionary in use. They are considered glossary items, instead of dictionary items, in our alignment algorithm.

CAP. 71 Control of Exemption Clauses	第 71 章 管制免責條款條例
PART II CONTROL OF EXEMPTION CLAUSES ORDINANCE Avoidance of liability for negligence, breach of contract, etc.	第 II 部 管制免責條款 逃避因疏忽、違約等而引致的法律責任
. . .	. . . . .
8. Liability arising in contract (1) This section applies as between contracting parties where one of them deals as consumer or on the other's written standard terms of business. (2) As against that party, the other cannot by reference to any contract term – (a) when himself in breach of contract, exclude or restrict any liability of his in respect of the breach; or (b) claim to be entitled – (i) to render a contractual performance substantially different from that which was reasonably expected of him; or (ii) in respect of the whole or any part of his contractual obligation, to render no performance at all, except in so far as (in any of the cases mentioned above in this subsection) the contract term satisfies the requirement of reasonableness.	8. 合約因致的法律責任 (1) 如立約一方以消費者身分交易，或按另一方的書面標準業務條款交易，則本條適用於處理立約各方之間的問題。 (2) 對上述的立約一方，另一方不能藉合約條款而 – (a) 在自己違反合約時，卸除或局限與違約有關的法律責任; 或 (b) 聲稱有權 – (i) 在履行合約時，所履行的與理當期望他會履行的有頗大的分別; 或 (ii) 完全不履行其依約應承擔的全部或部分法律義務，但在該合約條款 (於本款上述的任何情況下) 符合合理標準的範圍內，則不在此限。
(Enacted 1989) [cf. 1977 c. 50 s. 3 U.K.]	(1989年制定) [比照 1977 c. 50 s. 3 U.K.]

Table 1: An illustration of the structure of HK ordinances

Resource	HK Legal Glossary		Dictionary	
Size (K)	23.2		109.3 / 119.6	
Example	`abuse`	濫用	`sufficient`	足夠
	`registration procedure`	登記程序	`other`	其他
	`abandon ··· a petition`	放棄 ··· 呈請	`solicitor`	律師
	`absolute majority of the votes`	絕對多數票	`order`	命令
			`and`	及

Table 2: The size of the glossary and dictionary, and example entries

The apportioned part of any such rent, annuity, dividend, or other payment shall be payable or recoverable,	上述租金、年金、股息或其他付款的分攤部分,
in the case of a continuing rent, annuity, or other such payment,	如屬連續租金、年金或其他該等付款,
when the entire portion of which such apportioned part forms part becomes due and payable,	則於該分攤部分所屬的整份租金、年金或其他該等付款到期須支付時而非之前即須予支付或可予追討; ·····
and not before; ···	
All rents, annuities, dividends, and other periodical payments in the nature of income shall,	所有租金、年金、股息及其他屬收入性質的定期付款,
like interest on money lent,	
be considered as accruing from day to day,	須如同貸款利息而視作逐日累算,
and shall be apportionable in respect of time accordingly.	並據此按時間的長短而可予分攤。

Figure 1: Illustration of punctuation correspondence in BLIS corpus

3 Clause Identification

The first step towards clause alignment is proper identification of clauses in the bilingual corpus. From a linguistic point of view, a clause is a sub-structure within a sentence. A compound sentence may take more than one clause. One of the motivations for conducting text alignment at the clause level, instead of the sentence level, in our research is that, in the legal domain, a sentence can be notoriously long, containing many clauses and even sub-clauses within clauses. Often a long paragraph comprises only one sentence, consisting of a number of clauses and sub-clauses. Thus, it is reasonable to regard text alignment at the sentence level for the legal texts in use for our research as not very suitable for acquiring useful examples for the specific purpose of EBMT.

The most important information in the bilingual corpus that we can utilise to identify clauses is the punctuation marks in the two languages and their relatively reliable correspondence in the corpus. The punctuation pairs ";/;" and "./。" are reliable for *sentence* identification. They always mark the end of a sentence in both languages, respectively. The pair ",/," is useful for *clause* identification, but may be problematic, because a English comma "," does not necessarily mark the end of a clause, especially when it appears in a coordination structure. Therefore, it does not necessarily match a Chinese comma "，", which is a reliable clause ending in most cases in the BLIS corpus. For example, in the fragment of English-Chinese parallel text shown in Figure 1 some underlined commas "," in English match their counterparts "、", a slight-pause mark, in Chinese coordination structure, but the ones preceding "or" (或) and "and" (及) do not have an overt counterpart in Chinese.

We observe that some prepositional and adverbial phrases are identified as clauses. Though they are not authentic clauses in the linguistic sense, nevertheless, such results are quite useful in our research work, because we aim at acquiring not only examples at the clause level, as we do at this stage of our project, but also examples at sub-clause, phrase and word levels in later stages. It is thus entirely acceptable if some structures finer-grained than clauses, such as phrases or sub-clause chunks, are properly aligned at the clause alignment stage, as we have to align them up sooner or later. For example, we allow the last English comma in the first sentence in Figure 1 to be recognised as the clause end, and so single out a few words as a "clause" for our clause alignment. This clause will find its counterpart in the last Chinese clause of the sentence through a proper alignment based on the lexical pairs "not/非" and "before/之前".

4 Similarity Measures

Clause alignment is aimed at identifying pairs of clauses that are translations of each other in two languages. Any two paired clauses, also referred to as a true pair, are assumed to be more similar to each other than any other arbitrary clause pair, with respect to their meaning and interpretation, because translation is intended to preserve semantic and pragmatic equivalence in both source and target language. This section gives the definition of similarity measures used in our lexical-based clause alignment.

4.1 Four factors

The similarity of a clause pair gives an indication of how strongly the two clauses, each in a language, are correlated to each other. This indication suggests how likely the two clauses should be identified as the translation of each other.

In a statistical-based approach to text alignment, only non-lexical information such as co-occurrence statistics and clause length is utilised to compute a probabilistic score for a candidate pair of clauses. In a lexical-based approach, we can make use of the available lexical information to derive a more reliable semantic similarity based on lexical correspondence. For example, we can compute a similarity score for a clause pair in terms of the number of matched word pairs in the two clauses in question. Note that the word pairs are defined in a bilingual dictionary.

For the purpose of defining an effective similarity measure for lexical-based clause alignment, we consider the following four factors as most important:

1. Matched dictionary items: The more such items exist in a candidate pair of clauses, the more likely they are to be a true pair.

2. Matched glossary items: The more such items exist in a candidate pair of clauses, the more likely they are to be a true pair,

3. Clause length ratio: A candidate pair of clauses are considered more similar if they have a length ratio closer to the average clause ratio of the two languages in question.

4. Clause positions: A candidate pair of clauses are considered more similar if their positions are closer[2].

However, it is necessary to differentiate between matched dictionary items and matched glossary items, because a pair of matched glossary items generally gives a much stronger indication of the similarity of a pair of candidate clauses than a pair of matched dictionary items. By the term "item" we refer to an entry in the glossary or dictionary in use.

4.2 Empirical formulation of similarity measures

Following the observed impact of the four factors on the similarity of clause pairs, we propose the following empirical formula for measuring the similarity of a candidate pair of clauses $P=<s_i, s_j>$ in a given bilingual corpus:

$$sim(s_i, s_j) = \frac{\sum_{d \in P} f(d) + W \cdot \sum_{g \in P} f(g)}{|r \cdot |s_i| - |s_j|| \cdot |i - j|} \quad (1)$$

where i and j are clause positions in the corpus, d and g are, respectively, matched dictionary and glossary items found in the clause pair in question, $f(\cdot)$ is an evaluation function for the significance of a

[2]Notice that it is convenient to think of two clauses to be closer if their positions are closer in a bilingual corpus. However, it would be more precise to measure the position closeness in terms of the distance of two paired clauses' similarity score from the left-top to right-bottom diagonal of a given similarity matrix, e.g., as the one in Table 4.

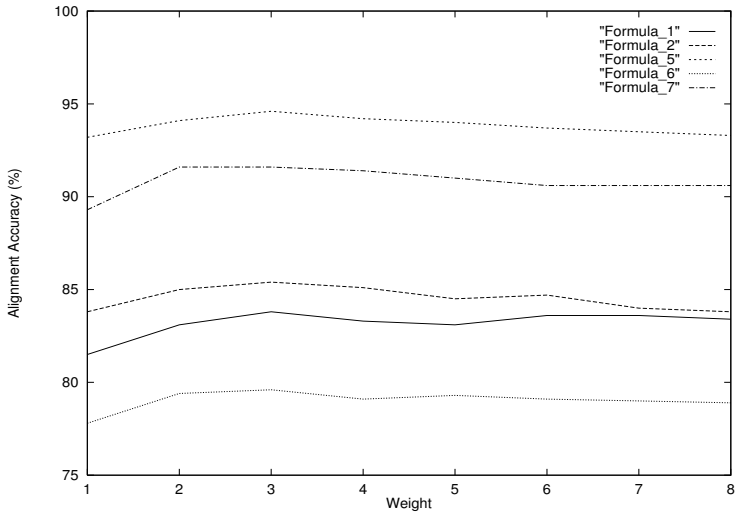

Figure 2: Experimental outcomes for determining an optimal value of W

pair of matched items, W is a weight indicating the significance of a matched glossary item in terms of a dictionary item (if other factors involved, e.g., word length, are equal), r is the average clause (or text) length ratio of the two languages in terms of the bilingual corpus in use, and $|\cdot|$ is the length of a string. The simplest evaluation function is $f(\cdot) = |\cdot|$, which means that we take the length, in the number of characters by default, of a pair of matched items as the measure for their significance contributed to the similarity of their matrix clause.

For a matched dictionary or glossary item $< u, v >$, the $f(\cdot)$ function is defined as follows, in terms of the average text length ratio for the two languages in question.

$$f(< u, v >) = r \cdot |u| + |v| \tag{2}$$

It is also necessary to differentiate between two different length functions. One, $|\cdot|_c$, represents the number of characters in a string; the other, $|\cdot|_w$, represents the number of words in a string. Choosing different length functions for (1), we have slightly different formulae for calculating similarity. However, their effectiveness turns out to be strikingly different. We have examined their effectiveness in a number of clause alignment experiments, reported in Table 3 below.

4.3 Optimal weight

The empirical coefficient W as a weight for terminology items must be experimentally determined for any particular corpus. According to our experimental results on clause alignment in the BLIS corpus with the similarity measure given in (1), the optimal value for W is set to 3 in our experiments, which indicates that a matched glossary item is, in general, 3 times as significant as a matched dictionary item. The experiments for determining this optimal coefficient are reported in Figure 2. Formulae 3 and 4 are not tested, because W is not involved.

4.4 The impact of the four factors

We have examined different combinations of the four factors in the similarity measure for clause alignment. The experimental results are presented in Table 3, which suggests the following:

1. There is only a marginally significant difference between defining the evaluation function $f(\cdot)$ in terms of characters or words. The difference is about 2 percentage points. Measuring clause length in terms of words leads to a slightly better alignment performance.

Similarity Formula	Resources Used				Alignment Accuracy	Correct Word / Char. Ratio
	Glos	Dict	Len	Pos		
(1) $\frac{\sum_{d\in P}\lvert d\rvert_c + W\cdot\sum_{g\in P}\lvert g\rvert_c}{\lvert r\cdot\lvert s_i\rvert_c-\lvert s_j\rvert_c\rvert\cdot\lvert i-j\rvert}$	√	√	√	√	83.1%	73.3% / 73.4%
(2) $\frac{\sum_{d\in P}\lvert d\rvert_w + W\cdot\sum_{g\in P}\lvert g\rvert_w}{\lvert r\cdot\lvert s_i\rvert_w-\lvert s_j\rvert_w\rvert\cdot\lvert i-j\rvert}$	√	√	√	√	85.4%	75.3% / 75.4%
(3) $\frac{\sum_{d\in P}\lvert d\rvert_w}{\lvert r\cdot\lvert s_i\rvert_w-\lvert s_j\rvert_w\rvert\cdot\lvert i-j\rvert}$		√	√	√	45.8%	47.3% / 47.9%
(4) $\frac{\sum_{g\in P}\lvert g\rvert_w}{\lvert r\cdot\lvert s_i\rvert_w-\lvert s_j\rvert_w\rvert\cdot\lvert i-j\rvert}$	√		√	√	80.1%	75.3% / 75.8%
(5) $\frac{\sum_{d\in P}\lvert d\rvert_w+W\cdot\sum_{g\in P}\lvert g\rvert_w}{\lvert i-j\rvert}$	√	√		√	94.6%	88.6% / 88.5%
(6) $\frac{\sum_{d\in P}\lvert d\rvert_w+W\cdot\sum_{g\in P}\lvert g\rvert_w}{\lvert r\cdot\lvert s_i\rvert_w-\lvert s_j\rvert_w\rvert}$	√	√	√		79.4%	70.4% / 70.5%
(7) $\sum_{d\in P}\lvert d\rvert_w + W\cdot\sum_{g\in P}\lvert g\rvert_w$	√	√			91.6%	85.1% / 84.9%

Table 3: Similarity measures and their performance

2. Interestingly, omitting both clause length and clause position only lowers the alignment accuracy by 3 percentage points from the best performance that we have achieved. This confirms our intuition that matched glossary and dictionary items give a much stronger indication of a good alignment in a lexical-based approach to clause alignment than do clause length and position.

3. Using a glossary as the sole lexical resource leads to an alignment accuracy double that of using a dictionary as the sole lexical resource, despite the fact that a dictionary is many times larger than a glossary. This result is consistent with our intuition that terminology plays a more crucial role than do normal words in text alignment, especially, for a thematic corpus such as the BLIS corpus.

4. An unexpected finding in our research is that neglecting clause length information leads to a performance about 15 percentage points better than does neglecting clause position information. It achieves the best alignment accuracy in our experiments, i.e., 94.6%. This can be explained by a feature in bilingual legal texts. Many long English sentences (and clause) translate into several short Chinese sentences (and clauses). Therefore, the clause length information brings a negative effect, instead of a positive effect.

5 Alignment Algorithms

5.1 Similarity matrix

With the similarity measure defined in (1), we can calculate the similarity score for any clause pair in a parallel bilingual corpus and derive a similarity matrix. For an English-Chinese corpus of m English clauses and n Chinese clauses, we can derive a similarity matrix that looks like the one given in Table 4, where $a_{ij} = sim(c_i, e_j)$. It is observed that most non-zero scores scatter around the diagonal line from the left-top to the right-bottom, and most scores away from this diagonal are zero or near zero.

Banding is a popular idea that many researchers follow in text alignment. It stems from the fact that in a parallel corpus a clause c_i in one language does not align with a clause e_j in the other language if a_{ij}, which situates at the *crossing point* of the two candidate clauses, is far away from the diagonal in the similarity matrix. With respect to this, our alignment algorithm can also work along the diagonal

	e_1	e_2	e_3	$\cdots$	e_j	$\cdots$	e_m
c_1	1.20	2.35	0.40	$\cdots$	0.00	$\cdots$	0.00
c_2	0.61	1.97	2.55	$\cdots$	0.00	$\cdots$	0.00
c_3	0.72	0.84	7.53	$\cdots$	0.05	$\cdots$	0.00
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$	$\cdots$	$\cdots$
c_i	0.00	0.07	$\cdots$	$\cdots$	a_{ij}	$\cdots$	a_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_n	0.00	0.00	0.03	$\cdots$	a_{nj}	$\cdots$	1.05

Table 4: An example of similarity matrix

in order to achieve better efficiency, instead of searching through all possible pairs in the matrix for an alignment. An empirical strategy for this purpose is as follows: a pair of clauses whose position difference is greater than a preset threshold is considered an impossible match. This threshold can be set as several times of the number of clauses in the largest paragraph in the BLIS corpus. An equivalent strategy is to specify that a clause does not match another clause outside the counterpart paragraph in the other language, assuming that paragraphs are well aligned.

The following two straightforward strategies for clause alignment are based on a given similarity score matrix. They are designed to test the performance of the lexical-based alignment approach. Our experiments are carried out with the second algorithm, evolving from the first.

5.2 Algorithms

Our first alignment algorithm for lexical-based clause alignment is a *best-first* algorithm:

1. Repeatedly pick the next greatest available score in the similarity matrix until all clauses in both languages are covered;

2. Turn the set of selected scores into alignments.

It looks reasonable, but unfortunately it doesn't work well. The problem is that this algorithm sometimes picks too many scores. We can take the first three clauses in the matrix in Table 4 as a tiny parallel corpus to illustrate the problem. A sound alignment algorithm should pick the scores a_{33}=7.55, a_{23}=2.55, a_{12}=2.35 and a_{11}=1.20, resulting in a reasonable alignment of two pairs:

$<[c_1], [e_1, e_2]>$ and $<[c_2, c_3], [e_3]>$.

However, the best-first algorithm also picks a_{22}=1.97, unexpectedly, resulting in an unreasonable alignment:

$<[c_1, c_2], [e_1, e_2]>$ and $<[c_2, c_3], [e_2, e_3]>$,

where both c_2 and e_2 appear in the two alignments.

We can improve the best-first algorithm by turning it into a *best-only* algorithm, with respect to the observation that a clause in one language tends to match its most similar clause in the other language, and vice versa. Accordingly, we implemented the best-only algorithm as follows:

1. Pick the greatest score in each row in the similarity matrix;

2. Pick the greatest score in each column in the similarity matrix;

3. Find the union of these two sets.

4. Turn the final score set into alignments.

This algorithm covers all clauses in both languages, and picks no more scores than necessary. For example, if we apply this algorithm to the above mentioned tiny corpus, it picks exactly the four scores

expected: a_{33}=7.55, a_{23}=2.55, a_{12}=2.35 and a_{11}=1.20, and consequently gives the following alignment result:

$<[c_1],[e_1,e_2]>$ and $<[c_2,c_3],[e_3]>$.

This algorithm is very simple but highly effective for lexical-based clause alignment, according to our experimental results. It has the capacity to output not only one-to-one alignment but also one-to-many and many-to-many (usually, two-to-two) alignments. However, according to our observation, most mistakes have to do with one-to-many clause correspondence, where clause length information has a negative effect. When we drop the clause length factor from the similarity formula, many such errors disappear. Even so, a significant number of such errors still remain in the output, leading to an error rate of 5.4%.

The time complexity of the best-first algorithm is $O(mn)$. If the idea of banding is somehow applied, it can be reduced to $O(cn)$, with c as the max band width in use, and $c \ll m$. The best-only algorithm is more efficient. Its time complexity is $O(m+n)$ only – the greatest attraction of this lexical-based approach to clause alignment.

6 Evaluation

6.1 Evaluation strategy

Since it is a laborious task to manually build a large-scale precisely clause-aligned bilingual corpus for testing, we take a more realistic approach to evaluating our clause alignment work: We conducted clause alignment experiments on a medium-sized portion of the BLIS corpus, and then randomly chose a reasonably small set of clause pairs in the output for manual examination for the purpose of evaluation. Information about the size of the entire parallel corpus in use is given in Table 5. Our program conducted clause alignment on all clauses in this data set each time with one of the similarity measures formulated above. On a PIII PC, of 1GHz and 256MB memory, it takes less than 21 minutes to finish the alignment (including similarity calculation), at an average speed of aligning 4.73K words (i.e., 2.46K Chinese and 2.27K English words) per second.

BLIS Corpus		English	Chinese	Total
Size	Clauses (K)	273	305	578
	Words (M)	2.86	3.09	5.95
	Characters (M)	17.8	10.2	28.0

Table 5: The size of the BLIS corpus used in experiments

For the purpose of estimating the alignment accuracy of each formula, however, the number of clause pairs we intended to randomly pick as an *evaluation set* is about 1000. The accuracy is defined, straightforwardly, as the proportion of correctly aligned clause pairs in this evaluation set. The evaluation process was conducted as follows.

1. Randomly pick an integer N around 1000.
2. Randomly pick N English clauses from the entire corpus of testing data.
3. Locate their Chinese counterparts in the alignment output produced by our alignment program with the similarity measure that we believed, intuitively, to be the best.
4. Manually check the correctness of all these N clause pairs:
5. If any clause in this set involves many-to-many alignment and some of its counterpart clauses are not included, add them in, resulting in an evaluation data set whose size is larger than N.

6. Alignment output by our program with any of the similarity measures is evaluated using this test set, yielding an estimation of alignment accuracy.

6.2 Evaluation measures

The alignment accuracy of each similarity measure has been reported in Table 3, together with correct word ratio and correct character ratio as a more comprehensive evaluation. The correct word (or character) ratio is defined as the proportion of words (or characters) within the correctly aligned clause pairs in the evaluation set. The intention of having these two ratios as additional evaluation measures is to remedy a certain roughness in the accuracy measure, which gives the same credit to correctly aligned clauses of different length.

In a sense, we have three kinds of clause alignment accuracy, one in terms of the number of clauses in correct alignment, and the other two in terms of the number of words and characters in correct alignment. They complement each other, making our evaluation not only more comprehensive but also more finer-grained.

The best alignment accuracy we have achieved is 94.6%, with Formula 5. Interestingly, we have also observed from Table 3 that both these two ratios are consistently significantly lower than the alignment accuracy over all similarity measures. This fact seems to suggest that shorter clauses have a better chance to find their counterparts than longer clauses within this lexical-based alignment approach. This observation goes counter to the intuition that longer clauses have a better chance of correct alignment, because they have a better chance to carry more true word pairs, resulting in, consequently, a higher similarity score. However, our experiments turned out differently, indicating that there are still issues within this simple and effective approach to clause alignment that invite more thorough review.

6.3 Alignment errors

Most mistakes in the experimental outputs involve one-to-many or many-to-many alignment, suggesting that the best-only algorithm is relatively weak in handling these types of alignment. Its weakness in the one-to-many alignment explains why clause length carries an unexpected negative effect on alignment performance, rather than the positive one that we would expect. Most such cases involve one long clause (especially, sentence) in English to be aligned with several short clauses (or, sentences) in Chinese. Errors of this type appear to be caused mostly by the enormous difference between the length of the long English sentence and that of any of its Chinese counterpart fragments. However, it seems more reasonable to attribute the cause to the unreliable similarity calculation, which is designed with a bias towards one-to-one alignment, in particular in its straightforward way of making use of clause length information. To compute the similarity for one-to-many alignment, those clauses on the "many" side need to be grouped together first and then fit into the similarity measure. Thus, a better treatment would be to also compute the similarity between a long English sentence and a candidate set of short Chinese sentences, in addition to the similarity between the long English sentence and each individual short Chinese sentence (or clauses). However, the first obstacle towards a better similarity calculation for one-to-many alignment is that there is no candidate set of short sentences during the alignment process. Thus, the problem of how to infer a set of candidate clauses for one-to-many alignment based on the one-to-one similarity calculation is an important topic for future research. Moreover, handling a similar issue for many-to-many alignment is also more demanding.

There are also many problematic cases involving not only the inadequacy of lexical resources and the imperfection of clause identification but also other problems, e.g., text chunk scrambling. For example, a chunk of text (a phrase, sub-clause or whatever) is embedded within one clause in the source text but its counterpart is attached to a different clause in the target text. Obviously, such a text scrambling phenomenon further increases the complexity of the clause alignment problem.

[e1]: The apportioned part of any such rent, annuity, dividend, or other payment shall be payable or recoverable,	[c1]: 上述租金、年金、股息或其他付款的分攤部分,
[e2]: in the case of a continuing rent, annuity, or other such payment,	[c2]: 如屬連續租金、年金或其他該等付款,
[e3]: when the entire portion of which such apportioned part forms part becomes due and payable, and not before;	[c3]: 則於該分攤部分所屬的整份租金、年金或其他該等付款到期須支付時而非之前即須支付或可予追討;
Unexpected alignment output: $<[c1],[e1]>, <[c2],[e2]>, <[c3],[e3]>$	Expected output: $<[c1,c3],[e1,e3]>, <[c2],[e2]>$

Figure 3: An example of problematic alignment output

Another example is shown in Figure 3. The underlined part in [e1] is the exact match for the underlined part in [c3]. The correct alignment should be $<[e1,e3],[c1,c3]>$. However, the alignment algorithm cannot detect the match of these two fragments in [e1] and [c3], because (1) they are not recognised as stand alone clauses and (2) there are inadequate lexical resources to match any words between them. In this case, our alignment program unexpectedly outputs the two pairs $<[e1],[c1]>$ and $<[e3],[c3]>$, and overlooks the underlined parts in [e1] and [c3] as a match. This alignment output is not totally wrong, but it leaves no chance for the scrambled parts to match each other in later alignment at a finer-grained level.

7 Conclusions and Future Work

In the previous sections we have presented our lexical-based approach to clause alignment together with experimental results on the BLIS corpus, a collection of English-Chinese bilingual legal texts of HK law. The lexical resources used in the alignment include a bilingual legal glossary and a large-scale bilingual dictionary. The similarity measure for two candidate paired clauses is defined in terms of the matched lexical items involved, their clause length, and clause positions.

The best-only alignment algorithm we implemented for the lexical-based alignment is as straightforward as collecting the maximum scores from each row and column in a similarity matrix, computed with our similarity measure on each candidate pair of clauses in a given corpus. The algorithm works in a simple way, but turns out to be a powerful and effective strategy for clause alignment. It covers not only one-to-one but also one-to-many and many-to-many alignments. With available lexical resources, this simple approach achieves an alignment accuracy of 94.6% on the BLIS corpus. This excellent performance can be accounted for by the fact that lexical information gives a more reliable indication of a proper alignment for two clauses than does non-lexical information.

Interestingly, this alignment strategy seems not to benefit from all non-lexical information. For example, it is surprising that integrating clause length information into the similarity measure lowers the alignment accuracy significantly. This indicates the need to look for a better way to harmonise the roles of lexical and non-lexical information in clause alignment than simply integrating them into a formula to compute similarity scores. Previous studies showed that non-lexical information is effective when little lexical information is available. Accordingly, we can incorporate this finding to enrich our alignment algorithm, e.g., applying non-lexical information to infer more correct alignments where lexical resources have been exhausted.

Since inadequate lexical resources continue to be one of the most critical problems in a lexical-based approach to text alignment, only a more effective strategy for combining lexical and non-lexical information can help further improve performance, in addition to making a fuller use of existing lexical

resources. The underlined parts in Figure 3 also illustrate the need to come up with a more flexible way of applying existing lexical resources. The entry for "payable" in our lexical resource matches the translation "可支付(的)", but not "應支付(的)" and "須支付(的)". The only difference among the three variant translations in Chinese is the auxiliary to the verb "支付" (pay) – the core of the term. If this core could match, the underlined parts in Figure 3 would have one pair of lexical items as a partial match, rather than none. This example gives rise to an interesting issue for our future research: whether existing lexical entries like this one can be used more fully to facilitate clause alignment, via partial match or some other flexible way of matching for inexact lexical items.

7.1 Future work

The ultimate goal of our work on text alignment is aimed at acquiring examples at various linguistic levels to facilitate EBMT, including clauses, phrases, words and many text chunks that are not necessarily linguistic constituents at any structural level. Clause alignment is the first step towards this goal.

Once clause alignment is accomplished successfully, our future work will go into two main directions. One is further improvement of the clause alignment performance. We need a better way of combining various types of available resources (especially lexical resources and non-lexical resources) and/or a valid machine learning approach to acquiring more lexical resources (e.g., the approach illustrated by Kay and Röscheisen). The other direction is text alignment in finer granularity, including word, phrase, and chunk alignment, all aimed at acquiring more examples from existing translations.

The second direction is more noteworthy because it exhibits a continuum from the resource-poor approach to the resource-rich approach. When available resources are exhausted, the resource-poor approach assumes a critical role. For example, after clause alignment, we have in each aligned clause pair a number of aligned lexical items that divide each clause into a number of non-aligned fragments. Then, the question is, how do we align these fragments, and then, words and phrases within these fragments? Since lexical resources have been exhausted, the only thing we can resort to is the lexical-poor approach, to make use of non-lexical information to carry out statistical inference about finer-grained alignment.

Chunk alignment in our future work is intended to achieve proper alignment on, and within, these fragments. It is not a trivial task. Whether only linguistic constituents should be considered valid chunks remains a problem. If only linguistic constituents are considered, sentence parsing or chunking (via partial parsing, for instance) is needed in order to identify such chunks. However, we prefer to be more open-minded: any strings that translate each other in two languages can be recognised as chunks. Whether they are linguistic constituents or not is not the most important concern, because translation is not always carried out in terms of constituency.

A bilingual dictionary is considered a set of examples at the word level. By chunk alignment, most likely to involve some unsupervised or semi-supervised learning methods after clause alignment, we are expecting to obtain aligned chunks (i.e., examples) at various linguistic levels to accomplish our ultimate goal of EBMT. Only when an adequate number of examples have been acquired can we move on to build up an EBMT translation model with the support of other MT and NLP technology, e.g., statistical MT and language modelling. It is noted that longer examples are less ambiguous but have a smaller chance to be re-used, due to their low frequency, and shorter examples (e.g. word pairs) are more frequent and thus have a higher chance to be re-used, but are more ambiguous. The EBMT model is intended to achieve a balance of these two aspects, to achieve quality translation.

When the phase of example acquisition is finished and an EBMT model is accordingly constructed, the next tasks include (1) the decomposition of input sentences into an optimal sequence of example chunks in terms of some optimality criterion (e.g., probability), and (2) the generation of output sentences from the example chunks. Re-ordering or re-organizing the example chunks in the target language side to enhance the readability of the output sentences is, of course, one of the main tasks in the generation.

References

[1] P. J. Arthern. Machine translation and computerized terminology systems: a translator's viewpoint. In B. M. Snell, editor, *Translating and the Computer: Proceedings of a Seminar*, pages 77–108, London, 1978.

[2] BLIS. Bilingual laws information system (BLIS). Info. Tech. and Resources Unit, Admin. Division, Dept. of Justice, HK Government, 1998. Detailed information available aty `http://www.justice.gov.hk/`.

[3] P. F Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lefferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85, 1990.

[4] P. F. Brown, J. C. Lai, and R. L. Mercer. Aligning sentences in parallel corpora. In *ACL-91*, pages 169–76, Berkeley, 1991.

[5] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. D. Lefferty, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

[6] R. D. Brown. Example-based machine translation in the Pangloss system. In *COLING-96*, pages 169–174, 1996.

[7] S. Chen. Aligning sentences in bilingual corpora using lexical information. In *ACL-93*, pages 9–13, Columbia, Ohio, 1993.

[8] K. Church. Char-Align: A program for aligning parallel texts at the character level. In *ACL-93*, pages 1–8, Columbia, Ohio, 1993.

[9] B. Collins and P. Cunningham. A methodology for example based machine translation. In *CSNLP-95: 4th Conf. on the Cognitive Science of NLP Proceedings*, Dublin, 1995.

[10] I. Dagan, K. W. Church, and W. A. Gale. Robust bilingual word alignment for machine aided translation. In *Proc. of the Workshop of Very Large Corpora*, pages 1–8, Columbus, OH, 1993.

[11] R. Evans and A. Kilgarriff. MRDs, standards and how to do lexical engineering. In *Proc. of 2nd Language Engineering Convention*, pages 125–32, London, 1995.

[12] P. Fung and K. McKeown. A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, 12:53–87, 1997.

[13] W. A. Gale and K. W. Church. A program for aligning sentence in bilingual corpora. In *ACL-91*, pages 177–84, Berkeley, 1991.

[14] W. J. Hutchins. The origins of the translator's workstation. *Machine Translation*, 13:287–307, 1998.

[15] M. Kay. The proper place of man and machines in language translation. *Machine Translation*, 12:3–23, 1997. First print as research report CSL-80-11, Xerox PARC, Palo Alto, CA., 1980.

[16] M. Kay and M. Röscheisen. Text translation alignment. *Computational Linguistics*, 19(1):75–102, 1993. First print as Technical Report P90-00143 in Xerox Palo Alto Research Center in 1988.

[17] C. Kit and J. J. Webster. Machine translation of idioms based on tokenization. In *Proceedings of 1st Singapore International Conference on Intelligent Systems*, Singapore, 1992.

[18] A. Melby. *The Possibility of Language: A Discussion of the Nature of Language*. John Benjamins, Amsterdam, 1995.

[19] M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*, pages 173–180. North-Holland, Amsterdam, 1984.

[20] S. Nirenburg, S. Beale, and C. Domashnev. A full-text experiment in example-based machine translation. In *Int'l Conf. on New Methods in Language Processing*, pages 78–87, Manchester, 1994.

[21] S. Nirenburg, C. Domashnev, and D. Grannes. Two approaches to matching in example-based translation. In *TMI'93*, pages 47–57, Kyoto, 1993.

[22] H. H. Pan. A machine translation system for scientific titles of English. In *Proc. of 1986 Int'l Conf. on Chinese Computing*, Singapore, 1986.

[23] H. H. Pan. Towrads understanding-based MT. In *Proc. of the Int'l Conf. on Chinese Information Processing*, Beijing, 1987.

[24] B. C. Pappegaaij, V. Sadler, and A. P. M. Witkam. *Word Export Semantics: An Interlingual Knowledge-Based Approach*. Reidel, Dordrecht, 1986.

[25] S. Sato. Example-based translation of technical terms. In *TMI'93*, pages 58–63, Kyoto, 1993.

[26] S. Sato and M. Nagao. Toward memory based translation. In *COLING-90*, pages 247–252, Helsinki, 1990.

[27] K. Schubert. Linguistic and extra-linguistic knowledge: a catalogue of language-related rules and their computational application in machine translation. *Computer and Translation*, 1:125–152, 1986.

[28] K. K. Sin and D. Roebuck. Language engineering for legal transplantation: conceptual problems in creating common law Chinese. *Language and Communication*, 16(3):235–254, 1996.

[29] F. A. Smadjia. Retrieving collocation from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[30] F. A. Smadjia and K. R. McKeown. Translating collocation for use in bilingual lexicon. In *Proc. of the ARPA Human Language Technology Workshop*, Princeton, N.J., 1994.

[31] H. L. Somers. "new paradigms" in MT: the state of play now that the dust has settled. In F. van Eynde, editor, *Machine Translation Workshop, ESSLLI-98*, pages 22–33, Saarbruecken, Germany, 1998.

[32] H. L. Somers. Example-based machine translation. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, pages 611–627. Marcel Dekker, New York, 2000.

[33] H. L. Somers. Machine translation. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, pages 329–346. Marcel Dekker, New York, 2000.

[34] E. Sumita, O. Furuse, and H. Iida. An example-based disambiguation of prepositional phrase attachment. In *TMI'93*, pages 80–90, Kyoto, 1993.

[35] E. Sumita, H. Iida, and H. Kohyama. Translating with examples: A new approach to machine translation. In *TMI'90*, pages 203–212, Texas, 1990.

[36] J. J. Webster and C. Kit. Tokenization for machine translation: What can be learned from Chinese word identification. In *Proc. of 3rd Int'l Conf. on Chinese Information Processing,*, Beijing, 1992.

[37] Y. Wilks and M. Stenvenson. The grammar of sense: Using part-of-speech tags as a first step in semantic discrimination. *Natural Language Engineering*, 4(2):135–144, 1998.

[38] D. Wu. Grammarless extraction of phrasal translation examples from parallel texts. In *TMI'95*, pages 354–372, Leuven, Belgium, 1995.

[39] D. Wu. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *IJCAI-95*, pages 1328–1335, Montreal, 1995.

[40] D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, 1997.

[41] D. Wu. Alignment. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, pages 415–458. Marcel Dekker, New York, 2000.

[42] D. Wu and X. Xia. Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 9(3-4):285–313, 1995.