http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (11): 2076-2083, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2076.2083

Extraction of Alignment Relationships in Comparable Corpora Based on Singular Value Decomposition

Francisco Oliveira, Fai Wong, Anna Ho, Yi-Ping Li and Sam Chao
Faculty of Science and Technology, University of Macau, China

Abstract: With the wide availability of information from the Internet nowadays, the process of automatically establishing alignment relationships between bilingual pairs has become a hot research topic. In this study, an alignment algorithm based on Singular Value Decomposition is proposed for handling crossing relationships between structurally dissimilar language pairs in comparable corpora. Portuguese and Chinese are used as example languages and different features are defined to model the relationship between them. Experiments results show that the proposed approach enhances the alignment accuracy compared with the baseline approach.

Key words: Text alignment, singular value decomposition, dissimilar languages

INTRODUCTION

By definition, text alignment is a procedure to match the corresponding translation from the source to the target language in a pair of bilingual documents. This process has involved the use of bilingual texts as inputs which are documents that contain domain equivalent contents with different languages.

Application area: The results generated by different alignment methods contribute a lot of valuable information in different areas. Véronis stated that the aligned text can be applied in a significantly diverse way. In general, it can be fall into four groups: Lexicography and Terminology, Machine Translation (MT), Word Sense Disambiguation (WSD) and Cross Language Information Retrieval (CLIR).

Lexicography relates to methods which analyze and describe the semantic and paradigmatic relationships within the lexicon of a language and linking the data in dictionaries according to the structures and dictionary components investigated. Modern techniques for the compilation of dictionaries involve employing bilingual corpora. If these have been sentence aligned in advance, lexicographers can improve the efficiency of discovering translations of a certain word and its corresponding collocations or other related information. Terminology is always related to the discovery of translations of complex or specialized terms. Specialist in compilation of dictionaries always find that the necessary resources do not exist or rare. Gaussier (1998) and Hiemstra (1998) designed algorithms that retrieve those translations from aligned bilingual texts.

In MT, aligned bilingual corpora is a valuable resource in Example Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) systems. EBMT systems produce the translation of the input source by searching through the translation example knowledge base that contains aligned translation pairs similar to the input source. SMT systems can use these corpora in estimating probabilities between the translation of words and the ordering of the sentences extracted from the corpora.

The task of WSD is to resolve the meaning or sense of a word unambiguously in a given context. Ng *et al.* (2003) illustrated an application of aligned corpus in solving WSD problem based on the aligned corpus. First, word alignments are identified by GIZA++ software and sense classes are defined for each noun based on defined criteria. All of these are then used to train a WSD classifier in identifying the most probable sense of words in new context.

The main objective of CLIR is to retrieve information written in a language different from the language of the user's input query. However, the query usually relies on one specific area, the shortcoming of MT and imperfection of bilingual dictionaries cannot get useful results. One of the solutions is to make use of the information extracted from the aligned parallel texts. Yang *et al.* (1998) proposed a system based on semantic latent indexing technique. The idea is to apply someone's query (source language) to a parallel text base and then the system returns the corresponding document of source language. The high ranked documents of the target language corresponding to the returned documents can

Corresponding Author: Francisco Oliveira, Faculty of Science and Technology, University of Macau, China

be the query to documents of the target language in a non-parallel text base. This method can utilize limited aligned resource to expand the search in unlimited non-aligned resources.

Levels of text alignment: Basically, text alignment can be divided into four main different levels: paragraph, sentence, phrase and word. In the paragraph alignment level, researchers are interested in matching the paragraph from a source text to its corresponding translation in the target text. This usually acts as a pre-processed step for further levels of alignment type. Gelbukh and Sidorov (2006) suggested that the alignment of bilingual texts in paragraph level can overcome the case which translators may split up or merge the translated fragments.

In the sentence alignment level, usually one-to-one, one-to-many, many-to-one or many-to-many mappings happen in the alignment process and the results can only represent which group of sentence is aligned to each other. Definitely, the group here does not mean as large as a paragraph, but consists of 1 or 2 sentences.

In the phrase and word level, they are responsible in matching groups of words and independent words, respectively. Zhang *et al.* (2003) derived a method of phrase alignment of Chinese and English bilingual texts. The author stated that most of the phrase alignment method performs the phrase segmentation procedure before the alignment can start. Named entities can also belong to phrase because they always consist of a group of words and Feng *et al.* (2004) proposed a method on field. Vogel *et al.* (1996) suggested an algorithm in word level alignment by employing Hidden Markov Model. The authors admitted that aligning words perfectly in bilingual text is difficult. However, by estimating the frequency of word correspondence can lead to a satisfactory result at sentence level.

Among different alignment levels, Chuang *et al.* (2002) stated that it is useful to perform sentence alignment with a very high precision. Moreover, since sentence alignment of bilingual corpora has a great effect in different application areas as mentioned previously, it has been considered as one of the hottest research topics now a days.

Related works: Various approaches have been published since the latest 80's. Those techniques can be generally classified into three main categories: lexical based, statistical based and hybrid based.

Lexical based method mainly makes use of lexical information, such as bilingual dictionary, to perform the alignment procedure. The research taken by Kay and Roscheisen (1993) employed a partial alignment of the word level to introduce a maximum likelihood alignment of the sentence level. The authors suggested that the words to be translated to each other should have similar distribution in the bilingual texts. Chen (1993) model constructed a word-to-word translation approach. The translation model estimates the cost of an alignment by dividing the bilingual text into sequence of sentence beads. Each bead contains zero or more sentences of each language and assuming the beads are independent to each other. The final result of the alignment is a sequence of beads with the maximized probability estimated.

In statistical approach, Brown *et al.* (1990) and Gale and Church (1991) showed a method of aligning sentences based on a statistical model of character length. Brown *et al.* (1991) made use of the major and minor anchor points to facilitate the alignment process. On the other hand, Gale and Church (1991) model assigned a probabilistic score to each proposed correspondence of sentence base on the difference of their lengths and the variance of the difference. Vogel *et al.* (1996) applied Hidden Markov Model in word alignment to an English-French corpus based on different features, such as length ratios, alignment probabilities, etc. given from the input bilingual documents but do not include any external lexical information. Smith *et al.* (2010) proposed a ranking model in which determines either a sentence in the source document is parallel or not to a sentence in the target document. Moreover, different features are defined and one of them, is based on Wikipedia markup.

Although, statistical approach gives a high performance, it is not robust enough to solve languages in different families, while lexical approach provides a lot of confirmation details of alignments but cannot solve the sparseness and efficiency problems with large lexical information. As a result, the combination of lexical and statistical based approaches leads to hybrid systems to avoid the intrinsic impediments of each approach.

However, the techniques mentioned in above cannot fulfill our cases. Unlike bilingual texts from similar language families, Portuguese and Chinese are different in their syntax and appearance. The position of constituents are very different and it is also not possible to do similarity measurement on words to assist in the alignment process as previously discussed language pairs case performs. Moreover, we do not limit the input for only parallel corpora but also comparable corpora in the process but many related works usually assume that the corpora are parallel.

With these concerns, in this study, an alignment algorithm that employs the Singular Value Decomposition (SVD) technique for handling bilingual texts in structurally dissimilar, Portuguese-Chinese comparable corpora is

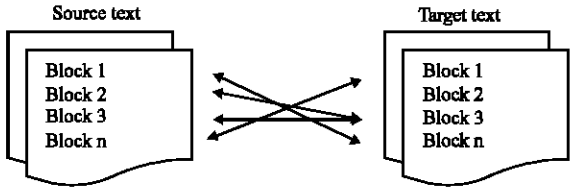

Fig. 1: Cross level alignment relationships

presented. The main objective is not only for handling one-to-one parallel mapping cases but also cross level alignment problems, as shown in Fig. 1 which often appear in Comparable corpora. We define cross level alignment as having blocks which are not aligned in parallel and one source may contain more than one alignment or vice-versa. Each block is a text which can be either a sentence or a paragraph. Block 1 of the source language may align to block n of the target instead of the first one and block 2 and 3 of the source may belong to block 3 of the target.

LANGUAGE DEVIATION PROBLEMS

In order to improve the accuracy in the generation of cross alignment results for structurally dissimilar languages, several language deviation problems between Portuguese and Chinese have been studied and concluded.

Word representation: Portuguese language belongs to the Latin language family and it is often easier for handling bilingual alignments between these languages by considering their shared characteristics. As an example, since their vocabularies are constructed by a group of alphabets, one simple approach is to compare their degree of similarity of words in sentences to tackle the problem. Moreover, since the words in a given text are separated by spaces, their sentence length can also be an important factor in considering their relationships. On the other hand, Chinese is a non-alphabetic language and its words are based on forms of hieroglyphs. Since Chinese sentences consist of several Chinese characters and they often do not have any space to identify themselves, it is always not easy to identify word boundaries in the literature. As a result, this implies that the meaning of a Portuguese word can be either a single Chinese character or a group of Chinese characters. Since their word representation is different, it is not possible to apply alphabetical similarity of lexicons or direct length ratio measures mentioned previously to deal with these languages.

Digit presentation: The presentation of numbers or digits in Portuguese and Chinese languages is different in some cases. When numbers are presented in digits, direct alignments cannot be established in many situations. As an example, suppose that the Portuguese sentence 216.5 mil milhões (216.5 億 thousand millions) is aligned with the Chinese sentence 2,165. The Chinese word億 (hundred million) cannot directly translate to mil (thousand) or "milhões" (million) in Portuguese and the representation of unit is totally different.

Lexicalization, discontinuity and syntactic structure: In terms of lexicalization, information in the corpora may often contain words which are unknown to the system's knowledge. As an example, suppose that there is a bilingual sentence "correio/郵件 (post) in the corpora. The system may not have any knowledge related to "correio" with 億 (post) besides 郵包 (mail parcel). Although, the expression in Chinese is a little bit different, there should be some relationship between them and in the alignment process. The second one is discontinuity relationships between the source and target words. It is often to have words in Portuguese which are composed by two or more Chinese words. As an example, the Portuguese word "emprestar" (lend) is composed two Chinese words 把 and 借給 but they are separated by intermediate Chinese words. Since Portuguese and Chinese do not belong to the same language family, the syntactic structure varies a lot and the position of the constituents is quite different in some sense. Therefore, it is not possible to directly apply some of the approaches mentioned to deal with Portuguese-Chinese alignments.

ALIGNMENT BASED ON SINGULAR VALUE DECOMPOSITION

We introduce the Singular Value Decomposition technique to overcome the cross alignment problem and the language deviations mentioned in the previous section. In a mathematical point of view, SVD is a technique to decompose a matrix W into the product of three matrices as shown in Eq. 1.

$$W = U \times S \times V^T \quad (1)$$

The columns of U are the left singular vectors describe the row entities of W. The columns of V are the right singular vectors which describe the column entities of W. S is a diagonal matrix contains singular values s_1, s_2, ..., s_n which are ordered in a descending way. There is also mathematical proof demonstrating that every matrix has its own decomposition. By formulating th problem into the view of matrix, SVD provides certain

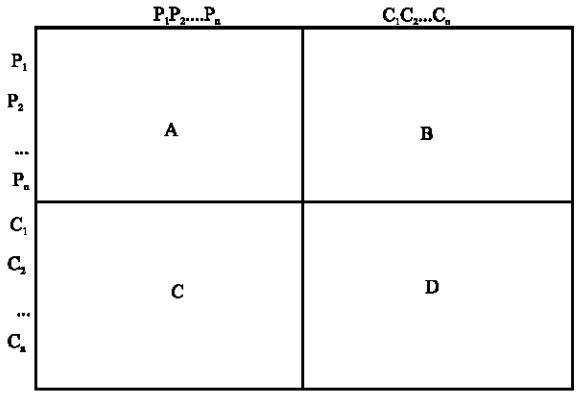

Fig. 2: Cross level alignment relationships

numerical answers. SVD has been applied into many areas such as Latent Semantic Analysis (LSA) and sense disambiguation analysis. In these application areas, they involve in filtering of useless information and relationship analysis. They also show that the application of SVD effectively reduces the dimensionality of the term-document matrix, removes random noise from the matrix and it is sensitive to high-order co-occurrence information which is ignored by the cosine similarity measures.

Statistical relationship matrix: In order to formulate the alignment problem so that SVD can assist the algorithm, a Statistical Relationship Matrix (SRM) is constructed. All the examples described in this section are at sentence level, but it can be also considered as long texts or paragraphs. Figure 2 shows a matrix describing how each Portuguese sentence is related to each Chinese sentence.

SRM is based on idea of occurrences matrix in LSA. The proposed algorithm keeps the matrix structure in estimating the similarity between each row and they are the relationship vectors of each sentence, therefore we need to concatenate the Portuguese and Chinese sentence together. There are four zones in SRM, Zone A consist of entries about only Portuguese sentence and Zone D for only Chinese sentences. These two zones have anchor characteristics for each relationship vectors. In fact, entries with identical sentence pair will assign 1 and others to be 0. For Zone B and Zone C, these are the entries with estimated relationship based on the feature scores of each sentence pair.

Scoring features: Each value in the SRM indicates the relationship r between a Portuguese sentence P and a Chinese sentence C, as shown in Eq. 2. Several features F are defined to deal with language deviation problems and different weights W are assigned accordingly based on their degree of usefulness.

$$r(P, C) = W_1F_1+W_2F_2+W_3F_3+W_4F_4 \le 1 \quad (2)$$

Lexical matching feature: Lexicon matching feature F_1, as shown in Eq. 3, is used to determine how well a given sentence P can be translated into sentence C based on a bilingual dictionary. F_1 is defined as the multiplication of FMatchScore and PenaltyScore and it has a value between 0 and 1.

$$F1 = F\ MatchScore \times PenaltyScore \quad (3)$$

For each word p in P, the corresponding meanings will be retrieved from the dictionary and they are compared with the characters in C. Assume that the set M contains all the Chinese meanings $\{m_1, m_2, ..., m_y\}$ where y denotes the number of meanings found in bilingual dictionary for a specific Portuguese word. Moreover, M_p is defined as the set containing all the meanings m extracted from the bilingual dictionary of word p. For all the words in P with length l_p, FMatchScore is defined in Eq. 4.

$$FMatchScore = \left(\sum_{a=1}^{l_p} \arg\max_{y} P(M_a \mid C)\right) \div l_p \quad (4)$$

For each word p in P, the corresponding meanings will be retrieved from the dictionary and they are compared with the characters in C. When a word in sentence P has a meaning in the bilingual dictionary that matches all the characters with the exact sequence in sentence C, it is treated as a Full Match case. As an example, in the bilingual pair "Direito de posse da conta de correio electrónico 電郵帳戶之擁有權 (Ownership of email account), if the word "conta" (account) with the meaning of 帳戶 (account) matches the characters in sequence, then it is considered as a Full Match case.

Ideally, if all words in P have only one meaning that fully matches the ones in C, then both sentences must be direct alignments between each other. However, this may not happen always in real situations. Each source word may contain several meanings in the dictionary and they do not have exact matches in sequential order with the words in the Chinese sentence. Suppose that the Portuguese word "correio" has two meanings: 郵包 (postal parcel) and 郵遞員 (postman) in the dictionary. When they are compared, none of them will be considered as a Full Match case although the partial meaning character 郵 matches the target pair.

In order to reinforce this shortcoming, based on the characteristics that N-gram has in preserving the sequential order, the proposed Partial Match model is based on N-gram mechanism instead of just considering

the number of characters in the meaning that match the Chinese sentence partially. As a result, Eq. 5 is used in preserving the sequential order of the Chinese meaning.

$$P(m|C)=\left(\sum_{i=1}^{l_m}\frac{exist(C,group_i)}{\#groups}\times w_i\right)\div\left(\sum_{i=1}^{l_m}w_i\right) \quad (5)$$

$group_i$ is a group array of the Chinese characters extracted from the Chinese meaning m, #groups indicates the total number of groups in a specific stage and l_m is the character length of m. The function exist returns the number of groups that matches the characters in C. Moreover, each gram is assigned with a weight w and more is given to higher n-gram matches to indicate a stronger relationship between the meaning and the sentence.

Penalty Score is considered to handle the sub-phrases problem. If a sentence is a sub-sentence of the other one, the system may accidentally align the results wrongly. This is because the FMatchScore only considers Portuguese lexicons that contain Chinese meanings in the dictionary. Moreover, there is a problem in the calculation of a shorter Portuguese sentence with the longer Chinese sentence. As a result, penalty is made on the remaining part of the Chinese characters, as shown in Eq. 6.

$$PenaltyScore = 1-\frac{\#\text{ non Matched Chinese Chars}}{\#\text{ Chinese Chars}} \quad (6)$$

#nonMatchedChineseChars indicates the remaining non-matched Chinese characters and #ChineseChars shows the total number of Chinese characters in the sentence. When #nonMatchedChineseChars is greater, the PenaltyScore becomes smaller and vice versa.

Length ratio feature: Feature F_2 estimates the relationship between bilingual sentences based on the length ratio. Since Portuguese and Chinese are different in terms of word representation, the length ratio cannot be estimated as a direct one to one relationship. According to the experiments performed by Li *et al.* (1999), Portuguese-Chinese alignment systems have good alignment relationships when the length ratio mean equals 0.694. Therefore, F_2 is defined in Eq. 7, in which $length_c$ is the length of Chinese sentence and $length_p$ is the length of Portuguese sentence. mean is a referenced normalized value, when re is greater than 1, F_2 is 0. On the other hand, if the length ratio between the source and target pair tends to the mean, they have a greater chance to be aligned. In this case, F_2 should tend to 1.

$$F_2 = 1-\begin{cases}1, re\geq 1\\ re\end{cases}, re=\left|\frac{length_c}{length_p}-mean\right| \quad (7)$$

Punctuation matching feature: Feature F_3 is responsible to determine the relationship between the bilingual sentences in the sense of punctuations. In the research done by Chuang and Yeh (2005), it shows that punctuations are sometimes an important key of showing relationships when lexicon knowledge is not enough. Since the presentation of punctuations is different in Portuguese and Chinese, a transformation of the Chinese punctuations is first performed. The punctuation matching feature F_3 used is shown in Eq. 8. #punc is the total number of punctuations found in the Portuguese sentence. The exist function returns 1 if the punctuation $punc_a$ in the Portuguese sentence exists in C, otherwise it returns 0. The idea is to retrieve the probability of a punctuation that exists in both Portuguese and Chinese sentence.

$$F_3=\left(\sum_a exist(C,punc_a)\right)\div(\#punc) \quad (8)$$

Digit groups matching feature: A special treatment for texts with many numeric descriptions and decimal numbers is considered to enhance the alignment relationships. The idea is to extract digit groups from the Portuguese sentence and compare them in the Chinese sentence. However, since different representations of units are used, the extracted digit groups should not contain any punctuation except the decimal point. This feature is calculated as in Eq. 9. d_a is the digit group extracted from P and #dg is the total number of digit groups extracted from P. The exist function determines if d_a is part of C or not. It returns 1 if d_a has exact matches in C and 0 otherwise.

$$F_4=\left(\sum_a exist(C,d_a)\right)\div(\#dg) \quad (9)$$

Dimensionality reduction: A dimensionality reduction is performed in the matrix SRM in the generation of a new matrix SRM', as shown in Eq. 10 which is the closest approximation to the original matrix in a k-dimensional space.

$$SRM' = U'\times S'\times V'^T \quad (10)$$

The purpose of k is to restrict the size of matrix S be only keeping the first k singular values and a new matrix S' is generated. Similarly, elimination of row vectors of U and the column vectors of V^T are performed in order to

generate SRM'. In this study, the rank is defined as the sum between the total number of Portuguese sentence (TS_p) and the total number of Chinese sentences (TS_c) divided by two, as shown in Eq. 11. This rank was chosen because in some sense, it represents approximately the number of sentences in the bilingual corpora so that the result can cover the whole text.

$$k = \lfloor (TS_p + TS_c) \div 2 \rfloor \quad (11)$$

The determination of the closeness between words can be based on the use of matrix SRM'. However, in practice, it uses the reduced dimensionality representation of the matrices generated based on SVD to identify sentence relationships. These are represented by the multiplication of U' and S'. Based on the characteristics of SVD, it helps in removing noisy information in the original matrix and reducing the complex work in the analysis of the syntactic issues between two languages. It is believed that this technique effectively weakens the redundant relationship of a wrong bilingual pair and strengthens correct relationships, as to improve the accuracy in cross sentence alignment.

Similarity measurement: Since every row in the matrix represents the characteristics in terms of vector for every Portuguese and Chinese sentence, similarity measurement is performed in order to determine the degree of alignment relationships among the vectors in the matrix. If a bilingual sentence pair is a translation to each other, they should have a higher degree of similarity and vice versa. In this study, we considered the cosine similarity function to calculate the overlapping between two vectors, as defined in Eq. 12.

$$\text{Similarity}(\vec{v}, \vec{w}) = (\sum_{i=1}^{n} v_i \times w_i) / \sqrt{\sum_{i=1}^{n} (v_i)^2 \times \sum_{i=1}^{n} (w_i)^2} \quad (12)$$

Vectors v and w are Portuguese and Chinese sentence vectors with n entries. Every possible bilingual sentence pair will produce a value by Eq. 12 and then the corresponding similarity value will insert into the Alignment Matrix (AM). It is a matrix made up of Portuguese sentence as rows and Chinese sentence as columns. For each entry, it stores the corresponding similarity value between the pair. The higher is the value the greater the chance that *v* and *w* are the translation of each other.

As an example, suppose that the Chinese sentence C_1 should align with P_1. Before the application of SVD with dimensionality reduction, the closeness between them is 0.3178. Once SVD is applied, this value increases sharply to 0.9689. On the other hand, the closeness between pairs which are not related will be decreased accordingly. The decision of the best alignment relationship between a bilingual pair is based on the following criteria: for each row in AM, the process chooses the highest value in the columns for each row if they are higher than the predefined threshold.

EVALUATIONS

In order to evaluate the effectiveness of the proposed framework, experiments are carried out based on different bilingual comparable corpora, including: Macao Year Book 2006 (A), parts of Educational Digest (B), administrative laws of Macao Special Administrative Region (C), Macau Special Administrative Region Policy Address 2007 (D) and internal staff regulations in a company (E). All of them are provided in Portuguese and Chinese covering different domains, as summarized in Table 1.

In these experiments, the corresponding weights are assigned to the features: 0.6 for Lexical Matching F_1, 0.2 for Length Ratio F_2, 0.1 for Punctuation Matching F_3 and 0.1 for Digit Groups Matching Feature F_4. Lexical matching retrieves the highest weight because it is the most reliable resource in this structurally dissimilar language scenario. The proposed framework is compared to a baseline approach which only considers lexical clues based on a bilingual dictionary in the identification of alignment relationships. Evaluation results are shown in Table 2.

Besides the baseline approach and the proposed framework, we also show the precision rate for the alignment task without considering applying the dimensionality reduction of SVD. Moreover, the threshold is defined to capture 70% of the sentences in each of the corresponding corpus. We believed that the sentence pair with a similarity degree higher or equal to the threshold is captured and confirmed to be possible aligned pairs. The accuracy increases sharply based on the SVD technique. Since it filters out the noise and useless relationship information among sentence pairs and makes similar entries in the statistical relationship vector become more similar to each other after the dimensionality reduction, it effectively increases the overall alignment accuracy.

There are still many rooms for improvement. When there are many bilingual sentences, the size of the SRM becomes very large and this directly affects the efficiency of the program in generating alignment relationships. We may either design another data representation model which keeps the matrix in a smaller size or first divide the whole corpus into manageable number of sentences before the SVD is performed by introducing classification

Table 1: Characteristics and size of experimented corpus

Corpus	Type/Domain	Crossing relationship	No. of Port. sentences	No. of Port. words	No. of Chin. sentences	No. of Chin. sentences
A	Statistics	High	89	2564	72	2945
B	Magazine	High	140	3823	108	5024
C	Law	Low	484	7850	484	12148
D	Report	High	145	3774	107	4707
E	Regulation	Low	146	3765	132	4888

Table 2: Comparison between the baseline, with and without SVD in terms of accuracy

Corpus	Sentences captured	Baseline (%)	Without SVD dimensionality reduction (%)	Proposed approach (%)
A	63	68.10	60.53	96.83
B	98	62.50	50.30	89.80
C	339	79.80	52.15	85.84
D	102	51.20	55.43	84.31
E	104	73.40	63.74	96.15

methodologies. On the other hand, since weights are assigned to different features, the proposed approach can be easily extended by taking more clues into consideration. As an example, since the current approach only employs a bilingual dictionary with fundamental vocabularies and translations, the addition of Name Entity lexicon and phrases lexicon can obtain more relationship information between bilingual pairs in terms of specific phrases and proper nouns. Another possible clue is to consider the number of Part-of-Speeches in the bilingual sentence. We believed that they should have the same or similar number disregarding the syntax order.

CONCLUSION

The proposed alignment algorithm based on Singular Value Decomposition in dealing Portuguese-Chinese bilingual comparable corpora is presented in this study. The whole process is divided into three stages: the construction of the Statistical Relationship Matrix and the calculation of the relationship scores; application of SVD; and sentence relationship analysis. The relationship score between each sentence pair is calculated based on different clues, represented by lexical matching, length ratio approximation, punctuation matching and digit groups matching. Moreover, to further improve the ability of handling the linguistic differentiation between Portuguese and Chinese, N-gram mechanism is applied in the lexical matching score. Dimensionality reduction of SVD is applied in strengthening relationships and in filtering noisy information in the original matrix. Similarity measurement based on cosine similarity is performed in revealing correct alignment relationships.

ACKNOWLEDGMENT

This study was partially supported by the Research Committee of University of Macau under grant UL019/09-Y2/EEE/LYP01/FST and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

REFERENCES

Brown, P.F., J. Cocke, S.A.D. Pietra, V.J.D. Pietra and F. Jelinek *et al*., 1990. A statistical approach to machine translation. Comput. Ling., 16: 79-85.

Brown, P.F., J.C. Lai and R.L. Mercer, 1991. Aligning sentences in parallel corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, June 18-21, Berkeley, CA., USA., pp: 169-176.

Chen, S.F., 1993. Aligning sentences in bilingual corpora using lexical information. Proceedings of 31st Annual Meeting of the Association for Computational Linguistics, June 22-26, Stroudsburg, PA., USA., pp: 9-16.

Chuang, T.C., G.N. You and J.S. Chang, 2002. Adaptive bilingual sentence alignment. Lect. Notes Comput. Sci., 2499: 21-30.

Chuang, T.C. and K.C. Yeh, 2005. Aligning parallel bilingual corpora statistically with punctuation criteria. Int. J. Comput. Ling. Chin. Lang. Process., 10: 95-122.

Feng, D., Y. Lv and M. Zhou, 2004. A new approach for English-Chinese named entity alignment. Proceedings of the International Conference on Empirical Methods in Natural Language Processing, Oct. 30, Barcelona, Spain, pp: 372-379.

Gale, W.A. and K.W. Church, 1991. A program for aligning sentences in bilingual corpora. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, June 18-21, Berkeley, CA., USA., pp: 177-184.

Gaussier, E., 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Aug. 10-14, Montreal, Canada, pp: 444-450.

Gelbukh, A.F. and G. Sidorov, 2006. Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming. Lect. Notes Comput. Sci., 4225: 824-833.

Hiemstra, D., 1998. Multilingual Domain Modeling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. In: Computational Linguistics in the Netherlands 1997, Coppen, P.A., H. van Halteren and L. Teunissen (Eds.). Rodopi B.V., Netherlands, pp: 41-58.

Kay, M. and M. Roscheisen, 1993. Text-translation alignment. Comput. Ling., 19: 121-142.

Li, Y.P., C.M. Pun and F. Wu, 1999. Portuguese-Chinese machine translation in Macao. Proceedings of 7th Machine Translation Sunmit, September 1999, Singapore, pp: 236-243.

Ng, H.T., B. Wang and Y.S. Chan, 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, July 2003, Sapporo, Japan, pp: 455-462.

Smith, J.R., C. Quirk and K. Toutanova, 2010. Extracting parallel sentences from comparable corpora using document level alignment. Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 1-6, Los Angeles, California, USA., pp: 403-411.

Vogel, S., H. Ney and C. Tillmann, 1996. HMM-based word alignment in statistical translation. Proc. 16th Int. Conf. Comput. Ling., 2: 836-841.

Yang, Y., J.G. Carbonell, R.D. Brown and R.E. Frederking, 1998. Translingual information retrieval: Learning from bilingual corpora. Artif. Intell., 103: 323-345.

Zhang, Y., S. Vogel and A. Waibel, 2003. Integrated phrase segmentation and alignment algorithm for statistical machine translation. Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, Oct. 26-29, Beijing, China, pp: 567-573.