Translation Relationship Quantification: A Cluster-Based Approach and its Application to Shakespeare's Sonnets

Fazlı Can, Ethem F. Can, Ceyhun Karbeyaz

Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
{canf, efcan, karbeyaz}@cs.bilkent.edu.tr

Abstract. We introduce a method for quantifying translation relationship between source and target texts. In this method, we partition source and target texts into corresponding blocks and cluster them separately using word phrases extracted by a suffix tree approach. We quantify the translation relationship by examining the similarity between source and target clustering structures. In this comparison we aim to observe that their similarity is meaningful, i.e., it is significantly different from random. The method is based on the hypothesis that similarities and dissimilarities among the source blocks will not be lost in translation and reappear among target blocks. For testing we use Shakespeare's sonnets and its translation in Turkish. The results show that our method successfully quantifies translation relationships.

1 Introduction

During translation a text in a particular source language is comprehended and expressed in another target language by retaining the same meaning. Different translations of a given source text can communicate the same message without losing the original meaning. Researchers define objective and quantitative metrics to evaluate the accuracy and quality of translations. In some fields, like automotive service information, there are some metrics for measuring translation quality. Such metrics usually focus on lexical, semantic, and syntactic errors.

In this work, we introduce a method that aims to quantify translation relationship with no information other than the texts themselves. The method is simple and intuitive. In order to quantify translation relationship, we divide source text and its translation into corresponding blocks and cluster source and target blocks separately by defining each text block by a document description vector as is done in information retrieval. After clustering, similar blocks become the member of joint clusters and dissimilar blocks appear in different clusters. Intuitively we expect to have significant similarity between source and target clustering structures, since source and target texts are supposed to have the same meaning and hence similarities/dissimilarities among source blocks

E. Gelenbe et al. (eds.), *Computer and Information Sciences*, Lecture Notes in Electrical Engineering 62, DOI 10.1007/978-90-481-9794-1_24,

reappear among target blocks. For meaningful translations we aim to quantify the similarity between source and target texts and show that their similarity is significant, i.e., not by chance and significantly different from random. In the experiments we use all 154 sonnets of Shakespeare [1] and their Turkish translations by Halman [2]. The experiments show that our findings match with the observations of [3, pp. 352-353].

The main contribution of this study is a statistical, language-independent translation relationship quantification method. The method works without requiring any knowledge about the languages involved. We perform an in depth analysis of Shakespeare's sonnets with Turkish translations. This paper provides some sample results from our findings.

Previous work on the problem of phrasal translation comparison is done by Munteanu and Marcu [4]. They use suffix trees for both source and target languages and benefit from a bilingual lexicon to provide correspondence between them. They successfully create alignments for the source and target languages that have the same word order (e.g. English and French). Unlike their approach of comparing the languages having a common word order, in this study we analyze English and Turkish literary works which have different grammar forms and we use sonnet numbers instead of words as our cluster members.

2 The Method

For translation relationship quantification we first divide the source text into blocks. In some cases blocking may be available in a natural way (for example in our experiments each sonnet of Shakespeare is treated as a block). Word phrases are identified before indexing operation performed. Phrases and words used for indexing are referred to as terms. We make use of a generalized suffix tree approach to identify the word phrases. Rather than simply using single words we aim to identify frequently used phrases, since we intuitively expect that word phrases will be a better reflection of translations. Next, we cluster the source documents and target documents using the k-means clustering algorithm for $k = \{3, 5, 7, ..., 51\}$. After obtaining the clusters we compare source and target clustering structures to measure the similarity between them.

We hypothesize that if target text is a (meaningful) translation of source text, then the clustering structures C_s (source) and C_t (target) should have a meaningful similarity, i.e., a similarity which is significantly different from random, and hence, this implies a meaningful translation relationship.

We use a modified form of the formula developed by Yao [5] for measuring the similarity between source and target clustering structures. Originally, Yao's formula determines the number of disk pages to be accessed to retrieve the related records of a query under the assumption that database records are randomly distributed among the same size pages. Later Can and Ozkarahan [6] adapted the formula for environments for pages (clusters) with different sizes. For using Yao's formula in our problem we treat the individual clusters of C_s as queries

and determine how their members (like the related documents of a query) are distributed in the clustering structure C_t.

We use the Monte Carlo approach [7] and generate a large number of random clustering structures for C_t (during randomization the number of clusters and the size of the individual clusters of C_t are kept the same as given to and determined by the k-means clustering algorithm). We obtain the baseline distribution for n_{tr} values (average number of target cluster in random clustering), which is the similarity between C_s and these randomized C_t clustering structures. We decide that n_t value (actual number of target clusters in C_t) is significantly different from the n_{tr} value if it is smaller than most of the random observations. In the experiments we generate 1000 random cases. We refer to the entity n_t as the Translation Relationship Index (TRI) and check the merit of the index by comparing it with the value of n_{tr}. The existence of n_{tr}, which can be directly computed by the modified Yao's formula [6], gives TRI the attribute of a measurement criterion, since n_{tr} provides a benchmark or a reference point. If the observed TRI value indicates that the relationship is different from random (i.e., if n_t is smaller than n_{tr}), we obtain the baseline distribution for n_{tr} using the Monte Carlo approach to decide if the difference is significant.

3 Experimental Results

Before performing indexing operation, we remove the punctuation marks in the blocks and all letters are converted to lower case. During indexing we use three stemming approaches. One is no stemming, in this approach words are used as they are. The others are a pseudo stemmer, which uses the first five characters of words (we prefer a pseudo stemming approach because of its simplicity) for Turkish translation, and Porter Stemming Algorithm [8] for the original documents.

Table 1. Similarity between original sonnets of Shakespeare and their Turkish translations by Halman: n_t vs. n_{tr}values ($k = 33$, μ: mean value, σ: standard deviation).

Stemming	n_t	n_{tr}: μ (σ)
No Stem.	3.76	4.16 (0.094)
With Stem.	3.63	4.24 (0.089)

Considering the different values of k during the clustering process, $k = 33$ provides the most significant difference, i.e., the lowest p-values which are computed using z-scores. The results for Shakespeare's sonnets and their Turkish translations by Halman can be seen in Table 1. The n_{tr} mean values are exactly the same as the ones obtained by the modified Yao's formula.

As provided in Table 1, TRI values (n_t) are smaller than those of the Monte Carlo n_{tr} values. This means that the cluster formations for the translator show

a similarity to the clustering structure of the original sonnets, which is different from random. In most of the cases considering the different values of k, the results of two-tailed z-tests indicate that, for both stemmed and unstemmed versions, those cases are statistically significantly different from random.

4 Conclusion and Future Work

In this study, we propose a cluster-based approach for quantifying translation relationship between source and target texts. Our method can only indicate that a translation retains the meaning of source text and does not say anything about the style. It can be used with any pair of languages without requiring any knowledge about them. In the experiments we successfully quantify relationship between Shakespeare's sonnets and its Turkish translations.

Our method can complement and be used in connection with other more conventional translation quality measurement methods which are based on lexical, semantic, and syntactic type information. As future work, further experiments with additional data can be interesting, e.g., doing a similar investigation using translations of the sonnets in other languages. It would be interesting to investigate the effects of block size granularity on performance. Furthermore, the method can be used for plagiarism detection both in intra- and inter-language platforms.

Acknowledgments. This research is supported by the Scientific and Technical Research Council of Turkey (TÜBİTAK) under the grant number 109E006. We would like to thank Pınar Duygulu, Jon M. Patton, and literary scholars Talât Sait Halman and Bülent Bozkurt for their helpful pointers and constructive comments.

References

1. Ledger, G.: Shakespeare's sonnets. http://www.shakespeares-sonnets.com/sonn01.htm (2010)
2. Halman, T.S.: William Shakespeare Soneler, 8th ed. Dünya Kitapları, İstanbul (2004)
3. Enginün, İ.: Türkçede Shakespeare. Dergah Yayınları, İstanbul (2008)
4. Munteanu, D.S., Marcu, D.: Processing comparable corpora with bilingual suffix trees. In: EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Morristown, NJ, USA (2002) 289–295
5. Yao, S.B.: Approximating block accesses in database organizations. Commun. ACM **20**(4) (1977) 260–261
6. Can, F., Ozkarahan, E.A.: Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. ACM Trans. Database Syst. **15**(4) (1990) 483–517
7. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988)
8. Porter, M.F. In: An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 313–316