Probabilistic Neural Network Based English-Arabic Sentence Alignment

Mohamed Abdel Fattah[1], Fuji Ren[1], Shingo Kuroiwa[1]
[1] Faculty of Engineering, University of Tokushima
2-1 Minamijosanjima
Tokushima, Japan 770-8506
{mohafi, ren, kuroiwa} @is.tokushima-u.ac.jp

Abstract. In this paper, we present a new approach to align sentences in bilingual parallel corpora based on a probabilistic neural network (P-NNT) classifier. A feature parameter vector is extracted from the text pair under consideration. This vector contains text features such as length, punctuation score, and cognate score values. A set of manually aligned training data was used to train the probabilistic neural network. Another set of data was used for testing. Using the probabilistic neural network approach, an error reduction of 27% was achieved over the length based approach when applied on English-Arabic parallel documents.

1 Introduction

Recently, much work has been reported on sentence alignment using different techniques [1]. Length-based approaches (length as a function of sentence characters [2] or sentence words [3]) were the most interesting. These approaches work quite well with clean input, such as the Canadian Hansards corpus, however they do not work well with noisy document pairs. Moreover, these approaches require that the paragraph boundaries be clearly marked, which is not the case for most of document pairs. Cognate approaches have also been proposed and have been combined with length-based approaches to improve alignment accuracy [4, 5]. They have used sentence cognates such as digits, alphanumerical symbols, punctuation, and alphabetical words. However both of Simard and Thomas did not take the text length between two successive cognates (Simard case) or punctuations (Thomas case) into account which increased the system confusion that leads to execution time increase and accuracy decrease (we have avoided this drawback in this work).

In this paper we present a non-traditional approach for the sentence alignment problem. In the sentence alignment problem, we may have 1-0 (One English sentence does not match any of the Arabic sentences), 0-1, 1-1, 1-2, 2-1, 2-2, 1-3 and 3-1. There may be more categories in bi-texts, however they are rare, hence we consider only the previous mentioned categories. As illustrated above, we have eight sentence alignment categories. Hence, we can consider sentence alignment as a classification problem. This classification problem may be solved by using a probabilistic neural network classifier.