Published: December 15, 2022 | Pages: 64-66

MACHINE TRANSLATION TO ALIGN PARALLEL TEXTS

Sukhrob Sobirovich Avezov

Lecturer Department Of Russian Literary Studies Bukhara State University, Uzbekistan

ABSTRACT: This article proposes a procedure for aligning parallel texts using an online translator of source text sentences. The result of the translation is compared with the translation made by a professional translator and these two translations are aligned using dynamic programming tools. The method was tested on parallel corpora of A. Kadiri's novel "Past Days" translated into English and Russian. Continuation of work involves the fragmentation of sentences into phrases and words.

KEYWORDS: Machine translation, dynamic programming, novel by A. Kadiri, "Past Days", parallel texts.

INTRODUCTION

Parallel linguistically meaningful texts are important in a number of areas of natural language processing and lexicographic applications, in particular in the field of example-based machine translation and in Translation Memory (TM) systems. TM looks for the best match between sentences in the source and target texts and stores a pair of sentences in the machine's memory. When trying to translate new text, the TM system searches for the closest source language (SL) sentence in the device's memory and outputs a parallel target language (TL) sentence. The problem, the solution of which is essential in this approach, is to establish correspondence between units of texts in different languages at the level of sentences, phrases, and even at the level of words. Several approaches have been proposed to solve the problem of matching text units at different levels.

THE MAIN FINDINGS AND RESULTS

[1] described a method based on the number of words contained in sentences, in addition, he considers some anchor points and paragraph markers. This method has been applied to Hansard Corpus with 96-97% accuracy. [2] proposed a method based on a simple statistical model of sentence lengths. The model is based on the observation that longer sentences in one language tend to be translated by longer sequences in another language, and vice versa. A probability value is assigned to each pair of sentences based on the ratio of their lengths and the variance of that

Published: December 15, 2022 | Pages: 64-66

ratio. Although GaleChurch's apparent efficiency has been tested in various languages, it runs into problems when handling complex alignments, i.e. when one sentence of the source text matches several sentences of the target text or vice versa, or when several sentences are translated by several but their boundaries do not match.

It should be noted that the proposed methods involve the widespread use of bilingual dictionaries for word-by-turn comparison of sentences in SL and TL [3]. At the same time, for most pairs of languages, there are no bilingual machine-readable dictionaries, and even paper-based dictionaries. In the case of the presence of the latter, the translation of the dictionary into a machine-readable form requires significant labor costs and does not always give an accurate result due to errors in recognition, editing, and reconciliation. The proposed approach is based on the use of a multilingual online translator. Such translators are produced by many leading Internet companies, Microsoft, Yandex, etc. We use Google translator, which currently processes more than 100 languages and, accordingly, about 10,000 language pairs. The list of languages is constantly expanding, and the quality of translation is also improving. It is important for us that the words SL are translated by the most frequent equivalents of TL. Further, by means of dynamic programming, the sentences of the same digital language are compared, namely, the sentences of the translation made by a professional translator and Google translation, thus eliminating the need to use bilingual dictionaries.

The correspondence between the sentences of the source and target texts is very often not one-to-one, i.e. one sentence of the source text can correspond to several sentences of the translation and vice versa; some sentences and entire paragraphs of the source text may be missing in the translation, sentence boundaries may not coincide, i.e. a group of words in the translation goes into the next sentence, and so on. [4]. Especially often the lack of a one-to-one correspondence between sentences and phrases in pairs of texts is typical for the translation of works of art. When leveling at the sentence level, purely structural (by length, number of words) and statistical methods (by the frequency of constituent words) are used, which can be used for languages with a small resource base [5]. Length justification methods are very sensitive to sentence gaps or insertions, in the sense that a single gap or insertion may result in incorrect subsequent alignment from the point of the gap or insertion to the end of the text. Statistical methods also often give erroneous alignment results, requiring costly manual verification and correction later on. For scientific texts, the transcription method is often used, since many scientific terms come from the same source - Greek, Latin, and later from English, German, French. The terms matched in this way serve as reference points for further alignment. The use of bilingual dictionaries for text alignment is less common and has been used mainly for specialized texts (English-French minutes of the Canadian Parliament, EC legal texts, program specifications, etc.). The alignment method proposed by us contains certain limitations, namely: a) the order of sentences in the Uzbek and foreign texts is the same; b) there are no significant (more than 200 words) omissions in the TL; c) the length of parallel texts is not too large - about 60 thousand word usages.

Published: December 15, 2022 | Pages: 64-66

First of all, it is necessary to divide the text of the SL (Uzbek language) into semantic meaningful parts, sentences or parts of a sentence, most often separated by punctuation marks. Period, question and exclamation marks, semicolon, colon, ellipsis are chosen as separators in the Uzbek text. Dots after abbreviations, initials, etc. should be excluded from the set of separators. The division into semantically significant parts is also carried out for the CL text with some modifications. In particular, in texts in Russian, the end of the sentence is marked with a colon (:).

CONCLUSION

A procedure for aligning parallel texts at the sentence level is proposed. The procedure uses a machine translation system (Google translation), which allows, in the absence of a bilingual machine-readable dictionary, to translate the source / target text by sentences and then compare this translation with the sentences of the target / source text. As a measure of proximity between sentences, one can use the number of words that coincide or are close to spelling without resorting to a morphological analysis of word forms. The dynamic programming procedure finds the optimal path (in the sense of the largest number of matching words) from the beginning of texts to their end. At the same time, 85% of all proposals are compared. The remaining gaps are caused, as a rule, by the translation of one sentence - two or more, or vice versa - two or more sentences of the SL are translated by one sentence of the TL. In these cases, the specified segments are merged.

REFERENCES

1. P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, R.L. Merceer. The mathematics of statistical machine translation: parameter estimation // Computational Linguistics. 1993. Vol. 19 (2).
2. Sobirovich A. S. Development of a Parallel Corpus of the Uzbek and Russian Languages //Vital Annex: International Journal of Novel Research in Advanced Sciences. – 2022. – Т. 1. – №. 5. – С. 152-155.
3. Khamidovna N. L. Expression of the Harmony of Language and Culture in World and Uzbek Lexicography //resmilitaris. – 2023. – Т. 13. – №. 1. – С. 233-244.
4. Sharipov S. S. " LEXICOGRAPHY" AND DICTIONARY COMPILATION //Scientific reports of Bukhara State University. – 2021. – Т. 5. – №. 2. – С. 52-66.
5. Sobirovich A. S. Lecturer at the Department of Russian Language and Literature Bukhara State University //Scientific reports of bukhara state university. – С. 86.