Corpus-Based Translation Study on English-Persian Verb Phrase Ellipsis

Summary

The present research adopted a descriptive corpus-based translation approach and focused on the patterns of translation of English Verb Phrase Ellipsis (VPE) into Persian. The goal was to find out how the observed translation behavior may be taken as advantageous information for improving English-Persian Machine Translation (MT) systems performances.

For this purpose, a bilingual English-Persian parallel corpus was used. It consisted of 1,600 movies´ subtitles, consisting of informal conversations, with about 4 million words for each language. Unitex finite-state tools were applied in order to detect the intended English VPE by defining certain search patterns. The extracted cases of VPE were then compared with their Persian counterparts.

Analysis of the Persian translations provided by the translators and the strategies utilized by them in dealing with English VPE was indicative of the fact that in those cases where Persian and English show some similar VPE constructions, the human translator keeps the translation quite close to the original text, especially by retaining the ellipsis. However, in many cases, the elliptical forms are language-specific. In such cases, it is not possible to keep the ellipsis and the translator has to render the text in a non-elliptical form in order to provide the appropriate text, so to comply with Persian grammatical norms; that is, the gap resultant of VPE in English is usually recovered by the antecedent verb or replaced by a pro-verb.

When the two languages present similar construction, Google translator (GT) also produces a quite reasonable translation. However, in cases where Persian does not allow ellipsis, GT fails to recover the gap left by zeroed material in the source text. Auxiliary verbs also pose some specific problems, as GT translate them into light or lexical verbs.

The analysis of data was based on the following order: VPE after auxiliary verbs; VPE after

complementizer `to´; and VPE in the presence of pro-forms.

The results indicate that human translator dealing with English VPE predominantly adopts the strategy of recovering the zeroed verb from its previous occurrence in discourse. Naturally, in some cases, instead of a verb, a pro-verb is used. For light verb constructions in Persian, the tendency is towards retaining the light verb and zeroing the nominal component. For a residual number of cases the strategies were non-literal.

This general behavior, however, depends on the auxiliary verb, used in the source language. Differences in the kind of the auxiliary verb in English VPE, thus, have a relevant bearing on the choice of the strategies the human translator adopted. For instance, English VPEs occurring after auxiliaries `do´, `be´, and `have´ cannot be translated into Persian by keeping the ellipsis; therefore, the gap is usually filled by the antecedent verb or a pro-verb. However, if the English sentence carries a VPE after auxiliary `be´ and the sentence is translated into Persian using passive voice, then the VPE can be kept. Persian allows keeping the gap produced by the English VPE when this involves the modal verbs `can´, `may´, or `must/have to´, if they are translated as مجبور بودن (majboor boodan) [OBLIGED+BE/GR]).

In case of English VPEs occurring after infinitival complementizer `to´, the translation is mostly by filling the gaps with the antecedent verb. For English VPEs with pro-form structures with `so/too/as well/neither/either´, the translation, for the most part, is by using pro-forms, and thus, keeping the ellipsis.

GT produces distorted translations when dealing with English VPEs occurring after tense operators, since it translate these auxiliaries literally, and also because it does not recover the gap resulting from the English elliptical sentence. For VPEs after modal verbs, GT performs quite acceptably but only after modal `can´; however it fails in dealing with other modal verbs.

GT, in all cases, retains the VPE after complementizer `to´; thus, the output is often unnatural. And, finally, GT, in dealing with VPE in presence of pro-forms, mostly produces inadequate translations.

As a statistical-based MT system, GT does not take into consideration the discourse previous to the sentence under processing. Therefore, it seems incapable to recover the gap induced by English VPE, which results in incorrect translation output in many cases.

The comparison between HT and GT of Persian texts indicates that a stronger effort should be invested in an anaphora resolution module, particularly, for certain English VPE patterns: those auxiliary verbs` do´, `be´, `have´, and `will´, and those after complementizer `to´.

The findings of this study may help devise better performing strategies for English-Persian MT systems, namely, by highlighting the relevance of an anaphora resolution module prior to the MT of this language pair.