An In-depth Comparative Analysis of the Filipino-Cebuano Parallel Corpus Utilizing the Contrastive, Typological, and Translation Mining Traditions

Shamier K. Animeta[1], Ma. Juliet Vasay[1], Kristine Mae M. Adlaon[2]

College of Arts and Humanities[1], Mindanao Natural Language Processing R and D Lab[2]

University of the Immaculate Conception

Abstract

Formulating an acceptable parallel corpus for conforming towards the authenticity of the language within the domain of translation is rather an arduous task. An existing parallel corpus was built as a data set to be used to train a neural machine translation system. The training of the NMT would likely undergo the requirement of expanding the inputs of syntax, semantic, grammatical rules and structure, in order to maximize its ability to comprehend what is being prompted in a pragmatic sense; within the nuances of Natural Language Processing (NLP) in each of the monolinguistic parallel corpora of both Filipino and Cebuano for translation. This paper conducted an in-depth comparative analysis of the parallel corpora utilizing the approaches of translation language representative traditions. These are segmented into three traditions, namely: the contrastive tradition, the typological tradition, and the translation-mining tradition. Within the three traditions, there are apparent relative inconsistencies in terms of the syntax and semantics in both parallel corpora data transcriptions, while the pragmatic nuances of the parallel corpora remain present. Therefore, the parallel corpora of Filipino-Cebuano deviates from the grammatical conformity and alignment of representativeness in the translation of one corpus to another, which this paper tackled comprehensively.

Keywords: Parallel Corpora, Translation, Natural Language Processing, Comparative Analysis, Filipino-Cebuano Language.

1 Introduction

Translation is a process of methodically rendering and transferring the entities of linguistic discourse from one language to another. During translating, comprehending the meaning of the source text is essential in order to have the acceptable equivalent and representativeness in the target language; hence, meaning is translated in connection to syntax, style, and sounds [1]. The More so, the utilization of the parallel corpora is of interest providing an opportunity into aiding translation particularly within machine translation systems. This approach can benefit translators in developing structured translation processes for words or phrases that do not have an immediate equivalent in the target language [2]. Furthermore, the translation direction is unlikely to be constant, thus some texts in a parallel corpus may have undergone translation on one source text to possible multiple target languages [3]. The prominence of Neural Machine Translation, led to maximize the capacity of translation performance into further comprehending the full structure of language, along with the aid of the Natural Language Processing to supplement the formulation of the parallel corpora to abide within authenticity. Moreover, aside from the authenticity of the language, another significant aspect of translation is the accuracy along with the representativeness of meaning between two monolingual parallel corpora in order to attain clarity and quality translation. However, in the collection and analysis of corpus data in usage-based approaches to linguistic study is predicated on a single significant assumption: the corpus is representative of the language phenomena under inquiry and consideration [4]. More so, a according to a relevant quote from Biber (1993:244) “representativeness refers to the extent to which a sample includes the full range of variability in a population.” To which it is referring towards the extralinguistic criteria of language that contributes the distribution to suffice the data needed for the corpus. Therefore, the availability of extensive data is rather limited within the bounds of both Cebuano and Filipino language. Filipino is widely spoken in the Philippines with a similar pattern that has complex morphology; with an adaptable order, that ranges its sentence forms that can be arranged in rather six distinctive approaches like SVO, VSO, and VOS. While Cebuano is asserted to follow the VSO [5] [6]. Due to the morphological complexities along with the varying syntax, semantics, and pragmatics would possibly hinder the target language representativeness; particularly when the monolinguistic parallel corpora came from a low-resource database. Hence, having the said parallel corpora of both undergo an evaluation of representativeness through the translation traditions before NLP.

Related work

The nuances of the Target language representativeness have been methodical way to evaluate the translated parallel corpora by utilizing the three strategic traditions since there are assumptions relating to translation: “translations are representative of their target language, and they convey the same meaning as their originals.” To which should not be taken in absolution, as the parallel corpora has taken the acknowledgement that translation differs [7]. More so, target language representativeness started with the foundational work of Langacker (1987;1991) [8] regarding with cognitive grammar and linguistics, along with the concern of the way of how users (native speakers) represent, analyze, and utilize language. Parallel to this, corpus linguistics has emerged as a single source of supporting data for enhancing descriptions of language structures and usage. It is a strategy that experimentally investigates the use of language in large and systematic

collections of authentic texts using automated and/or computerized technologies and a combination of both qualitative and quantitative methods [9]. However, in the early 2000s, the same empirical results emerged, which prompted some linguists to abandon parallel corpus research, relying on the topic of representativeness [10]. The collection and analysis of corpus data in usage-based approaches to linguistic study is predicated on one main assumption: that the corpus is representative of the language phenomena under investigation. Of certainly, corpus representativeness is a structure, both theoretically and empirically [11].

Methods

A comparative analysis of the two monolinguistic parallel corpora of both the Filipino and Cebuano Translation; to which the data transcription utilized in this paper is gathered differently. The Filipino corpus sample was sourced by perusing through a plethora of existing machine translated dataset, obtained within various translation websites (web-crawled) that already contained the Filipino language dataset for translation, it is then later gathered and complied into one large transcript. While the Cebuano corpus sample was crowd sourced; in which it is gathered through snowball sampling, by having a group of individuals with a range of average to fluent proficiency within the Cebuano Language. To which they are recorded and undergo extensive transcribing to formulate the transcription of the Cebuano corpora. When the corpus of each language is complete, it is then sent for natural language processing. However, in order to achieve natural language level translation. The Parallel corpora of Filipino and Cebuano should undergo an evaluation of representativeness before the word alignment.

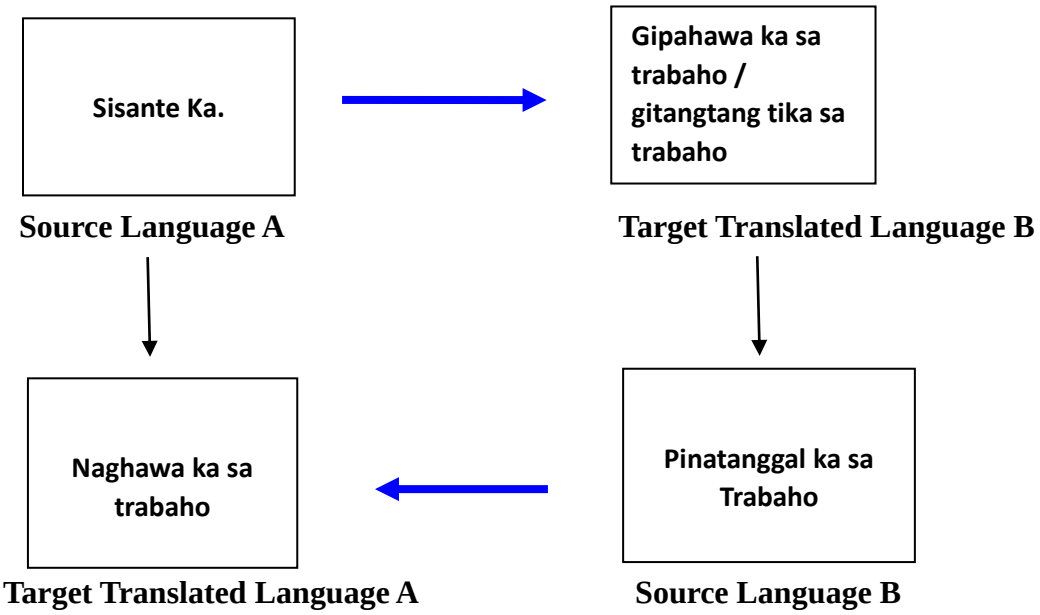
As each of the monolinguistic parallel corpus of Filipino and Cebuano would likely constrict the corpora in terms of the conformity, and alignment, as they equally vary within the ambiguity of the vocabulary between both languages. Moreover, the fact that both languages differ greatly in terms of the given affixes, to which they also differ in functioning morphologically, that could possibly hinder the meaning representativeness, that could alter the syntax of both language causing to have significant impact in both word and sentence alignment through the NLP.

Contrastive Tradition

The contrastive tradition deals with the issue of representativeness, the comparison among the parallel corpora is translation based; to which the given is then the translation equivalence of language. The comparison is done through a process of having to compare multilingual data to the extracted monolingual data, between the source and the translation. During the evaluation of the parallel corpora of Filipino Corpora, the issue of representativeness is rather clear within the bounds of the structural aspect of language, such as syntax and morphology. As stated previously, a difference in terms of the linguistic structure between two parallel corpora could alter the meaning of a particular word or sentence; especially when they are interpreted differently in another data. To which it is quite evident within the given data transcript of both Filipino and Cebuano.

The web-crawled translation of Filipino as evaluated, the structure was overall grammatically erroneous; there is an evident alteration in terms of the meaning equivalence, and representativeness within the semantic of the language. While the transcript of the Cebuano as evaluated, is syntactically and morphologically intact. The representativeness and the equivalence do not par with the semantics; to which has the tendency to borrow particular words from another language to a point that the Cebuano translation is rather narrowed and simplified. however, the transcript of both the Filipino and Cebuano Corpora has a rather narrowed fluency. To which it is typical to word loan, just to fill the gap of the word equivalent to the within the target language.

Contrastive Architecture.



(Figure 1.)

This figure illustrates the comparison and contrast between the translated corpora of both Filipino and Cebuano. Within the date set given, we only have one source of each language; Hence, to create a contrast, we gather other sources that are purely written in either Filipino or Cebuano as a basis for the source be contrast. In this figure, we took the phrase "*Sisante ka*" (Filipino A) from the given data set. However, the supposed Cebuano equivalent is" naghawa *ka sa trabaho*" (Cebuano A); to which it is then evident that it is not the exact equivalent; there are nuances of comprehending the word (pragmatic sense). It is then pointed out that it is syntactically incorrect (grammatical sense). Affixations play a role in Languages that could alter what word or phrase could mean. More so, there would be a variance in meaning, e.g. Cebuano A suggests that the person is quitting rather than being fired, than what Filipino A intended to

which it to Fire the person. Therefore, it does not match, cannot be deemed as the exact equivalent or the aligned translation. In Contrast to the phrase *"Sisante ka"* is another phrase that pars with Filipino A which is *"pinataggal ka sa trabaho"* (Filipino B) a simple and comprehensible phrase that is used to fire a person. The equivalent is *"gipahawa ka sa trabaho"* (Cebuano B). To find the representative for Filipino B is to dissect the morphology and syntax and pay attention of the equivalent root word and the corresponding affixes leading to the change in the morpheme of a word.

Typological Tradition

Typological tradition acknowledges the issue of representativeness yet does not actively control within the issue. Relies on the contextual details of what is being translated; therefore, focusing on semantical and pragmatic structure in the aspect of the language. To which presumably the typological tradition dotes on the ambiguity of language within the bounds of translation, to provide a full lexical analysis in the domain of context, safeguarding from the full influence of the translation approach of word for word. In the evaluation of the parallel corpora of Filipino and Cebuano, it is evident that hey both par within the pragmatical structures.

Even though the Filipino Corpus is a Web-crawled data transcript that is syntactically inaccurate; it still pars with the semantics of the language, hence, remained comprehensible yet it still indefinite to be an exact output for translation, as ambiguity comes with variance in the lexicon of the language. While the Cebuano corpus was pragmatically and semantically intact. However, while comparing the Cebuano Corpus to the Filipino Corpus, the equivalence, often are not in par with one another, as there are factors that hinders the alignment of context within both monolinguistic parallel corpora. It is worth pointing out the possible key factors that cause misalignment within the context between the source and the target translated language, which is: the morphology, the syntax, sentences structures, and the affixations of a language. Furthermore, aligning the context relies on the structure of the said factors.

Translation Mining Tradition

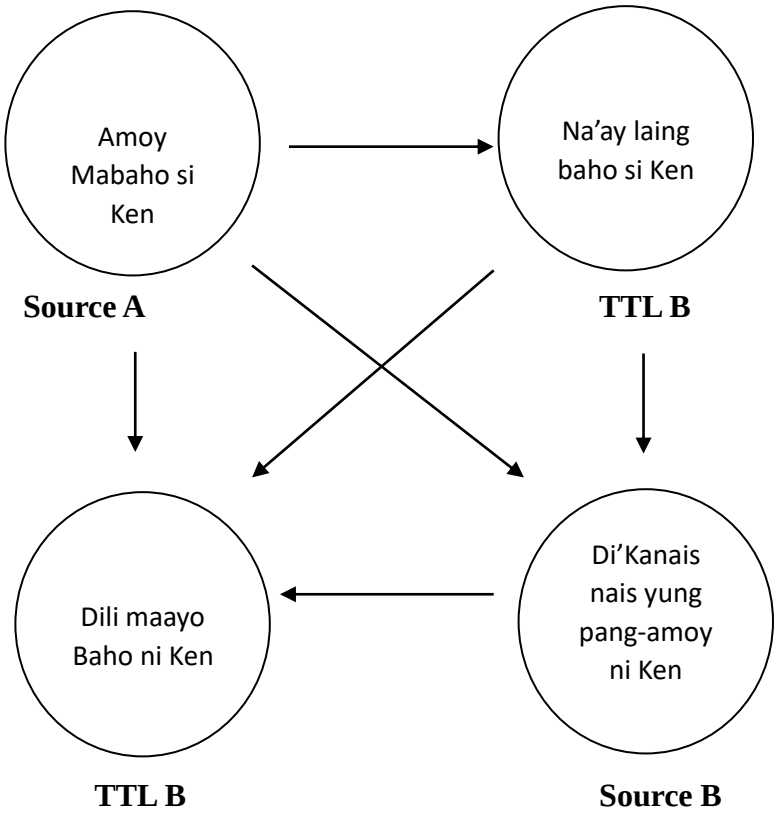
This tradition is a mix of both contrastive and typological tradition; to which it puts both the translated and untranslated text at the same level. While putting individual items in individual languages. More so, it still succeeds in dealing with representativeness by relying on the judgement of the native speakers as the initial check on the acceptability and naturalness of the data. Furthermore, the tradition pars within the "Have-Perfect" approach as it deals whether the translation is accurate and within the sufficient equivalence that the parallel corpus could possibly emulate that of the native speaker to achieve the state of "natural language". To which when the data is then fed to an A.I. what is being translated could finally be par the accuracy. However, as evaluating the Filipino corpus, it is deemed to sound incorrect for native speakers. How come it is incorrect? The linguistic structure is inaccurate, thus, hindering the given context,

especially that we should also acknowledge the fact that the corpus was web-crawled and underwent a series of MT.

Therefore, the translation would be completely different form how Filipino is often utilized. In comparison to the unnatural state of the given Filipino transcript; while the given data of the Cebuano translation came from set of participants who has the capabilities to articulate the Cebuano language. However, in the judgment of a native speaker the translation is to be deemed inaccurate. As previously stated in the past tradition, it is rather evident that the state of the corpus; it does not reach the accuracy of the language. The corpus is narrowed and simplified, perpetuating the basic linguistic structure of the Cebuano Language

Translation Mining Architecture

English: "*Ken smells unpleasant*"



(Figure 2)

This Figure illustrates a vast variation of the supposed aligned translation of both Filipino and Cebuano. one source is Web-crawled, while the other was garnered through native speakers of the languages involved. Through the Translation Mining Tradition, this strategy would rely on the judgement of the native speaker of the language; especially towards the representativeness of word. In the figure we use one phrase e.g., "*Ken smells unpleasant"* there is a variety of

translation with the phrase, even in both Filipino and Cebuano. Source A leans towards the literal translation of the Phrase; e.g., *Amoy Mabaho = smells unpleasant* in contrast to source B, it is then determined by a native speaker of Filipino, a deep rooted variant of Filipino that is formal and well aligned. E.g., *Di' kanais nais = unfavorable*, *pang-amoy = smells*. And within the judgement of the Filipino, Both are acceptable and reliable translation, the only thing that hinder one of them is the grammatical aspect in particular Source A. In the case of TTL A (Cebuano A) is the literal translation of the English phrase, e.g., *Dili Maayo = Unpleasant, Baho=smells* while on the other hand the TTL B (Cebuano B) leans to another context, e.g., *laing baho=other smells*. Natives speakers would likely to choose the TTL B, since it is more contextually fair, rather than being literal. More so, this proves to why Translation Mining is a complementary study as it should base on the Native speaker's judgement and satisfaction.

Conclusion

Each monolinguistic parallel corpus of Filipino and Cebuano will likely constrain the corpora in terms of conformance and alignment since they differ similarly within the ambiguity of the lexicon between both languages. They both have distinct language functions. As one must examine the issue of their variation in structure, which runs from the set of morphemes to the language's provided semantics. This is far too large to be supplied to A.I. It could need more work, but the three ways could limit the extensibility of a language's data collection. Particularly, the issue of representativeness truly revolves around the Parallel Corpora of Filipino and Cebuano; for the translation to conform to one another. It is not centered around the root word as there are aspects of both languages. The morphology, syntax, sentence patterns, and affixations of a language are all possibly significant variables that produce misalignment within the context of the source and destination translated languages. In order to avoid these variables, a study that lines with a contrastive strategy with the supplement study of translation mining. The contrastive tradition strategy deals with the grammatical aspect of language to par within the sematic and pragmatic nuances.

References

[1] Ghazala, Hasan, (1995) Translation as problems and solutions (4th ed.) Syria: Dar Elkalem ElArabi.

[2] https://gotranscript.com/blog/the_importance_of_equivalence_in_translation

[3] Sinclair.J, n.d. University of Birmingham, School of English
http://www.ilc.cnr.it/EAGLES96/corpustyp/node3.html

[4] Sophie Raineri and Camille Debras (2019) Corpora and Representativeness: Where to go from now?
https://doi.org/10.4000/cognitextes.1311

[5] Biber, Douglas. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257
http://dx.doi.org/10.1093/llc/8.4.243

[6] Jenn Leana Fernandez and Kristine Mae Adlaon (2022) Exploring Word Alignment Towards an Efficient Sentence Aligner for Filipino and Cebuano Languages
https://aclanthology.org/2022.loresmt-1.13.pdf

[7] Jun Tariman. 2010. Cebuano 101: The cebuano language sentence structure. pages 22–26.

[8] Le Bruyn, Bert et.al (2022) Parallel Corpus Research and Target Language Representativeness: The Contrastive, Typological, and Translation Mining Traditions

[9] Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Vol. II: Descriptive application.* Stanford: Stanford University Press.
DOI : 10.1515/9780804764469

[10] Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
DOI : 10.1017/CBO9780511804489

[11] McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. Corpus-Based Language Studies: An Advanced Resource Book. London and New York: Routledge.

[12] Halliday, Michael A. K. 2005. Computational and quantitative studies, volume 6. In *The collected works of M. A. K. Halliday*. Hong Kong: Continuum