The neural machine translation of dislocations

Behnoosh Namdarzadeh, Nicolas Ballier
School of English Studies, CLILLAC-ARP, Université Paris Cité, France

https://doi.org/10.36505/ExLing-2022/13/0035/000577

Abstract

This paper investigates neural machine translation (NMT) outputs for dislocated constructions from French into English. Dislocations are often considered to be "substandard in formal registers" (Lambrecht 1994: 12). In French, multiple copies of the subject are licit in spoken data, whereas translations into English preclude them (De Cat 2007). We analysed 436 translations of French dislocated segments in the novel *Voyage au bout de la nuit* (Céline 1932) and a contemporary corpus for spoken data from Corpus de Français Parlé Parisien (CFPP) (Branca-Rosoff & Lefeuvre 2016) by DeepL and Google. Beyond prototypical *X, c'est* dislocations, translation toolkits continue to misfire, and this might be due to the lack of spoken data in training sets of NMT.

Keywords: dislocation, language pairs French English, neural machine translation

Introduction

This paper replicates the challenge set approach proposed by Pierre Isabelle and colleagues for English into French (Isabelle et al. 2017). The idea is to target the difficult linguistic features and observe the output of machine translation systems like Google Translate and DeepL. Our focus for our challenge set is dislocations; structures where double constituents are licit in the source text (like two grammatical subjects in French) but not in English.

Dislocations are universal (Lambrecht 1994) and all languages seem to have identical forms of topic-marking. Syntactically, two positions can be considered. One is called *theme* or *Left Dislocation* (LD), a clause and a constituent to its left. In the example "[Peter] I've known him for a long time" (Westbury 2016), Peter is a dislocated segment which occurs before the clause. The other is right dislocation (RD). In the example "He lived in Africa, [the wizard] (Lambrecht 1994) right dislocation has the wizard at the right edge of the sentence. Pragmatically, a linear arrangement of linguistic elements in a sentence affects information packaging. Our dislocation challenge set of French examples encompasses a pragmatic need (usually expressed by thematization) and a syntactic constraint (only one subject) for the translations into English.

Methodology

Our corpus includes *Voyage au bout de la nuit* from the INTERSECT parallel corpus (Salkie 2022). The reference translation in English of the French sentences in which dislocation occurs was obtained using AntPConc software.

In addition to this classic subcorpus, a contemporary corpus for spoken data from the Corpus de Français Parlé Parisien (CFPP) des années 2000 (Branca-Rosoff & Lefeuvre 2016) was also searched using the Universal Dependency (UD) annotation to retrieve the possible dislocations with the dependency relation. Annotation was performed with the {UDpipe} package in R. The corpora yielded 2,546 occurrences, out of which we analysed 436 translations by DeepL and Google translate. We briefly report the discrepancies in the translations of dislocations observed in our corpus.

Results

While we do not report recall for the automatic detection of dislocations with UD, the precision of our retrieval method was pretty accurate for the 218 analysed dislocations (91% and 98 %). Few false positives were detected but more frequently for the written data (appositions and parentheticals mistaken as dislocations) than for the spoken data (repairs and repetitions).

Table 1. Distribution of main dislocation types and success rate for the *c'est* dislocation in our data.

corpora	multiple	c'est	"subject copy" in the translation
Voyage (n=109)	15	62	Google = 17, DeepL = 11
CFPP (n=109)	22	73	Google = 30, DeepL = 32

The detailed typology of dislocations observed in the data is beyond the remit of this paper. We describe the complexity of the dislocations in Table 1 by reporting multiple cases of dislocations within sentences, much more frequent in our spoken data. We focus on the dominant type of dislocations (*<left dislocated item>, c'est* dislocations) and compare the two toolkits on their ability to produce translations that avoid the repetition of the subject (we call it "subject copy"). *C'est* constructions often have *ça* as a left dislocated constituent like *ça c'est vrai* (Céline 1932), which is translated as *that is true*. Despite the high number of occurrences of *c'est* construction in both of our corpora (62 occurrences out of 109 sentences for the novel and 73 occurrences out of 109 sentences for CFPP), there is still a deficiency in translating this construction by the toolkits. Google tends to produce more "subject copies" in typical examples. The picture is more blurred for more complex cases like *l'amour c'est elle la misère …* (Céline 1932), Google outperforms DeepL and translates it as *love is misery …* with the suppression of the extra subject which is not required in English, whereas DeepL output is closer to the ST translation suggesting *the love it is it the misery… .*

Left dislocated items can also be stacked as instances with more than two constituents for topic-marking function (Raquel 2002). For example, *in Lui, le père, je l'apercevais…*(Céline 1932), which includes double topicalization, all the initial subjects are translated by the two translation toolkits as *him, the father, I*

…. . In the other example, *moi ça m'a toujours semblé… normal…* (Branca-Rosoff, Lefeuvre 2016), Google keeps the structure of double topicalization as in the French source text and translates it into *me it always seemed to me… normal*, whereas, DeepL omits the extra subject and translates it into *I always thought it was… normal.*

Many dislocations have a tonic pronoun as the left dislocated item, prototypically the *moi, je* construction in French sentences. It can be tricky for the MT systems. In translating the source text *mais autrement non moi je trouve j'aime bien* (Branca-Rosoff, Lefeuvre 2016), Google follows word-by-word strategy translating it into *but otherwise not me I find I like.* DeepL omits this part of the source text and suggests *but otherwise I like it.*

Discussion and conclusion

We retained the original absence of punctuation (commas) of the CFPP, which is even more distinct from the canonical training data of the toolkit. For instance, translating the French ST *le parc Mabille[,] c'est parc des Beaumonts maintenant[,] ils l'ont bien aménagé* (Branca-Rosoff & Lefeuvre 2016), DeepL suggests *the park Mabille it is park of Beaumonts now they arranged it well.* The zero article for *Parc de Beaumonts* may account for the absence of recognition of the pattern. We revised the transcription of CFPP and added commas where appropriate to check the ability of the MT toolkits. Re-punctuating the sentences (see our [,]) did not solve the *it* subject copy issue.

Analysing only 436 translations of dislocated constructions produced by Google and DeepL still outlines meaningful patterns for the toolkit translations of this tricky structure. The partial success with *c'est* dislocation suggests training data is crucial for the results. Les frequent structures tend to be mistranslated, especially for spoken data. Overall, the challenging dislocated segments mainly originate from spoken language and this might suggest that more spoken data should be included in the training sets of neural machine translation.

While parentheticals seem to ease the translation of dislocations, more complex structures with stacking remain an issue for NMT toolkits. The topic-marking function and some tropicalized object constructions can also be challenging.

The overall patterns differ in the two toolkits. On the one hand, Google seems to keep the paratactic structure of the French source sentence, i.e., to produce a word-by-word translation of the structure and to reiterate the words. If anything, Google tends to be more source-based for the translation of dislocations and more systematically preserves the original punctuation. On the other hand, DeepL outputs are hypotactic in a sense that even for double constituents, the toolkit suppresses reiterations or links the constituents using subordinating conjunctions.

Acknowledgements

Part of this research was financed within the SPECTRANS project, funded under the 2020 émergence research project, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

References

Branca-Rosoff, S., Lefeuvre, F. 2016. Le Corpus de Français Parlé Parisien des années 2000: Constitution, outils et analyses. Le cas des interrogatives indirectes. Corpus 15, 265–284. https://doi.org/10.4000/corpus.3043

Céline, L.-F. 1932. Voyage au bout de la nuit. Paris, Gallimard.

De Cat, C. 2007. French Dislocation: Interpretation, Syntax, Acquisition. Oxford, Oxford University Press.

Isabelle, P., Cherry, C., Foster, G. 2017. A Challenge Set Approach to Evaluating Machine Translation. Association for Computational Linguistics, 2486–2496. https://doi.org/10.18653/v1/D17-1263

Lambrecht, K. 1994. Information structure and sentence form: Topic, focus, and the mental representations of discourse referents. Cambridge, Cambridge University Press.

Raquel, H. 2002. Establishing topic in conversation: A contrastive study of left-dislocation in English and Spanish. Circle of Linguistics Applied to Communication, 31–50, Department of Spanish Philology, Universidad Complutense de Madrid.

Salkie, R. 2022. INTERSECT. http://arts.brighton.ac.uk/staff/raf-salkie/intersect

Westbury, J. 2016. Left dislocation: A typological overview. Stellenbosch Papers in Linguistics Plus, 50(0), 21-45. https://doi.org/10.5842/50-0-715