Automatically Extracting Challenge Sets for Non-local Phenomena in Neural Machine Translation - Supplementary

Leshem Choshen[1] and Omri Abend[1,2]
[1]School of Computer Science and Engineering, [2] Department of Cognitive Sciences
The Hebrew University of Jerusalem
leshem.choshen@mail.huji.ac.il, oabend@cs.huji.ac.il

1 Manual Evaluation of Extraction Procedure

Manual evaluation of the sentences extracted using our procedure was performed using two proficient annotators (authors of this paper), one for each source language. These include 180 source German sentences extracted from Books, and 81 English sentences including all the instances extracted from News and 45 extracted from Books. Within Books, sentences are distributed uniformly across phenomena and d values $d \in \{1, 2, 5\}$.

In German, phrasal verbs are detected with high precision: 96% of the sentences indeed include a phrasal verb LDD. For reflexive verbs, 63% include reflexive verbs with a distance of at least 1, and two thirds of the remaining cases (25%) include verbal non-reflexive pronouns with $d \geq 1$ (in German, some pronouns may be used both as reflexive and non-reflexive). While non-reflexive verbal pronouns are not lexical LDDs (as they can mostly be translated word by word), they do challenge the system to disambiguate them from reflexive verbs. Our analysis in the next section shows that the extracted non-reflexive cases present similar trends as the reflexive cases.

In English (Table 1) detection precision is high for reflexive and phrasal verbs. Preposition stranding detection precision is lower. However, wrongly extracted examples mostly involve prepositional objects that are elided or difficult to detect. We therefore consider the difficulties such cases pose as sufficiently similar to the ones posed by preposition stranding.

2 Manual MT Performance evaluation

Using the same sample of 180 sentences used for German detection (See Manual Validation in the paper), we analyze the performance of the Transformer using in-house annotators. One annotator (an author of the paper), proficient in English and German, was presented with the German source in which the relevant tokens were marked. The annotator was asked to locate and mark the corresponding part in the English reference. Places in which the gold translation did not contain a translation of the phenomena, usually due to alignment errors in the corpus or complete omission by the human translator, are removed from the analysis. Then, two annotators (a different author and a non-author), proficient in English, were asked to judge whether the Transformer output conveys the meaning marked in the reference. Inter-annotator agreement was computed to be $\kappa = 0.79$.

Results (Table 2) show a decrease in performance when increasing the distance d. With reflexive verbs, this effect is smaller between $d = 1$ and $d = 2$. However, looking at each category separately (reflexive or non-reflexive pronouns) shows that performance decreases with d in all cases (Table 3).

	News	Books
Reflexive	0.91	0.87
Preposition Stranding	0.75	0.60
Particle	0.94	1.00

Table 1: Ratio of extracted sentences that indeed present the target lexical LDD in English. Rows correspond to various lexical LDD types, and column correspond to corpora.

	Amount			Accuracy		
$d =$	1	2	5	1	2	5
Particle	28	26	26	0.68	0.58	0.42
Reflexive	20	24	14	0.50	0.50	0.29
All	48	50	40	0.60	0.54	0.38

Table 2: Results of manual annotation of translation quality per lexical LDD phenomena in German to English translation with the Transformer. Left: amount of sentences annotated for each type. Right: accuracy (ratio of cases deemed to be correctly translated). Columns correspond to the distance d, as judged by the annotators. Numbers reported are after removing extraction errors and disagreements.

	Amount			Accuracy		
$d =$	1	2	5	1	2	5
Non-reflexive	3	15	4	0.67	0.53	0.25
Reflexive	17	9	10	0.47	0.44	0.30

Table 3: Results of manual annotation of translation quality per sub-type of reflexive verbs in German to English translation with the Transformer. Left: amount of sentences annotated for each type. Right: accuracy (ratio of cases deemed to be correctly translated). Columns correspond to the distance d, as judged by the annotators. Numbers reported are after removing extraction errors and disagreements.