

ate in-house PBMT and NMT systems as well as Google’s GNMT system.

In addition to proposing the novel idea of a challenge set evaluation, our contribution includes our annotated English–French challenge set, which we provide in both formatted text and machine-readable formats (see supplemental materials). We also supply further evidence that NMT is systematically better than PBMT, even when BLEU score differences are small. Finally, we give an analysis of the challenges that remain to be solved in NMT, an area that has received little attention thus far.

2 Related Work

A number of recent papers have evaluated NMT using broad performance metrics. The WMT 2016 News Translation Task (Bojar et al., 2016) evaluated submitted systems according to both BLEU and human judgments. NMT systems were submitted to 9 of the 12 translation directions, winning 4 of these and tying for first or second in the other 5, according to the official human ranking. Since then, controlled comparisons have used BLEU to show that NMT outperforms strong PBMT systems on 30 translation directions from the United Nations Parallel Corpus (Junczys-Dowmunt et al., 2016a), and on the IWSLT English-Arabic tasks (Durrani et al., 2016). These evaluations indicate that NMT performs better on average than previous technologies, but they do not help us understand what aspects of the translation have improved.

Some groups have conducted more detailed error analyses. Bentivogli et al. (2016) carried out a number of experiments on IWSLT 2015 English-German evaluation data, where they compare machine outputs to professional post-edits in order to automatically detect a number of error categories. Compared to PBMT, NMT required less post-editing effort overall, with substantial improvements in lexical, morphological and word order errors. NMT consistently outperformed PBMT, but its performance degraded faster as sentence length increased. Later, Toral and Sánchez-Cartagena (2017) conducted a similar study, examining the outputs of competition-grade systems for the 9 WMT 2016 directions that included NMT competitors. They reached similar conclusions regarding morphological inflection and word order, but found an even greater degradation in NMT performance as sentence length increased, perhaps due

to these systems’ use of subword units.

Most recently, Sennrich (2016) proposed an approach to perform targeted evaluations of NMT through the use of contrastive translation pairs. This method introduces a particular type of error automatically in reference sentences, and then checks whether the NMT system’s conditional probability model prefers the original reference or the corrupted version. Using this technique, they are able to determine that a recently-proposed character-based model improves generalization on unseen words, but at the cost of introducing new grammatical errors.

Our approach differs from these studies in a number of ways. First, whereas others have analyzed sentences drawn from an existing bitext, we conduct our study on sentences that are manually constructed to exhibit canonical examples of specific linguistic phenomena. We focus on phenomena that we expect to be more difficult than average, resulting in a particularly challenging MT test suite (King and Falkedal, 1990). These sentences are designed to dive deep into linguistic phenomena of interest, and to provide a much finer-grained analysis of the strengths and weaknesses of existing technologies, including NMT systems.

However, this strategy also necessitates that we work on fewer sentences. We leverage the small size of our challenge set to manually evaluate whether the system’s actual output correctly handles our phenomena of interest. Manual evaluation side-steps some of the pitfalls that can come with Sennrich (2016)’s contrastive pairs, as a ranking of two contrastive sentences may not necessarily reflect whether the error in question will occur in the system’s actual output.

3 Challenge Set Evaluation

Our challenge set is meant to measure the ability of MT systems to deal with some of the more difficult problems that arise in translating English into French. This particular language pair happened to be most convenient for us, but similar sets can be built for any language pair.

One aspect of MT performance excluded from our evaluation is robustness to sparse data. To control for this, when crafting source and reference sentences, we chose words that occurred at least 100 times in our training corpus (section 4.1).¹

¹With two exceptions: *spilt* (58 occurrences), which is

The challenging aspect of the test set we are presenting stems from the fact that the source English sentences have been chosen so that their closest French equivalent will be *structurally divergent* from the source in some crucial way. Translational divergences have been extensively studied in the past—see for example (Vinay and Darbelnet, 1958; Dorr, 1994). We expect the level of difficulty of an MT test set to correlate well with its density in divergence phenomena, which we classify into three main types: morpho-syntactic, lexico-syntactic and purely syntactic divergences.

3.1 Morpho-syntactic divergences

In some languages, word morphology (e.g. inflections) carries more grammatical information than in others. When translating a word towards the richer language, there is a need to recover additional grammatically-relevant information from the context of the target language word. Note that we only include in our set cases where the relevant information is available in the *linguistic* context.²

One particularly important case of morpho-syntactic divergence is that of *subject-verb agreement*. French verbs typically have more than 30 different inflected forms, while English verbs typically have 4 or 5. As a result, English verb forms strongly underspecify their French counterparts. Much of the missing information must be filled in through forced agreement in person, number and gender with the grammatical subject of the verb. But extracting these parameters can prove difficult. For example, the agreement features of a coordinated noun phrase are a complex function of the coordinated elements: a) the gender is feminine if all conjuncts are feminine, otherwise masculine wins; b) the conjunct with the smallest person ($p1 < p2 < p3$) wins; and c) the number is always plural when the coordination is “et” (“and”) but the case is more complex with “ou” (“or”).

A second example of morpho-syntactic divergence between English and French is the more explicit marking of the *subjunctive mood* in French

part of an idiomatic phrase, and *guitared* (0 occurrences), which is meant to test the ability to deal with “nonce words” as discussed in section 5.

²The so-called Winograd Schema Challenges (en.wikipedia.org/wiki/Winograd_Schema_Challenge) often involve cases where common-sense reasoning is required to correctly choose between two potential antecedent phrases for a pronoun. Such cases become En → Fr translation challenges if the relevant English pronoun is *they* and its alternative antecedents happen to have different grammatical genders in French: *they* → *ils/elles*.

subordinate clauses. In the following example, the verb “partiez”, unlike its English counterpart, is marked as subjunctive:

He demanded that you leave immediately. → Il a exigé que vous *partiez* immédiatement.

When translating an English verb within a subordinate clause, the context must be examined for possible subjunctive triggers. Typically these are specific lexical items found in a governing position with respect to the subordinate clause: verbs such as “exiger que”, adjectives such as “regrettable que” or subordinate conjunctions such as “à condition que”.

3.2 Lexico-syntactic divergences

Syntactically governing words such as verbs tend to impose specific requirements on their complements: they *subcategorize* for complements of a certain syntactic type. But a source language governor and its target language counterpart can diverge on their respective requirements. The translation of such words must then trigger adjustments in the target language complement pattern. We can only examine here a few of the types instantiated in our challenge set.

A good example is *argument switching*. This refers to the situation where the translation of a source verb V_s as V_t is correct but only provided the arguments (usually the subject and the object) are flipped around. The translation of “to miss” as “manquer à” is such a case:

John misses Mary → Mary *manque à* John.

Failing to perform the switch results in a severe case of mistranslation.

A second example of lexico-syntactic divergence is that of “crossing movement” verbs. Consider the following example:

Terry swam across the river → Terry *a traversé* la rivière *à la nage*.

The French translation could be glossed as, “Terry crossed the river by swimming.” A literal translation such as “Terry a nagé à travers la rivière,” is ruled out.

3.3 Syntactic divergences

Some syntactic divergences are not relative to the presence of a particular lexical item but rather stem from differences in the set of available syntactic patterns. Source-language instances of structures missing from the target language must be mapped onto equivalent structures. Here are some of the types appearing in our challenge set.

The position of French pronouns is a major case of divergence from English. French is basically an SVO language like English but it departs from that canonical order when post-verbal complements are pronominalized: the pronouns must then be rendered as *proclitics*, that is phonetically attached to the verb on its left side.

He gave Mary a book. → Il a donné un livre à Marie.

He gave_i it_j to her_k. → Il le_j lui_k a donné_i.

Another example of syntactic divergence between English and French is that of *stranded prepositions*. In both languages, an operation known as “WH-movement” will move a relativized or questioned element to the front of the clause containing it. When this element happens to be a prepositional phrase, English offers the option to leave the preposition in its normal place, fronting only its pronominalized object. In French, the preposition is always fronted alongside its object:

The girl whom_i he was dancing with_j is rich. → La fille avec_j qui_i il dansait est riche.

A final example of syntactic divergence is the use of the so-called *middle voice*. While English uses the passive voice in agentless generic statements, French tends to prefer the use of a special pronominal construction where the pronoun “se” has no real referent:

Caviar is eaten with bread. → Le caviar se mange avec du pain.

This completes our exemplification of morpho-syntactic, lexico-syntactic and purely syntactic divergences. Our actual test set includes several more subcategories of each type. The ability of MT systems to deal with each such subcategory is then tested using at least three different test sentences. We use short test sentences so as to keep

the targeted divergence in focus. The 108 sentences that constitute our current challenge set can be found in Appendix B.

3.4 Evaluation Methodology

Given the very small size of our challenge set, it is easy to perform a human evaluation of the respective outputs of a handful of different systems. The obvious advantage is that the assessment is then absolute instead of relative to one or a few reference translations.

The intent of each challenge sentence is to test one and only one system capability, namely that of coping correctly with the particular associated divergence subtype. As illustrated in Figure 1, we provide annotators with a question that specifies the divergence phenomenon currently being tested, along with a reference translation with the areas of divergence highlighted. As a result, judgments become straightforward: was the targeted divergence correctly bridged, yes or no?³ There is no need to mentally average over a number of different aspects of the test sentence as one does when rating the global translation quality of a sentence, e.g. on a 5-point scale. However, we acknowledge that measuring translation performance on complex sentences exhibiting many different phenomena remains crucial. We see our approach as being complementary to evaluations of overall translation quality.

One consequence of our divergence-focused approach is that faulty translations will be judged as successes when the faults lie outside of the targeted divergence zone. However, this problem is mitigated by our use of short test sentences.

4 Machine Translation Systems

We trained state-of-the-art neural and phrase-based systems for English-French translation on data from the WMT 2014 evaluation.

4.1 Data

We used the LIUM shared-task subset of the WMT 2014 corpora,⁴ retaining the provided tokenization

³Sometimes the system produces a translation that circumvents the divergence issue. For example, it may dodge a divergence involving adverbs by reformulating the translation to use an adjective instead. In these rare cases, we instruct our annotators to abstain from making a judgment, regardless of whether the translation is correct or not.

⁴<http://www.statmt.org/wmt14/translation-task.html>
<http://www-lium.univ-lemans.fr/~schwenk/nmt-shared-task>

corpus	lines	en words	fr words
train	12.1M	304M	348M
mono	15.9M	—	406M
dev	6003	138k	155k
test	3003	71k	81k

Table 1: Corpus statistics. The WMT12/13 eval sets are used for dev, and the WMT14 eval set is used for test.

and corpus organization, but mapping characters to lowercase. Table 1 gives corpus statistics.

4.2 Phrase-based systems

To ensure a competitive PBMT baseline, we performed phrase extraction using both IBM4 and HMM alignments with a phrase-length limit of 7; after frequency pruning, the resulting phrase table contained 516M entries. For each extracted phrase pair, we collected statistics for the hierarchical reordering model of Galley and Manning (2008).

We trained an NNJM model (Devlin et al., 2014) on the HMM-aligned training corpus, with input and output vocabulary sizes of 64k and 32k. Words not in the vocabulary were mapped to one of 100 mkcls classes. We trained for 60 epochs of $20k \times 128$ minibatches, yielding a final dev-set perplexity of 6.88.

Our set of log-linear features consisted of forward and backward Kneser-Ney smoothed phrase probabilities and HMM lexical probabilities (4 features); hierarchical reordering probabilities (6); the NNJM probability (1); a set of sparse features as described by Cherry (2013) (10,386); word-count and distortion penalties (2); and 5-gram language models trained on the French half of the training corpus and the French monolingual corpus (2). Tuning was carried out using batch lattice MIRA (Cherry and Foster, 2012). Decoding used the cube-pruning algorithm of Huang and Chiang (2007), with a distortion limit of 7.

We include two phrase-based systems in our comparison: PBMT-1 has data conditions that exactly match those of the NMT system, in that it does not use the language model trained on the French monolingual corpus, while PBMT-2 uses both language models.

4.3 Neural systems

To build our NMT system, we used the Nematus toolkit,⁵ which implements a single-layer neural sequence-to-sequence architecture with attention (Bahdanau et al., 2015) and gated recurrent units (Cho et al., 2014). We used 512-dimensional word embeddings with source and target vocabulary sizes of 90k, and 1024-dimensional state vectors. The model contains 172M parameters.

We preprocessed the data using a BPE model learned from source and target corpora (Sennrich et al., 2016). Sentences longer than 50 words were discarded. Training used the Adadelta algorithm (Zeiler, 2012), with a minibatch size of 100 and gradients clipped to 1.0. It ran for 5 epochs, writing a checkpoint model every 30k minibatches. Following Junczys-Dowmunt et al. (2016b), we averaged the parameters from the last 8 checkpoints. To decode, we used the AmuNMT decoder (Junczys-Dowmunt et al., 2016a) with a beam size of 4.

While our primary results will focus on the above PBMT and NMT systems, where we can describe replicable configurations, we have also evaluated Google’s production system,⁶ which has recently moved to NMT (Wu et al., 2016). Notably, the “GNMT” system uses (at least) 8 encoder and 8 decoder layers, compared to our 1 layer for each, and it is trained on corpora that are “two to three decimal orders of magnitudes bigger than the WMT.” The evaluated outputs were downloaded in December 2016.

5 Experiments

The 108-sentence English–French challenge set presented in Appendix B was submitted to the four MT systems described in section 4: PBMT-1, PBMT-2, NMT, and GNMT. Three bilingual native speakers of French rated each translated sentence as either a success or a failure according to the protocol described in section 3.4. For example, the 26 sentences of the subcategories S1–S5 of Appendix B are all about different cases of subject-verb agreement. The corresponding translations were judged successful if and only if the translated verb correctly agrees with the translated subject.

The different system outputs for each source sentence were grouped together to reduce the bur-

⁵<https://github.com/rsennrich/nematus>

⁶<https://translate.google.com>

den on the annotators. That is, in figure 1, annotators were asked to answer the question for each of four outputs, rather than just one as shown. The outputs were listed in random order, without identification. Questions were also presented in random order to each annotator. Appendix A in the supplemental materials contains the instructions shown to the annotators.

5.1 Quantitative comparison

Table 2 summarizes our results in terms of percentage of successful translations, globally and over each main type of divergence. For comparison with traditional metrics, we also include BLEU scores measured on the WMT 2014 test set.

As we can see, the two PBMT systems fare very poorly on our challenge set, especially in the morpho-syntactic and purely syntactic types. Their somewhat better handling of lexico-syntactic issues probably reflects the fact that PBMT systems are naturally more attuned to lexical cues than to morphology or syntax. The two NMT systems are clear winners in all three categories. The GNMT system is best overall with a success rate of 68%, likely due to the data and architectural factors mentioned in section 4.3.⁷

WMT BLEU scores correlate poorly with challenge-set performance. The large gap of 2.3 BLEU points between PBMT-1 and PBMT-2 corresponds to only a 1% gain on the challenge set, while the small gap of 0.4 BLEU between PBMT-2 and NMT corresponds to a 21% gain.

Inter-annotator agreement (final column in table 2) is excellent overall, with all three annotators agreeing on almost 90% of system outputs. Syntactic divergences appear to be somewhat harder to judge than other categories.

5.2 Qualitative assessment of NMT

We now turn to an analysis of the strengths and weaknesses of neural MT through the microscope of our divergence categorization system, hoping that this may help focus future research on key issues. In this discussion we ignore the results obtained by PBMT-2 and compare: a) the results obtained by PBMT-1 to those of NMT, both systems having been trained on the same dataset; and b) the

⁷We cannot offer a full comparison with the pre-NMT Google system. However, in October 2016 we ran a smaller 35-sentence version of our challenge set on both the Google system and our PBMT-1 system. The Google system only got 4 of those examples right (11.4%) while our PBMT-1 got 6 right (17.1%).

results of these two systems with those of Google NMT which was trained on a much larger dataset.

In the remainder of the present section we will refer to the sentences of our challenge set using the subcategory-based numbering scheme S1-S26 as assigned in Appendix B. A summary of the category-wise performance of PBMT-1, NMT and Google NMT is provided in Table 3.

Strengths of neural MT

Overall, both neural MT systems do much better than PBMT-1 at bridging divergences. In the case of morpho-syntactic divergences, we observe a jump from 16% to 72% in the case of our two local systems. This is mostly due to the NMT system’s ability to deal with many of the more complex cases of subject-verb agreement:

- *Distractors*. The subject’s head noun agreement features get correctly passed to the verb phrase across intervening noun phrase complements (sentences S1a–c).
- *Coordinated verb phrases*. Subject agreement marks are correctly distributed across the elements of such verb phrases (S3a–c).
- *Coordinated subjects*. Much of the logic that is at stake in determining the agreement features of coordinated noun phrases (cf. our relevant description in section 3.1) appears to be correctly captured in the NMT translations of S4.
- *Past participles*. Even though the rules governing French past participle agreement are notoriously difficult (especially after the “avoir” auxiliary), they are fairly well captured in the NMT translations of (S5b–e).

The NMT systems are also better at handling lexico-syntactic divergences. For example:

- *Double-object verbs*. There are no such verbs in French and the NMT systems perform the required adjustments flawlessly (sentences S8a–S8c).
- *Overlapping subcat frames*. NMT systems manage to discriminate between an NP complement and a sentential complement starting with an NP: cf. *to know NP* versus *to know NP is VP* (S11b–e)
- *NP-to-VP complements*. These English infinitival complements often need to be rendered as finite clauses in French and the NMT systems are better at this task (S12a–c).

Divergence type	PBMT-1	PBMT-2	NMT	Google NMT	Agreement
Morpho-syntactic	16%	16%	72%	65%	94%
Lexico-syntactic	42%	46%	52%	62%	94%
Syntactic	33%	33%	40%	75%	81%
Overall	31%	32%	53%	68%	89%
WMT BLEU	34.2	36.5	36.9	—	—

Table 2: Summary performance statistics for each system under study, including challenge set success rate grouped by linguistic category (aggregating all positive judgments and dividing by total judgments), as well as BLEU scores on the WMT 2014 test set. The final column gives the proportion of system outputs on which all three annotators agreed.

Finally, NMT systems also turn out to better handle purely syntactic divergences. For example:

- *Yes-no question syntax*. The differences between English and French yes-no question syntax are correctly bridged by the two NMT systems (S17a–c).
- *French proclitics*. NMT systems are significantly better at transforming English pronouns into French proclitics, i.e. moving them before the main verb and case-inflecting them correctly (S23a–e).
- Finally, we note that the Google system manages to overcome several additional challenges. It correctly translates *tag questions* (S18a–c), constructions with *stranded prepositions* (S19a–f), most cases of the *in-alienable possession* construction (S25a–e) as well as *zero relative pronouns* (S26a–c).

The large gap observed between the results of the in-house and Google NMT systems indicates that current neural MT systems are extremely data hungry. But given enough data, they can successfully tackle some challenges that are often thought of as extremely difficult. A case in point here is that of stranded prepositions (see discussion in section 3.3), in which we see the NMT model capture some instances of WH-movement, the textbook example of long-distance dependencies.

Weaknesses of neural MT

In spite of its clear edge over PBMT, NMT is not without some serious shortcomings. We already mentioned the degradation issue with long sentence which, by design, could not be observed with our challenge set. But an analysis of our results will reveal many other problems. Globally, we note that even using a staggering quantity of data and a highly sophisticated NMT model, the

Google system fails to reach the 70% mark on our challenge set. The fine-grained error categorization associated with the challenge set will help us single out precise areas where more research is needed. Here are some relevant observations.

Incomplete generalizations. In several cases where partial results might suggest that NMT has correctly captured some basic generalization about linguistic data, further instances reveals that this is not fully the case.

- *Agreement logic*. The logic governing the agreement features of coordinated noun phrases (see section 3.1) has been mostly captured by the NMT systems (cf. the 12 sentences of S4), but there are some gaps. For example, the Google system runs into trouble with mixed-person subjects (sentences S4d1–3).
- *Subjunctive mood triggers*. While some subjunctive mood triggers are correctly registered (e.g. “demander que” and “malheureux que”), the case of such a highly frequent subordinate conjunction as *provided that* → *à condition que* is somehow being missed (sentence S6a–c).
- *Noun compounds*. The French translation of an English compound $N_1 N_2$ is usually of the form $N_2 \text{ Prep } N_1$. For any given headnoun N_2 the correct preposition *Prep* depends on the semantic class of N_1 . For example *steel/ceramic/plastic knife* → *couteau en acier/céramique/plastique* but *butter/meat/steak knife* → *couteau à beurre/viande/steak*. Given that neural models are known to perform some semantic generalizations, we find their performance disappointing on our compound noun examples (S14a–i).

Category	Subcategory	#	PBMT-1	NMT	Google NMT
Morpho-syntactic	Agreement across distractors	3	0%	100%	100%
	through control verbs	4	25%	25%	25%
	with coordinated target	3	0%	100%	100%
	with coordinated source	12	17%	92%	75%
	of past participles	4	25%	75%	75%
	Subjunctive mood	3	33%	33%	67%
Lexico-syntactic	Argument switch	3	0%	0%	0%
	Double-object verbs	3	33%	67%	100%
	Fail-to	3	67%	100%	67%
	Manner-of-movement verbs	4	0%	0%	0%
	Overlapping subcat frames	5	60%	100%	100%
	NP-to-VP	3	33%	67%	67%
	Factitives	3	0%	33%	67%
	Noun compounds	9	67%	67%	78%
	Common idioms	6	50%	0%	33%
	Syntactically flexible idioms	2	0%	0%	0%
Syntactic	Yes-no question syntax	3	33%	100%	100%
	Tag questions	3	0%	0%	100%
	Stranded preps	6	0%	0%	100%
	Adv-triggered inversion	3	0%	0%	33%
	Middle voice	3	0%	0%	0%
	Fronted should	3	67%	33%	33%
	Clitic pronouns	5	40%	80%	60%
	Ordinal placement	3	100%	100%	100%
	Inalienable possession	6	50%	17%	83%
	Zero REL PRO	3	0%	33%	100%

Table 3: Summary of scores by fine-grained categories. “#” reports number of questions in each category, while the reported score is the percentage of questions for which the divergence was correctly bridged. For each question, the three human judgments were transformed into a single judgment by taking system outputs with two positive judgments as positive, and all others as negative.

- The so-called French “inalienable possession” construction arises when an agent performs an action on one of her body parts, e.g. *I brushed my teeth*. The French translation will normally replace the possessive article with a definite one and introduce a reflexive pronoun, e.g. *Je me suis brossé les dents* (‘I brushed myself the teeth’). In our dataset, the Google system gets this right for examples in the first and third persons (sentences S25a,b) but fails to do the same with the example in the second person (sentence S25c).

Then there are also phenomena that current NMT systems, even with massive amounts of data, appear to be completely missing:

- *Common and syntactically flexible idioms*. While PBMT-1 produces an acceptable translation for half of the idiomatic expressions of

S15 and S16, the local NMT system misses them all and the Google system does barely better. NMT systems appear to be short on raw memorization capabilities.

- *Control verbs*. Two different classes of verbs can govern a subject NP, an object NP plus an infinitival complement. With verbs of the “object-control” class (e.g. “persuade”), the object of the verb is understood as the semantic subject of the infinitive. But with those of the “subject-control” class (e.g. “promise”), it is rather the subject of the verb which plays that semantic role. None of the systems tested here appear to get a grip on subject control cases, as evidenced by the lack of correct feminine agreement on the French adjectives in sentences S2b–d.
- *Argument switching verbs*. All systems tested

here mistranslate sentences S7a–c by failing to perform the required argument switch: $NP_1 \text{ misses } NP_2 \rightarrow NP_2 \text{ manque à } NP_1$.

- *Crossing movement verbs*. None of the systems managed to correctly restructure the regular manner-of-movement verbs e.g. *swim across X* \rightarrow *traverser X à la nage* in sentences S10a–c. Unsurprisingly, all systems also fail on the even harder example S10d, in which the “nonce verb” *guitared* is a spontaneous derivation from the noun *guitar* being cast as an ad hoc manner-of-movement verb.⁸
- *Middle voice*. None of the systems tested here were able to recast the English “generic passive” of S21a–c into the expected French “middle voice” pronominal construction.

6 Conclusions

We have presented a radically different kind of evaluation for MT systems: the use of challenge sets designed to stress-test MT systems on “hard” linguistic material, while providing a fine-grained linguistic classification of their successes and failures. This approach is not meant to replace our community’s traditional evaluation tools but to supplement them.

Our proposed error categorization scheme makes it possible to bring to light different strengths and weaknesses of PBMT and neural MT. With the exception of idiom processing, in all cases where a clear difference was observed it turned out to be in favor of neural MT. A key factor in NMT’s superiority appears to be its ability to overcome many limitations of n -gram language modeling. This is clearly at play in dealing with subject-verb agreement, double-object verbs, overlapping subcategorization frames and last but not least, the pinnacle of Chomskyan linguistics, WH-movement (in this case, stranded prepositions).

But our challenge set also brings to light some important shortcomings of current neural MT, regardless of the massive amounts of training data it may have been fed. As may have been already known or suspected, NMT systems struggle with the translation of idiomatic phrases. Perhaps more interestingly, we notice that neural MT’s impressive generalizations still seem somewhat brittle. For example, the NMT system can appear to have

mastered the rules governing subject-verb agreement or inalienable possession in French, only to trip over a rather obvious instantiation of those rules. Probing where these boundaries are, and how they relate to the neural system’s training data and architecture is an obvious next step.

7 Future Work

It is our hope that the insights derived from our challenge set evaluation will help inspire future MT research, and call attention to the fact that even “easy” language pairs like English–French still have many linguistic issues left to be resolved. But there are also several ways to improve and expand upon our challenge set approach itself.

First, though our human judgments of output sentences allowed us to precisely assess the phenomena of interest, this approach is not scalable to large sets, and requires access to native speakers in order to replicate the evaluation. It would be interesting to see whether similar scores could be achieved through automatic means. The existence of human judgments for this set provides a gold-standard by which proposed automatic judgments may be meta-evaluated.

Second, the construction of such a challenge set requires in-depth knowledge of the structural divergences between the two languages of interest. A method to automatically create such a challenge set for a new language pair would be extremely useful. One could imagine approaches that search for divergences, indicated by atypical output configurations, or perhaps by a system’s inability to reproduce a reference from its own training data. Localizing a divergence within a difficult sentence pair would be another useful subtask.

Finally, we would like to explore how to train an MT system to improve its performance on these divergence phenomena. This could take the form of designing a curriculum to demonstrate a particular divergence to the machine, or altering the network structure to capture such generalizations.

Acknowledgments

We would like to thank Cyril Goutte, Eric Joanis and Michel Simard, who graciously spent the time required to rate the output of four different MT systems on our challenge sentences. We also thank Roland Kuhn for valuable discussions, and comments on an earlier version of the paper.

⁸ On the concept of nonce word, see https://en.wikipedia.org/wiki/Nonce_word.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*. San Diego, USA. <http://arxiv.org/abs/1409.0473>.
- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 257–267. <https://aclweb.org/anthology/D16-1025>.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. <http://www.aclweb.org/anthology/W16-2301>.
- Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 22–31. <http://www.aclweb.org/anthology/N13-1003>.
- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 427–436. <http://www.aclweb.org/anthology/N12-1047>.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. <http://www.aclweb.org/anthology/D14-1179>.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1370–1380. <http://www.aclweb.org/anthology/P14-1129>.
- Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics* 20:4. <http://aclweb.org/anthology/J/J94/J94-4004.pdf>.
- Nadír Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI machine translation systems for IWSLT 16. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.
- Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 848–856. <http://www.aclweb.org/anthology/D08-1089>.
- Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 144–151. <http://www.aclweb.org/anthology/P07-1019>.
- Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.
- Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016b. The amu-uedin submission to the wmt16 news translation task: Attention-based nmt models as feature functions in phrase-based smt. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 319–325. <http://www.aclweb.org/anthology/W16-2316>.
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709. <http://www.aclweb.org/anthology/D13-1176>.
- Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 1990 Conference on Computational Linguistics*. Association for Computational Linguistics, Helsinki, Finland. <http://aclweb.org/anthology/C/C90/C90-2037.pdf>.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. <https://doi.org/10.3115/1073083.1073135>.
- Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. *CoRR* abs/1612.04629. <http://arxiv.org/abs/1612.04629>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. <http://www.aclweb.org/anthology/P16-1162>.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pages 3104–3112.
- Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus statistical machine translation for 9 language directions. In *Proceedings of the The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain, pages 1063–1073. <http://aclweb.org/anthology/E/E17/E17-1100.pdf>.
- Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l’anglais*, volume 1. Didier, Paris.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. <http://arxiv.org/abs/1609.08144>.
- Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. <http://arxiv.org/abs/1212.5701>.

A Instructions to Annotators

The following instructions were provided to annotators:

You will be presented with 108 short English sentences and the French translations produced for them by each of four different machine translation systems. You will not be asked to provide an overall rating for the machine-translated sentences. Rather, you will be asked to determine whether or not a highly specific aspect of the English sentence is correctly rendered in each of the different translations. Each English sentence will be accompanied with a yes-no question which precisely specifies the targeted element for the associated translations. For example, you may be asked to determine whether or not the main verb phrase of the translation is in correct grammatical agreement with its subject.

In order to facilitate this process, each English sentence will also be provided with a French reference (human) translation in which the particular elements that support a yes answer (in our example, the correctly agreeing verb phrase) will be highlighted. Your answer should be “yes” if the question can be answered positively and “no” otherwise. Note that this means that any translation error which is unrelated to the question at hand should be disregarded. Using the same example: as long as the verb phrase agrees correctly with its subject, it does not matter whether or not the verb is correctly chosen, is in the right tense, etc. And of course, it does not matter if unrelated parts of the translation are wrong.

In most cases you should be able to quickly determine a positive or negative answer. However, there may be cases in which the system has come up with a translation that just does not contain the phenomenon targeted by the associated question. In such cases, and only in such cases, you should choose “not applicable” regardless of whether or not the translation is correct.

B Challenge Set

We include a rendering of our challenge set in the pages that follow, along with system output for the PBMT-1, NMT and Google systems.⁹ Sentences are grouped by linguistic category and subcategory. For convenience, we also include a reference

translation, which is a manually-crafted translation that is designed to be the most straightforward solution to the divergence problem at hand. Needless to say, this reference translation is seldom the only acceptable solution to the targeted divergence problem. Our judges were provided these references, but were instructed to use their knowledge of French to judge whether the divergence was correctly bridged, regardless of the translation’s similarity to the reference.

In all translations, the locus of the targeted divergence is highlighted in boldface and it is specifically on that portion that our annotators were asked to provide a judgment. For each system output, we provide a summary of our annotator’s judgments on its handling of the phenomenon of interest. We label the translation with a ✓ if two or more annotators judged the divergence to be correctly bridged, and with an ✗ otherwise.

We also release a machine-readable version of this same data, including all of the individual judgments, in the hope that others will find interesting new uses for it.

⁹A machine-readable version is provided in the file `Challenge_set-v2hA.json` in the supplemental materials.

Morpho-Syntactic

S-V agreement, across distractors

Is subject-verb agreement correct? (Possible interference from distractors between the subject's head and the verb).

S1a	Source	The repeated calls from his mother should have alerted us.
	Ref	Les appels répétés de sa mère auraient dû nous alerter.
	PBMT-1	Les appels répétés de sa mère aurait dû nous a alertés. ✗
	NMT	Les appels répétés de sa mère devraient nous avoir alertés. ✓
	Google	Les appels répétés de sa mère auraient dû nous alerter. ✓
S1b	Source	The sudden noise in the upper rooms should have alerted us.
	Ref	Le bruit soudain dans les chambres supérieures aurait dû nous alerter.
	PBMT-1	Le bruit soudain dans les chambres supérieures auraient dû nous a alertés. ✗
	NMT	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
	Google	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
S1c	Source	Their repeated failures to report the problem should have alerted us.
	Ref	Leurs échecs répétés à signaler le problème auraient dû nous alerter.
	PBMT-1	Leurs échecs répétés de signaler le problème aurait dû nous a alertés. ✗
	NMT	Leurs échecs répétés pour signaler le problème devraient nous avoir alertés. ✓
	Google	Leur échec répété à signaler le problème aurait dû nous alerter. ✓

S-V agreement, through control verbs

Does the flagged adjective agree correctly with its subject? (Subject-control versus object-control verbs).

S2a	Source	She asked her brother not to be arrogant .
	Ref	Elle a demandé à son frère de ne pas se montrer arrogant .
	PBMT-1	Elle a demandé à son frère de ne pas être arrogant. ✓
	NMT	Elle a demandé à son frère de ne pas être arrogant. ✓
	Google	Elle a demandé à son frère de ne pas être arrogant. ✓
S2b	Source	She promised her brother not to be arrogant .
	Ref	Elle a promis à son frère de ne pas être arrogante .
	PBMT-1	Elle a promis son frère à ne pas être arrogant. ✗
	NMT	Elle a promis à son frère de ne pas être arrogant. ✗
	Google	Elle a promis à son frère de ne pas être arrogant. ✗
S2c	Source	She promised her doctor to remain active after retiring.
	Ref	Elle a promis à son médecin de demeurer active après s'être retirée.
	PBMT-1	Elle a promis son médecin pour demeurer actif après sa retraite. ✗
	NMT	Elle a promis à son médecin de rester actif après sa retraite. ✗
	Google	Elle a promis à son médecin de rester actif après sa retraite. ✗
S2d	Source	My mother promised my father to be more prudent on the road.
	Ref	Ma mère a promis à mon père d'être plus prudente sur la route.
	PBMT-1	Ma mère, mon père a promis d'être plus prudent sur la route. ✗
	NMT	Ma mère a promis à mon père d'être plus prudent sur la route. ✗
	Google	Ma mère a promis à mon père d'être plus prudent sur la route. ✗

S-V agreement, coordinated targets

Do the marked verbs/adjective agree correctly with their subject? (Agreement distribution over coordinated predicates)

S3a	Source	The woman was very tall and extremely strong .
	Ref	La femme était très grande et extrêmement forte .
	PBMT-1	La femme était très gentil et extrêmement forte. ✗
	NMT	La femme était très haute et extrêmement forte. ✓
	Google	La femme était très grande et extrêmement forte. ✓
S3b	Source	Their politicians were more ignorant than stupid .
	Ref	Leurs politiciens étaient plus ignorants que stupides .
	PBMT-1	Les politiciens étaient plus ignorants que stupide. ✗
	NMT	Leurs politiciens étaient plus ignorants que stupides. ✓
	Google	Leurs politiciens étaient plus ignorants que stupides. ✓
S3c	Source	We shouted an insult and left abruptly.
	Ref	Nous avons lancé une insulte et nous sommes partis brusquement.
	PBMT-1	Nous avons crié une insulte et a quitté abruptement. ✗
	NMT	Nous avons crié une insulte et nous avons laissé brusquement. ✓
	Google	Nous avons crié une insulte et nous sommes partis brusquement. ✓

S-V agreement, feature calculus on coordinated source

Do the marked verbs/adjective agree correctly with their subject? (Masculine singular ET masculine singular yields masculine plural).

S4a1	Source	The cat and the dog should be watched .
	Ref	Le chat et le chien devraient être surveillés .
	PBMT-1	Le chat et le chien doit être regardée. ✗
	NMT	Le chat et le chien doivent être regardés. ✓
	Google	Le chat et le chien doivent être surveillés. ✓
S4a2	Source	My father and my brother will be happy tomorrow.
	Ref	Mon père et mon frère seront heureux demain.
	PBMT-1	Mon père et mon frère sera heureux de demain. ✗
	NMT	Mon père et mon frère seront heureux demain. ✓
	Google	Mon père et mon frère seront heureux demain. ✓
S4a3	Source	My book and my pencil could be stolen .
	Ref	Mon livre et mon crayon pourraient être volés .
	PBMT-1	Mon livre et mon crayon pourrait être volé. ✗
	NMT	Mon livre et mon crayon pourraient être volés. ✓
	Google	Mon livre et mon crayon pourraient être volés. ✓

Do the marked verbs/adjectives agree correctly with their subject? (Feminine singular ET feminine singular yields feminine plural).

S4b1	Source	The cow and the hen must be fed .
	Ref	La vache et la poule doivent être nourries .
	PBMT-1	La vache et de la poule doivent être nourris. ✗
	NMT	La vache et la poule doivent être alimentées. ✓
	Google	La vache et la poule doivent être nourries. ✓

S4b2	Source	My mother and my sister will be happy tomorrow.
	Ref	Ma mère et ma sœur seront heureuses demain.
	PBMT-1	Ma mère et ma sœur sera heureux de demain. ✗
	NMT	Ma mère et ma sœur seront heureuses demain. ✓
	Google	Ma mère et ma sœur seront heureuses demain. ✓
S4b3	Source	My shoes and my socks will be found .
	Ref	Mes chaussures et mes chaussettes seront retrouvées .
	PBMT-1	Mes chaussures et mes chaussettes sera trouvé. ✗
	NMT	Mes chaussures et mes chaussettes seront trouvées. ✓
	Google	Mes chaussures et mes chaussettes seront trouvées. ✓
Do the marked verbs/adjectives agree correctly with their subject? (Masculine singular ET feminine singular yields masculine plural.)		
S4c1	Source	The dog and the cow are nervous .
	Ref	Le chien et la vache sont nerveux .
	PBMT-1	Le chien et la vache sont nerveux. ✓
	NMT	Le chien et la vache sont nerveux. ✓
	Google	Le chien et la vache sont nerveux. ✓
S4c2	Source	My father and my mother will be happy tomorrow.
	Ref	Mon père et ma mère seront heureux demain.
	PBMT-1	Mon père et ma mère se fera un plaisir de demain. ✗
	NMT	Mon père et ma mère seront heureux demain. ✓
	Google	Mon père et ma mère seront heureux demain. ✓
S4c3	Source	My refrigerator and my kitchen table were stolen .
	Ref	Mon réfrigérateur et ma table de cuisine ont été volés .
	PBMT-1	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	NMT	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	Google	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
Do the marked verbs/adjectives agree correctly with their subject? (Smallest coordinated grammatical person wins.)		
S4d1	Source	Paul and I could easily be convinced to join you.
	Ref	Paul et moi pourrions facilement être convaincus de se joindre à vous.
	PBMT-1	Paul et je pourrais facilement être persuadée de se joindre à vous. ✗
	NMT	Paul et moi avons facilement pu être convaincus de vous rejoindre. ✓
	Google	Paul et moi pourrait facilement être convaincu de vous rejoindre. ✗
S4d2	Source	You and he could be surprised by her findings.
	Ref	Vous et lui pourriez être surpris par ses découvertes.
	PBMT-1	Vous et qu'il pouvait être surpris par ses conclusions. ✗
	NMT	Vous et lui pourriez être surpris par ses conclusions. ✓
	Google	Vous et lui pourrait être surpris par ses découvertes. ✗

S4d3	Source	We and they are on different courses.
	Ref	Nous et eux sommes sur des trajectoires différentes.
	PBMT-1	Nous et ils sont en cours de différents. ✗
	NMT	Nous et nous sommes sur des parcours différents. ✗
	Google	Nous et ils sont sur des parcours différents. ✗

S-V agreement, past participles

Are the agreement marks of the flagged participles the correct ones? (Past participle placed after auxiliary AVOIR agrees with verb object iff object precedes auxiliary. Otherwise participle is in masculine singular form).

S5a	Source	The woman who saw a mouse in the corridor is charming.
	Ref	La femme qui a vu une souris dans le couloir est charmante.
	PBMT-1	La femme qui a vu une souris dans le couloir est charmante. ✓
	NMT	La femme qui a vu une souris dans le couloir est charmante. ✓
	Google	La femme qui a vu une souris dans le couloir est charmante. ✓
S5b	Source	The woman that your brother saw in the corridor is charming.
	Ref	La femme que votre frère a vue dans le couloir est charmante.
	PBMT-1	La femme que ton frère a vu dans le couloir est charmante. ✗
	NMT	La femme que votre frère a vu dans le couloir est charmante. ✗
	Google	La femme que votre frère a vue dans le couloir est charmante. ✓
S5c	Source	The house that John has visited is crumbling.
	Ref	La maison que John a visitée tombe en ruines.
	PBMT-1	La maison que John a visité est en train de s'écrouler. ✗
	NMT	La maison que John a visitée est en train de s'effondrer. ✓
	Google	La maison que John a visité est en ruine. ✗
S5d	Source	John sold the car that he had won in a lottery.
	Ref	John a vendu la voiture qu'il avait gagnée dans une loterie.
	PBMT-1	John a vendu la voiture qu'il avait gagné à la loterie. ✗
	NMT	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓
	Google	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓

Subjunctive mood

Is the flagged verb in the correct mood? (Certain triggering verbs, adjectives or subordinate conjunctions, induce the subjunctive mood in the subordinate clause that they govern).

S6a	Source	He will come provided that you come too.
	Ref	Il viendra à condition que vous veniez aussi.
	PBMT-1	Il viendra à condition que vous venez aussi. ✗
	NMT	Il viendra lui aussi que vous le faites. ✗
	Google	Il viendra à condition que vous venez aussi. ✗
S6b	Source	It is unfortunate that he is not coming either.
	Ref	Il est malheureux qu'il ne viene pas non plus.
	PBMT-1	Il est regrettable qu'il n'est pas non plus à venir. ✗
	NMT	Il est regrettable qu'il ne soit pas non plus. ✗
	Google	Il est malheureux qu'il ne vienne pas non plus. ✓

S6c	Source	I requested that families not be separated.
	Ref	J'ai demandé que les familles ne soient pas séparées.
	PBMT-1	J'ai demandé que les familles ne soient pas séparées. ✓
	NMT	J'ai demandé que les familles ne soient pas séparées. ✓
	Google	J'ai demandé que les familles ne soient pas séparées. ✓

Lexico-Syntactic

Argument switch

Are the experiencer and the object of the “missing” situation correctly preserved in the French translation? (Argument switch).

S7a	Source	Mary sorely misses Jim .
	Ref	Jim manque cruellement à Mary .
	PBMT-1	Marie manque cruellement de Jim. ✗
	NMT	Mary a lamentablement manqué de Jim. ✗
	Google	Mary manque cruellement à Jim. ✗
S7b	Source	My sister is really missing New York .
	Ref	New York manque beaucoup à ma sœur .
	PBMT-1	Ma sœur est vraiment absent de New York. ✗
	NMT	Ma sœur est vraiment manquante à New York. ✗
	Google	Ma sœur manque vraiment New York. ✗
S7c	Source	What he misses most is his dog .
	Ref	Ce qui lui manque le plus, c'est son chien .
	PBMT-1	Ce qu'il manque le plus, c'est son chien. ✗
	NMT	Ce qu'il manque le plus, c'est son chien. ✗
	Google	Ce qu'il manque le plus, c'est son chien. ✗

Double-object verbs

Are “gift” and “recipient” arguments correctly rendered in French? (English double-object constructions)

S8a	Source	John gave his wonderful wife a nice present.
	Ref	John a donné un beau présent à sa merveilleuse épouse .
	PBMT-1	John a donné sa merveilleuse femme un beau cadeau. ✗
	NMT	John a donné à sa merveilleuse femme un beau cadeau. ✓
	Google	John a donné à son épouse merveilleuse un présent gentil. ✓
S8b	Source	John told the kids a nice story.
	Ref	John a raconté une belle histoire aux enfants .
	PBMT-1	John a dit aux enfants une belle histoire. ✓
	NMT	John a dit aux enfants une belle histoire. ✓
	Google	John a raconté aux enfants une belle histoire. ✓
S8c	Source	John sent his mother a nice postcard.
	Ref	John a envoyé une belle carte postale à sa mère .
	PBMT-1	John a envoyé sa mère une carte postale de nice. ✗
	NMT	John a envoyé sa mère une carte postale de nice. ✗
	Google	John envoya à sa mère une belle carte postale. ✓

Fail to

Is the meaning of “fail to” correctly rendered in the French translation?

S9a	Source	John failed to see the relevance of this point.
	Ref	John n’a pas vu la pertinence de ce point.
	PBMT-1	John a omis de voir la pertinence de ce point. ✗
	NMT	John n’a pas vu la pertinence de ce point. ✓
	Google	John a omis de voir la pertinence de ce point. ✗
S9b	Source	He failed to respond.
	Ref	Il n’a pas répondu .
	PBMT-1	Il n’a pas réussi à répondre. ✓
	NMT	Il n’a pas répondu. ✓
	Google	Il n’a pas répondu. ✓
S9c	Source	Those who fail to comply with this requirement will be penalized.
	Ref	Ceux qui ne se conforment pas à cette exigence seront pénalisés.
	PBMT-1	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	NMT	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	Google	Ceux qui ne respectent pas cette exigence seront pénalisés. ✓

Manner-of-movement verbs

Is the movement action expressed in the English source correctly rendered in French? (Manner-of-movement verbs with path argument may need to be rephrased in French).

S10a	Source	John would like to swim across the river.
	Ref	John aimerait traverser la rivière à la nage .
	PBMT-1	John aimerait nager dans la rivière. ✗
	NMT	John aimerait nager à travers la rivière. ✗
	Google	John aimerait nager à travers la rivière. ✗
S10b	Source	They ran into the room.
	Ref	Ils sont entrés dans la chambre à la course .
	PBMT-1	Ils ont couru dans la chambre. ✗
	NMT	Ils ont couru dans la pièce. ✗
	Google	Ils coururent dans la pièce. ✗
S10c	Source	The man ran out of the park.
	Ref	L’homme est sorti du parc en courant .
	PBMT-1	L’homme a manqué du parc. ✗
	NMT	L’homme s’enfuit du parc. ✗
	Google	L’homme sortit du parc. ✗

Hard example featuring spontaneous noun-to-verb derivation (“nonce verb”).

S10d	Source	John guitared his way to San Francisco.
	Ref	John s’est rendu jusqu’à San Francisco en jouant de la guitare .
	PBMT-1	John guitared son chemin à San Francisco. ✗
	NMT	John guitared sa route à San Francisco. ✗
	Google	John a guité son chemin à San Francisco. ✗

Overlapping subcat frames

Is the French verb for “know” correctly chosen? (Choice between “savoir”/“connaître” depends on syntactic nature of its object)

S11a	Source	Paul knows that this is a fact.
	Ref	Paul sait que c’est un fait.
	PBMT-1	Paul sait que c’est un fait. ✓
	NMT	Paul sait que c’est un fait. ✓
	Google	Paul sait que c’est un fait. ✓
S11b	Source	Paul knows this story.
	Ref	Paul connaît cette histoire.
	PBMT-1	Paul connaît cette histoire. ✓
	NMT	Paul connaît cette histoire. ✓
	Google	Paul connaît cette histoire. ✓
S11c	Source	Paul knows this story is hard to believe.
	Ref	Paul sait que cette histoire est difficile à croire.
	PBMT-1	Paul connaît cette histoire est difficile à croire. ✗
	NMT	Paul sait que cette histoire est difficile à croire. ✓
	Google	Paul sait que cette histoire est difficile à croire. ✓
S11d	Source	He knows my sister will not take it.
	Ref	Il sait que ma soeur ne le prendra pas.
	PBMT-1	Il sait que ma soeur ne prendra pas. ✓
	NMT	Il sait que ma soeur ne le prendra pas. ✓
	Google	Il sait que ma soeur ne le prendra pas. ✓
S11e	Source	My sister knows your son is reliable.
	Ref	Ma sœur sait que votre fils est fiable.
	PBMT-1	Ma soeur connaît votre fils est fiable. ✗
	NMT	Ma sœur sait que votre fils est fiable. ✓
	Google	Ma sœur sait que votre fils est fiable. ✓

NP to VP

Is the English “NP to VP” complement correctly rendered in the French translation? (Sometimes one needs to translate this structure as a finite clause).

S12a	Source	John believes Bill to be dishonest .
	Ref	John croit que Bill est malhonnête .
	PBMT-1	John estime que le projet de loi soit malhonnête. ✓
	NMT	John croit que le projet de loi est malhonnête. ✓
	Google	John croit que Bill est malhonnête. ✓
S12b	Source	He liked his father to tell him stories .
	Ref	Il aimait que son père lui raconte des histoires .
	PBMT-1	Il aimait son père pour lui raconter des histoires. ✗
	NMT	Il aimait son père pour lui raconter des histoires. ✗
	Google	Il aimait son père à lui raconter des histoires. ✗

S12c	Source	She wanted her mother to let her go .
	Ref	Elle voulait que sa mère la laisse partir .
	PBMT-1	Elle voulait que sa mère de lui laisser aller. ✗
	NMT	Elle voulait que sa mère la laisse faire. ✓
	Google	Elle voulait que sa mère la laisse partir. ✓

Factitives

Is the English verb correctly rendered in the French translation? (Agentive use of some French verbs require embedding under “faire”).

S13a	Source	John cooked a big chicken.
	Ref	John a fait cuire un gros poulet.
	PBMT-1	John cuit un gros poulet. ✗
	NMT	John cuit un gros poulet. ✗
	Google	John a fait cuire un gros poulet. ✓
S13b	Source	John melted a lot of ice.
	Ref	John a fait fondre beaucoup de glace.
	PBMT-1	John fondu a lot of ice. ✗
	NMT	John a fondu beaucoup de glace. ✗
	Google	John a fondu beaucoup de glace. ✗
S13c	Source	She likes to grow flowers.
	Ref	Elle aime faire pousser des fleurs.
	PBMT-1	Elle aime à se développer des fleurs. ✗
	NMT	Elle aime à cultiver des fleurs. ✓
	Google	Elle aime faire pousser des fleurs. ✓

Noun Compounds

Is the English nominal compound rendered with the right preposition in the French translation?

S14a	Source	Use the meat knife.
	Ref	Utilisez le couteau à viande.
	PBMT-1	Utilisez le couteau de viande. ✗
	NMT	Utilisez le couteau à viande. ✓
	Google	Utilisez le couteau à viande. ✓
S14b	Source	Use the butter knife.
	Ref	Utilisez le couteau à beurre.
	PBMT-1	Utilisez le couteau à beurre. ✓
	NMT	Utilisez le couteau au beurre. ✗
	Google	Utilisez le couteau à beurre. ✓
S14c	Source	Use the steak knife.
	Ref	Utilisez le couteau à steak.
	PBMT-1	Utilisez le steak couteau. ✗
	NMT	Utilisez le couteau à steak. ✓
	Google	Utilisez le couteau de steak. ✗

S14d	Source	Clean the water filter.
	Ref	Nettoyez le filtre à eau.
	PBMT-1	Nettoyez le filtre à eau. ✓
	NMT	Nettoyez le filtre à eau. ✓
	Google	Nettoyez le filtre à eau. ✓
S14e	Source	Clean the juice filter.
	Ref	Nettoyez le filtre à jus.
	PBMT-1	Nettoyez le filtre de jus. ✗
	NMT	Nettoyez le filtre de jus. ✗
	Google	Nettoyez le filtre à jus. ✓
S14f	Source	Clean the tea filter.
	Ref	Nettoyez le filtre à thé.
	PBMT-1	Nettoyez le filtre à thé. ✓
	NMT	Nettoyez le filtre de thé. ✗
	Google	Nettoyez le filtre à thé. ✓
S14g	Source	Clean the cloth filter.
	Ref	Nettoyez le filtre en tissu.
	PBMT-1	Nettoyez le filtre en tissu. ✓
	NMT	Nettoyez le filtre en tissu. ✓
	Google	Nettoyez le filtre en tissu. ✓
S14h	Source	Clean the metal filter.
	Ref	Nettoyez le filtre en métal.
	PBMT-1	Nettoyez le filtre en métal. ✓
	NMT	Nettoyez le filtre en métal. ✓
	Google	Nettoyez le filtre métallique. ✓
S14i	Source	Clean the paper filter.
	Ref	Nettoyez le filtre en papier.
	PBMT-1	Nettoyez le filtre en papier. ✓
	NMT	Nettoyez le filtre en papier. ✓
	Google	Nettoyez le filtre à papier. ✗

Common idioms

Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?

S15a	Source	Stop beating around the bush .
	Ref	Cessez de tourner autour du pot .
	PBMT-1	Cesser de battre la campagne. ✗
	NMT	Arrêtez de battre autour de la brousse. ✗
	Google	Arrêter de tourner autour du pot. ✓

S15b	Source	You are putting the cart before the horse .
	Ref	Vous mettez la charrue devant les bœufs .
	PBMT-1	Vous pouvez mettre la charrue avant les bœufs. ✓
	NMT	Vous mettez la charrue avant le cheval. ✗
	Google	Vous mettez le chariot devant le cheval. ✗
S15c	Source	His comment proved to be the straw that broke the camel's back .
	Ref	Son commentaire s'est avéré être la goutte d'eau qui a fait déborder le vase .
	PBMT-1	Son commentaire s'est révélé être la goutte d'eau qui fait déborder le vase. ✓
	NMT	Son commentaire s'est avéré être la paille qui a brisé le dos du chameau. ✗
	Google	Son commentaire s'est avéré être la paille qui a cassé le dos du chameau. ✗
S15d	Source	His argument really hit the nail on the head .
	Ref	Son argument a vraiment fait mouche .
	PBMT-1	Son argument a vraiment mis le doigt dessus. ✓
	NMT	Son argument a vraiment frappé le clou sur la tête. ✗
	Google	Son argument a vraiment frappé le clou sur la tête. ✗
S15e	Source	It's no use crying over spilt milk .
	Ref	Ce qui est fait est fait .
	PBMT-1	Ce n'est pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait haché. ✗
	Google	Ce qui est fait est fait. ✓
S15f	Source	It is no use crying over spilt milk .
	Ref	Ce qui est fait est fait .
	PBMT-1	Il ne suffit pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait écrémé. ✗
	Google	Il est inutile de pleurer sur le lait répandu. ✗

Syntactically flexible idioms

Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?

S16a	Source	The cart has been put before the horse.
	Ref	La charrue a été mise devant les bœufs .
	PBMT-1	On met la charrue devant le cheval. ✗
	NMT	Le chariot a été mis avant le cheval. ✗
	Google	Le chariot a été mis devant le cheval. ✗
S16b	Source	With this argument, the nail has been hit on the head .
	Ref	Avec cet argument, la cause est entendue .
	PBMT-1	Avec cette argument, l'ongle a été frappée à la tête. ✗
	NMT	Avec cet argument, l'ongle a été touché à la tête. ✗
	Google	Avec cet argument, le clou a été frappé sur la tête. ✗

Syntactic

Yes-no question syntax

Is the English question correctly rendered as a French question?		
S17a	Source	Have the kids ever watched that movie?
	Ref	Les enfants ont-ils déjà vu ce film?
	PBMT-1	Les enfants jamais regardé ce film? ✗
	NMT	Les enfants ont-ils déjà regardé ce film? ✓
	Google	Les enfants ont-ils déjà regardé ce film? ✓
S17b	Source	Hasn't your boss denied you a promotion?
	Ref	Votre patron ne vous a-t-il pas refusé une promotion?
	PBMT-1	N'a pas nié votre patron vous un promotion? ✗
	NMT	Est-ce que votre patron vous a refusé une promotion? ✓
	Google	Votre patron ne vous a-t-il pas refusé une promotion? ✓
S17c	Source	Shouldn't I attend this meeting?
	Ref	Ne devrais-je pas assister à cette réunion?
	PBMT-1	Ne devrais-je pas assister à cette réunion? ✓
	NMT	Est-ce que je ne devrais pas assister à cette réunion? ✓
	Google	Ne devrais-je pas assister à cette réunion? ✓

Tag questions

Is the English "tag question" element correctly rendered in the translation?		
S18a	Source	Mary looked really happy tonight, didn't she ?
	Ref	Mary avait l'air vraiment heureuse ce soir, n'est-ce pas ?
	PBMT-1	Marie a regardé vraiment heureux de ce soir, n'est-ce pas elle? ✗
	NMT	Mary s'est montrée vraiment heureuse ce soir, ne l'a pas fait? ✗
	Google	Mary avait l'air vraiment heureuse ce soir, n'est-ce pas? ✓
S18b	Source	We should not do that again, should we ?
	Ref	Nous ne devrions pas refaire cela, n'est-ce pas ?
	PBMT-1	Nous ne devrions pas faire qu'une fois encore, faut-il? ✗
	NMT	Nous ne devrions pas le faire encore, si nous? ✗
	Google	Nous ne devrions pas recommencer, n'est-ce pas? ✓
S18c	Source	She was perfect tonight, was she not ?
	Ref	Elle était parfaite ce soir, n'est-ce pas ?
	PBMT-1	Elle était parfait ce soir, elle n'était pas? ✗
	NMT	Elle était parfaite ce soir, n'était-elle pas? ✗
	Google	Elle était parfaite ce soir, n'est-ce pas? ✓

WH-MVT and stranded preps

Is the dangling preposition of the English sentence correctly placed in the French translation?		
S19a	Source	The guy that she is going out with is handsome.
	Ref	Le type avec qui elle sort est beau.
	PBMT-1	Le mec qu'elle va sortir avec est beau. ✗
	NMT	Le mec qu'elle sort avec est beau. ✗
	Google	Le mec avec qui elle sort est beau. ✓

S19b	Source	Whom is she going out with these days?
	Ref	Avec qui sort-elle ces jours-ci?
	PBMT-1	Qu'est-ce qu'elle allait sortir avec ces jours? ✗
	NMT	À qui s'adresse ces jours-ci? ✗
	Google	Avec qui sort-elle de nos jours? ✓
S19c	Source	The girl that he has been talking about is smart.
	Ref	La fille dont il a parlé est brillante.
	PBMT-1	La jeune fille qu'il a parlé est intelligent. ✗
	NMT	La fille qu'il a parlé est intelligente. ✗
	Google	La fille dont il a parlé est intelligente. ✓
S19d	Source	Who was he talking to when you left?
	Ref	À qui parlait-il au moment où tu es parti?
	PBMT-1	Qui est lui parler quand vous avez quitté? ✗
	NMT	Qui a-t-il parlé à quand vous avez quitté? ✗
	Google	Avec qui il parlait quand vous êtes parti? ✓
S19e	Source	The city that he is arriving from is dangerous.
	Ref	La ville d'où il arrive est dangereuse.
	PBMT-1	La ville qu'il est arrivé de est dangereuse. ✗
	NMT	La ville qu'il est en train d'arriver est dangereuse. ✗
	Google	La ville d'où il vient est dangereuse. ✓
S19f	Source	Where is he arriving from ?
	Ref	D'où arrive-t-il?
	PBMT-1	Où est-il arrivé? ✗
	NMT	De quoi s'agit-il? ✗
	Google	D'où vient-il? ✓

Adverb-triggered inversion

Is the adverb-triggered subject-verb inversion in the English sentence correctly rendered in the French translation?

S20a	Source	Rarely did the dog run.
	Ref	Rarement le chien courait-il .
	PBMT-1	Rarement le chien courir. ✗
	NMT	Il est rare que le chien marche. ✗
	Google	Rarement le chien courir. ✗
S20b	Source	Never before had she been so unhappy.
	Ref	Jamais encore n'avait-elle été aussi malheureuse.
	PBMT-1	Jamais auparavant, si elle avait été si malheureux. ✗
	NMT	Jamais auparavant n'avait été si malheureuse. ✗
	Google	Jamais elle n'avait été aussi malheureuse. ✓

S20c	Source	Nowhere were the birds so colorful.
	Ref	Nulle part les oiseaux n'étaient si colorés.
	PBMT-1	Nulle part les oiseaux de façon colorée. ✗
	NMT	Les oiseaux ne sont pas si colorés. ✗
	Google	Nulle part les oiseaux étaient si colorés. ✗

Middle voice

Is the generic statement made in the English sentence correctly and naturally rendered in the French translation?

S21a	Source	Soup is eaten with a large spoon.
	Ref	La soupe se mange avec une grande cuillère
	PBMT-1	La soupe est mangée avec une grande cuillère. ✗
	NMT	La soupe est consommée avec une grosse cuillère. ✗
	Google	La soupe est consommée avec une grande cuillère. ✗
S21b	Source	Masonry is cut using a diamond blade.
	Ref	La maçonnerie se coupe avec une lame à diamant.
	PBMT-1	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
	NMT	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
	Google	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
S21c	Source	Champagne is drunk in a glass called a flute.
	Ref	Le champagne se boit dans un verre appelé flûte.
	PBMT-1	Le champagne est ivre dans un verre appelé une flûte. ✗
	NMT	Le champagne est ivre dans un verre appelé flûte. ✗
	Google	Le Champagne est bu dans un verre appelé flûte. ✗

Fronted "should"

Fronted "should" is interpreted as a conditional subordinator. It is normally translated as "si" with imperfect tense.

S22a	Source	Should Paul leave, I would be sad.
	Ref	Si Paul devait s'en aller, je serais triste.
	PBMT-1	Si le congé de Paul, je serais triste. ✗
	NMT	Si Paul quitte, je serais triste. ✗
	Google	Si Paul s'en allait, je serais triste. ✓
S22b	Source	Should he become president, she would be promoted immediately.
	Ref	S'il devait devenir président, elle recevrait immédiatement une promotion.
	PBMT-1	S'il devait devenir président, elle serait encouragée immédiatement. ✓
	NMT	S'il devait devenir président, elle serait immédiatement promue. ✓
	Google	Devrait-il devenir président, elle serait immédiatement promue. ✗
S22c	Source	Should he fall, he would get up again immediately.
	Ref	S' il venait à tomber, il se relèverait immédiatement.
	PBMT-1	S'il devait tomber, il allait se lever immédiatement de nouveau. ✓
	NMT	S'il tombe, il serait de nouveau immédiatement. ✗
	Google	S'il tombe, il se lèvera immédiatement. ✗

Clitic pronouns

Are the English pronouns correctly rendered in the French translations?

S23a	Source	She had a lot of money but he did not have any .
	Ref	Elle avait beaucoup d'argent mais il n' en avait pas.
	PBMT-1	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
	NMT	Elle avait beaucoup d'argent, mais il n'a pas eu d'argent. ✓
	Google	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
S23b	Source	He did not talk to them very often.
	Ref	Il ne leur parlait pas très souvent.
	PBMT-1	Il n'a pas leur parler très souvent. ✗
	NMT	Il ne leur a pas parlé très souvent. ✓
	Google	Il ne leur parlait pas très souvent. ✓
S23c	Source	The men are watching each other .
	Ref	Les hommes se surveillent l'un l'autre
	PBMT-1	Les hommes se regardent les uns les autres. ✓
	NMT	Les hommes se regardent les uns les autres. ✓
	Google	Les hommes se regardent. ✗
S23d	Source	He gave it to the man.
	Ref	Il le donna à l'homme.
	PBMT-1	Il a donné à l'homme. ✗
	NMT	Il l'a donné à l'homme. ✓
	Google	Il le donna à l'homme. ✓
S23e	Source	He did not give it to her .
	Ref	Il ne le lui a pas donné.
	PBMT-1	Il ne lui donner. ✗
	NMT	Il ne l'a pas donné à elle. ✗
	Google	Il ne lui a pas donné. ✗

Ordinal placement

Is the relative order of the ordinals and numerals correct in the French translation?

S24a	Source	The first four men were exhausted.
	Ref	Les quatre premiers hommes étaient tous épuisés.
	PBMT-1	Les quatre premiers hommes étaient épuisés. ✓
	NMT	Les quatre premiers hommes ont été épuisés. ✓
	Google	Les quatre premiers hommes étaient épuisés. ✓
S24b	Source	The last three candidates were eliminated.
	Ref	Les trois derniers candidats ont été éliminés.
	PBMT-1	Les trois derniers candidats ont été éliminés. ✓
	NMT	Les trois derniers candidats ont été éliminés. ✓
	Google	Les trois derniers candidats ont été éliminés. ✓

S24c	Source	The other two guys left without paying.
	Ref	Les deux autres types sont partis sans payer.
	PBMT-1	Les deux autres mecs ont laissé sans payer. ✓
	NMT	Les deux autres gars à gauche sans payer. ✓
	Google	Les deux autres gars sont partis sans payer. ✓

Inalienable possession

Is the French translation correct and natural both in: a) its use of a particular determiner on the body part noun; and b) the presence or absence of a reflexive pronoun before the verb?

S25a	Source	He washed his hands.
	Ref	Il s' est lavé les mains.
	PBMT-1	Il se lavait les mains. ✓
	NMT	Il a lavé ses mains. ✗
	Google	Il se lava les mains. ✓
S25b	Source	I brushed my teeth.
	Ref	Je me suis brossé les dents.
	PBMT-1	J'ai brossé mes dents. ✗
	NMT	J'ai brossé mes dents. ✗
	Google	Je me suis brossé les dents. ✓
S25c	Source	You brushed your teeth.
	Ref	Tu t' es brossé les dents
	PBMT-1	Vous avez brossé vos dents. ✗
	NMT	vous avez brossé vos dents. ✗
	Google	Tu as brossé les dents. ✗
S25d	Source	I raised my hand.
	Ref	J'ai levé la main.
	PBMT-1	J'ai levé la main. ✓
	NMT	J'ai soulevé ma main. ✗
	Google	Je levai la main. ✓
S25e	Source	He turned his head.
	Ref	Il a tourné la tête.
	PBMT-1	Il a transformé sa tête. ✗
	NMT	Il a tourné sa tête. ✗
	Google	Il tourna la tête. ✓
S25f	Source	He raised his eyes to heaven.
	Ref	Il leva les yeux au ciel.
	PBMT-1	Il a évoqué les yeux au ciel. ✓
	NMT	Il a levé les yeux sur le ciel. ✓
	Google	Il leva les yeux au ciel. ✓

Zero REL PRO

Is the English zero relative pronoun correctly translated as a non-zero one in the French translation?

S26a	Source	The strangers the woman saw were working.
	Ref	Les inconnus que la femme vit travaillaient.
	PBMT-1	Les étrangers la femme vit travaillaient. ✗
	NMT	Les inconnus de la femme ont travaillé. ✗
	Google	Les étrangers que la femme vit travaillaient. ✓
S26b	Source	The man your sister hates is evil.
	Ref	L'homme que votre sœur déteste est méchant.
	PBMT-1	L'homme ta soeur hait est le mal. ✗
	NMT	L'homme que ta soeur est le mal est le mal. ✓
	Google	L'homme que votre sœur hait est méchant. ✓
S26c	Source	The girl my friend was talking about is gone.
	Ref	La fille dont mon ami parlait est partie.
	PBMT-1	La jeune fille mon ami a parlé a disparu. ✗
	NMT	La petite fille de mon ami était révolue. ✗
	Google	La fille dont mon ami parlait est partie. ✓

A Challenge Set for French \rightarrow English Machine Translation

Pierre Isabelle and Roland Kuhn
National Research Council Canada
first.last@nrc-cnrc.gc.ca

June 18, 2018

Abstract

We present a *challenge set* for French \rightarrow English machine translation based on the approach introduced in [1]. Such challenge sets are made up of sentences that are expected to be relatively difficult for machines to translate correctly because their most straightforward translations tend to be linguistically divergent. We present here a set of 506 manually constructed French sentences, 307 of which are targeted to the same kinds of structural divergences as in the paper mentioned above. The remaining 199 sentences are designed to test the ability of the systems to correctly translate difficult grammatical words such as prepositions. We report on the results of using this challenge set for testing two different systems, namely Google Translate and DEEPL, each on two different dates (October 2017 and January 2018). All the resulting data are made publicly available.

1 Introduction

Isabelle, Cherry and Foster[1] introduce a *challenge set* approach to evaluating machine translation (MT) systems. This approach is not meant as a substitute for traditional evaluation methods such as average BLEU or human scores on a held out portion of the training corpus. It is rather meant to supplement these methods with tools that directly measure the extent to which MT systems manage to tackle some of the more difficult translation problems. Thus, unlike traditional metrics, challenge sets provide developers with a fine-grained view of the remaining obstacles.

Ideally, one would like challenge sets to be constructed automatically. This is all the more desirable in that such sets are intrinsically language-pair dependent. But until automatic construction methods become available, we can turn to human experts for developing limited sets of challenging sentences. This is what [1] did for English \rightarrow French machine translation (MT). However, challenge sets are not only language-pair dependent: they are also direction dependent. For example, in English \rightarrow French translation there is a need to choose between the French verbs *savoir* and *connaître* as the correct translation for the English verb *to know*. As it turns out, this depends on the syntactic nature of the complement of the verb. But in the opposite direction this problem does not arise: both *savoir* and *connaître* simply translate as *to know*. This kind of asymmetry led us to develop a new challenge set that specifically targets French \rightarrow English MT.

In section 2, we describe the makeup of our new challenge set. In section 3 we report on the results of subjecting both Google Translate and DEEPL to the resulting challenge on two different dates: 5 October 2017 and 16 January 2018. As we will see, this constitutes an interesting way to track the systems' evolution.

2 Makeup of the New Challenge Set

In developing our French \rightarrow English challenge set we closely followed the practices described in [1]. In particular:

- We used short sentences that are each meant to bring into focus a single linguistic issue.
- All sentences are based on "common general vocabulary" because our goal is not to test lexical coverage but rather the system's ability to bridge specific linguistic divergences.
- We provide one reference translation for each sentence, notwithstanding the fact that other acceptable translations are usually possible.

- Each sentence is accompanied with a yes-no question that focuses the attention of evaluators on the particular issue that the sentence is intended to test.
- The evaluator’s responses to each yes/no question completely determine the outcome of the evaluation or the relevant sentence. Consequently, translation errors that lie outside the scope of the yes-no questions will be ignored as irrelevant.
- We use the same major classes of structural divergences as in the paper mentioned above: morpho-syntactic, lexico-syntactic and purely syntactic divergences. Each class is further subdivided into a set of subclasses which has a large overlap with those presented in the same paper.

One important difference with the earlier paper is that, in addition to testing the system’s ability to deal with structural divergences, this new dataset includes some 199 examples that are intended to probe the systems’ ability to cope with the notoriously difficult translation of grammatical words. This is achieved using groups of examples in which the same French grammatical word needs to be rendered in different ways in the English translation.

In the remainder of this section we describe and illustrate the classes of linguistic difficulties that are built into our challenge set.

2.1 Morpho-Syntactic divergences

We use the term *morpho-syntactic divergences* to refer to cases where the two languages differ in which grammatical features are overtly marked in the morphology of corresponding words. Whenever a target language word requires a feature marking that is not explicitly marked in its source, the MT system needs to infer the relevant feature from the context. Our challenge set is probing that capability for the following cases:

- **Proclitic pronouns.** French complement pronouns often need to be procliticized, that is, moved to the left of the verb and phonetically attached to it. When translated into English, these pronouns need to be repositioned in their normal complement position. Moreover, the French clitic frequently underdetermines the grammatical features of the complement, such its gender, person and number and case. In the following example, the French clitic *se* is not marked for gender but it needs to be translated as a neutral reflexive pronoun *itself* because it co-refers with the neutral noun *machine*.

Cette machine *se* répare elle-même. → This machine repairs *itself*.

The following two examples illustrate that the French clitic *leur* underdetermines the case/preposition marking of the corresponding English complement:

Je *leur* ai parlé. → I spoke *to them*.

Je *leur* ai emprunté un livre. → I borrowed a book *from them*.

- **Chez soi.** The French form *chez moi/toi/lui...* can translate as *at home* but only when it is being used reflexively:

Mon fils est demeuré *chez lui*. → My son stayed {*at his place* | *at home*}.

Ma fille est demeurée *chez lui*. → My daughter stayed {*at his place* | **at home*}. ¹

- **Verb tense.** The overt French verb tense marking frequently underdetermines its English counterpart. In the following example, the French verb form is compatible with both the indicative and subjunctive mood while its English counterpart is explicitly marked as subjunctive. As a result the MT system must be able to determine whether or not the context is triggering the subjunctive mood:

Il est essentiel qu’il *arrive* à temps. → It is essential that he *arrive* on time.

¹We use the asterisk to mark a translation as incorrect.

- **Verb tense concordance.** In sentences expressing two events with a specific time dependency, French and English often feature different tense concordance constraints. In the following example both verbs are semantically future but English, contrary to French, requires the subordinate verb to be in the grammatical present tense:

Max *partira* dès que tu te *lèveras*. → Max *will leave* as soon as you { **will get up* | *get up* }.

2.2 Lexico-syntactic divergences

We now turn to *lexico-syntactic divergences*. We place under that heading all cases where the corresponding governing words of the two languages happen to organize their respective dependents in different ways. As a result, whenever such a governing word gets translated the system must be able to reorganize its dependents accordingly.

- **Argument switch.** In some cases, the most straightforward translation of a given verb requires a change in the order of the verb's arguments. This is the case when the French verb *manquer à* is translated as *to miss*:

Mary *manque* beaucoup à John. → John *misses Mary* a lot.

- **Manner of movement verbs.** In English, a completed movement is often expressed by using a verb that expresses a manner of moving (*walk, climb, swim, etc.*) and combining it with a prepositional phrase that expresses the endpoint of the movement. In French, this is normally expressed by a more generic movement verb together with an adverbial expressing the manner of that movement. Here are two examples:

John *aimerait traverser* l'océan à la *nage*. → John would like to *swim across* the ocean.

John *entra dans* la salle *en courant*. → John *ran into* the room.

- **Verb/adverb transposition.** Some French verbs tend to be expressed as adverbs in English. This involves a reorganization of the sentence in which a verb which is subordinate in French becomes the main verb in English.

Max a *fini par comprendre* la difficulté. → Max *finally understood* the difficulty.

- **Non-finite to finite clause.** It is quite common for a non-finite clause of French to be translated as a finite clause in English. This raises the difficulty of introducing an adequate subject as well as an adequate verb tense in the English translation.

Max *croit connaître* la vérité. → Max *thinks he knows* the truth.

Mary *croyait connaître* la vérité. → Mary *thought she knew* the truth.

Aussitôt *son travail terminé*, Mary *partit*. → As soon as *her work was over*, Mary left.

- **"fact" insertion.**

Cela *provient de ce que* Max a trop dormi. → That *arises from the fact that* Max has slept too much.

- **"how" insertion.**

Max *sait réparer* une cafetière. → Max *knows how to repair* a coffee maker.

- **Middle voice.** The so-called middle-voice of French involves a pseudo-pronominal verb form whose interpretation is related to that of a passive sentence, often with a generic interpretation. It is most often translated in English with a passive form.

Ce type de moteur *se répare* facilement. → This type of engine *can be repaired* easily.

- **Control verbs.** So called *subject-control verbs* take an infinitival complement whose subject is understood to co-refer with the subject of the control verb. In contrast, the understood subject of *object-control verbs* is the object of the control verb. This difference can be brought to light when the infinitival complement is reflexive.

Max a convaincu sa fille de *se* sacrifier. → Max convinced his daughter to sacrifice {**himself* | *herself*}.

Max a promis à sa fille de ne pas *se* sacrifier. → Max promised his daughter not to sacrifice {*himself* | **herself*}.

- **Mass versus count nouns.** Both languages make a grammatical distinction between count nouns (e.g. *book, table, idea*) versus mass nouns (e.g. *wine, butter, fear*). However, there are cases where a French noun and its English counterpart happen to fall on different sides of the divide. In such cases, a partitive noun may need to be introduced or deleted in the translation.

Max lui a donné *un conseil*. → Max gave him {**an advice* | *a piece of advice*}.

- **Factitives.** Some French verbs require the use of the auxiliary verb "faire" in order to receive an agentive reading. In such cases, that auxiliary must disappear in the translation.

Max a {**fondu* | *fait fondre*} la glace. → Max melted the ice.

Max a {**explosé* | *fait exploser*} un rocher. → Max blew up a rock.

- **Two-position adjectives.** The correct translation of several French adjectives depends on whether they are placed before or after the noun they modify.

Une idée *simple* n'est pas forcément mauvaise. → A {*simple* | **mere*} idea is not necessarily bad.

La *simple* idée de partir la terrorisait. → The {**simple* | *mere*} idea of leaving terrorized her.

- **Genitive.** Contrary to French, the preferred way to express genitives in English is not to use a prepositional phrase but rather the case marking 's.

Il a pris le livre *de mon frère aîné*. → He took *my elder brother's* book.

2.3 Purely syntactic divergences

The third type of divergence considered stems from the fact that some syntactic constructions only exist in one or the other language. Whenever a French sentence contains a construction that has no direct counterpart in English, the MT system needs to be able to recast the source language material into a different construction.

In fact, we have already seen one such case above, namely that of French proclitics. While we listed them under the heading of morpho-syntactic divergences, they do exemplify both types of divergence at once. Since there are no proclitics in English, a French object proclitic needs to be relocated in the standard post-verbal position in the English translation: *Il la voit*. → *He sees her*. Here are some other subtypes of purely syntactic divergences.

- **Yes-no question syntax.** French and English differ in the way yes-no questions are formed. Basically, French questions are obtained as follows: if the subject is a proclitic, move it after the verb; otherwise insert a particle (either *est-ce que* at the beginning of the sentence or *-il* after the verb). In contrast, English questions are obtained by fronting an auxiliary verb.

As-tu lu ce livre? → *Have you read this book?*

Max partira-*t-il* à temps? → *Will* Max leave on time?

- **Tag questions.** The so-called *tag-question* construction of English does not exist in French, but the French *n'est-ce pas?* sentence-final question is normally translated as an appropriate tag question, which involves selecting the right auxiliary verb.

Il a vu la photo hier, *n'est-ce pas?* → He saw the picture yesterday, *didn't he?*

Nous devrions vérifier le niveau d'huile, *n'est-ce pas?* → We should check the oil level, *shouldn't we?*

- **WH-movement: relative clauses.** When a relative clause is formed, its internal relativized element gets fronted, typically in the form of a "WH-word". For example, in *The man whom you saw is my brother* the word "whom" is understood to refer to the object of the verb "saw", which we will call its *native site*. French and English relative constructions are often parallel enough that an MT system can get away with a superficial process that falls short of explicitly relating the WH word to its native site. However, such a superficial approach breaks down in the case of **stranded prepositions**. In French, whenever a prepositional phrase is relativized, its preposition must be fronted alongside the WH-word: *la fille avec qui tu as dansé*. In contrast, English will often leave the preposition stranded: *the girl you danced with*. Note that in the French → English direction, the MT system does not have to move the preposition to its native site, since preposition fronting is also permitted in English. However, if the system does move the preposition to its correct native site, then this provides nice evidence that it is able to perform some deeper processing.

L'homme à qui Max a donné un livre est parti. → The man *whom* Max gave a book *to* is gone.

La fille dont il a parlé est brillante. → The girl *that* he talked *about* is brilliant.

- **WH-movement: interrogatives.** Question formation and relative formation are highly parallel in both French and English. As a result, stranded prepositions raise the same translation issues with questions as with relatives.

À qui Max a-t-il donné un livre? → *Whom* did Max gave a book *to*?

Pour quelle compagnie travaille-t-il? → *What* company does he work *for*?

- **Negation.** In French, negation is typically expressed using a discontinuous form such as *ne ... pas/jamais/plus/nullement* while in English this is typically done using a single word. MT systems often run into difficulty with this phenomenon. In our first example below the system needs to recognize that *ne* is being used in an "expletive" (i.e. non-negative) way and therefore should not be translated. In our second example, the French negation is to be rendered by the single negation word *not*, but while reinforcing it with the intensifying adverbial *at all*.

Je crains que Max *ne* vienne nous voir. → I'm afraid Max is coming to see us.

Max *ne* comprend *nullement* cette idée. → Max does *not* understand this idea *at all*.

- **Double negation.** Double negations are sometimes used for stylistical effect and some MT systems appear to have difficulty coping with that.

Ce politicien n'est pas capable de ne pas mentir. → This politician is not able not to lie.

C'est le docteur dont il est *impossible* que vous n'ayez *pas* entendu parler. → It is the doctor of whom it is *impossible* that you have *not* heard.

- **Other doubled concepts.** Some MT systems appear to experience some difficulty with sentences that contain two tokens for the same concept.

Il a commis *faute* sur *faute*. → He committed *mistake* after *mistake*.

C'est *beaucoup beaucoup* mieux. → This is *much much* better.

2.4 Purely lexical divergences.

The kinds of structural divergences described above closely mirror what was done in [1] for English→French machine translation. However, in that work idiomatic phrases and support verbs were placed under the broader category of lexico-syntactic divergences. In the present work, we instead introduce an additional top-level category, namely that of purely lexical divergences. Alongside testing material for idioms and support verbs, this category will include a substantial amount of additional material meant to test the ability of MT systems to translate common grammatical words such as prepositions.

- **Common idioms – fixed.** Some phrases need to be translated as a group because they happen to have a language-specific idiomatic meaning. The simpler case is that of fixed idioms, those that always appear under one and the same form.

Ils sont déterminés à continuer *envers et contre tous*. → They are determined to continue *in spite of all opposition*.

Ils partiront *entre chien et loup*. → They will leave *at dusk*.

- **Common idioms – variable.** Many idioms exhibit some morphological and/or syntactic flexibility. As a result, there is a need for MT systems to generalize over a range of different surface forms.

Cessez de *tourner autour du pot*. → Stop *beating around the bush*.

Il *tournait* constamment *autour du pot*. → He was constantly *beating around the bush*.

Vous *mettez la charrue devant les boeufs*. → You *put the cart before the horse*.

La *charrue* a été *mise avant les bœufs*. → The *cart* was *put before the horse*.

- **Support verbs.** These verbs (also known as "light verbs") carry little meaning in themselves. Rather they combine with their complement to express what can often be expressed as a single verb. For example, *to walk* and *to take a walk* are roughly equivalent. But even though the support verb - here, *take* - carries little meaning in itself, its choice is not free. In this example, **make a walk* is not an acceptable substitute. Support verbs must be translated as a whole with their complements.

Max a *fait campagne contre* le maire hier. → Max *campaigned against* the mayor yesterday.

Ceci *apporte la preuve qu'il était au courant*. → This *is proof that* he was aware.

Unacceptable, literal translations for these two examples would be:

Max **made a campaign against* the mayor yesterday.

This **brings proof that* he was aware.

- **Grammatical words.** Grammatical words such as prepositions are notoriously difficult to translate. Our challenge set includes testing material for some 28 different grammatical words or phrases that are relatively difficult to translate correctly because they each have multiple uses. For each one we provide sets of sentences where the word needs to receive different translations as a result of these different uses. Consider for example some different uses/translations of the French preposition *en*:

Il lui a offert un foulard *en* soie. → He offered her a silk scarf.

Il est docteur *en* philosophie. → He's a doctor *of* philosophy.

En semaine, je travaille. → *On* weekdays, I work.

J'ai payé mes études *en* vendant du café. → I paid my tuition *by* selling coffee.

En travaillant, j'aime écouter de la musique. → *While* working, I like to listen to music.

Another good example is the multiple uses and translations of the preposition *par*:

Il a été averti *par* Paul. → He was warned *by* Paul.

Un lundi *par* mois, il se rend au marché. → One Monday *per* month, he goes to the market.

Il a fait cela *par* plaisir. → He did it *for* pleasure.

Il a fait cela *par* habitude. → He did it *out of* habit.

Le bateau a coulé *par* cent mètres de fond. → The boat sank *to* a depth of a hundred meters.

2.5 Our New Challenge Set.

We manually developed a set of 506 different challenging examples populating the main categories discussed above with the distribution shown in Table 1.

In addition to making use of our own personal experience in machine translation, we were able to draw many examples from Morris Salkoff's highly detailed and insightful French-English contrative grammar [2].

3 Testing Google Translate and DEEPL

Armed with this new French → English challenge set, we decided to evaluate the performance of the Google Translate and DEEPL neural machine translation systems. We submitted all 506 sentences to each system on two different dates: 5 October 2017 and 16 January 2018. We collected the results and proceeded to evaluate them.

The evaluation protocol was as follows. The human evaluator looks at each test case in turn, being provided with: a) the source-language sentence; b) one reference translation; c) the machine-translated sentence to be evaluated; and d) a single yes/no question about the translation and its relationship to the source-language sentence. The evaluator simply provides an answer the yes/no question associated with each translated example. Figure 1 provides two examples of material being presented to the evaluator together with his/her response (either "Yes" or "No").

The first author made an initial pass at responding to each one of the 2024 relevant questions (506 for each one of the four machine translations). The second author checked all these judgments and noted all disagreements. Each difference was then discussed by the two authors and a joint decision was made.

Thus, unlike in [1] where three independent evaluators were used, the results presented below only rely on the authors' judgments. However, we are making these judgments available alongside the new challenge set so that interested parties can compare them with their own judgments.

The main results are presented in Table 2. The outcome of October 2017 was similar to that presented in [1] for the English→French direction: in both cases DEEPL turned out to deal with the challenge set quite a bit better than Google. On the present challenge set, DEEPL's overall

Category	No. of examples	Percent
Morpho-syntactic	43	8.5%
Lexico-syntactic	79	15.6%
Purely syntactic	84	16.6%
Purely lexical	300	59.3%
Total	506	100%

Table 1: Distribution of challenge set examples across main categories.

Src	La femme s'est regardée dans le miroir.
Ref	The woman looked at herself in the mirror.
Sys	The woman looked at herself in the mirror.
Is the French highlighted pronoun correctly translated (y/n)? Yes	

Src	Je le suppose.
Ref	I suppose so.
Sys	I'm guessing.
Is the French highlighted pronoun correctly translated (y/n)? No	

Figure 1: Example challenge set questions.

rate of success was almost 13% higher than that of Google. This advantage holds in all categories of examples except for morpho-syntax where both systems are tied.

We can also see that the overall performance of both systems turned out to be somewhat better in January 2017. The Google system achieved an overall improvement of 2.6% for a relative error reduction of 3.6%, while the DEEPL system got a 1% improvement for a relative error reduction of 1.3%. The rate of progress varied across categories. In the case of morpho-syntax, Google managed to gain 7%, significantly bettering DEEPL which turned out to lose 4.6%. Conversely, in the case of pure syntax Google lost 3.5% while DEEPL's performance remained unchanged. For the other two categories (purely lexical and lexico-syntactic) both systems progressed but Google did so more markedly.

Table 3 provides a breakdown of the same results in terms of our finer-grained subcategories.

3.1 Conclusion

We have presented a new challenge set for evaluating machine translation systems in the French→English direction based on the principles presented in [1]. This new set includes 506 different sentences spread across four categories: morpho-syntactic, lexico-syntactic, purely syntactic and purely lexical. The first three categories mirror those of [1] but the last one is novel. Each sentence is meant to test the ability of MT systems to bridge one specific divergence issue between the two languages.

Our 506 challenge sentences have been submitted to the Google and DEEPL MT systems on two different dates: 5 October 2017 and 16 January 2018. The results have been evaluated according to the method presented in [1], which amounts to responding to the yes-no questions attached to each challenge sentence.

In this case the evaluators were the co-authors of this paper, which is not optimal. However, we are making all the data available so that readers can compare our judgments with theirs.

References

- [1] Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach for evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, September 2017. Association for Computational Linguistics.
- [2] Morris Salkoff. *A French-English grammar: contrastive grammar on translation principles*, volume 1. John Benjamins, Amsterdam, 1999.

Divergence type	GNMT-1	GNMT-2	DEEPL-1	DEEPL-2
Morpho-syntactic	76.7%	83.7%	76.7%	72.1%
Lexico-syntactic	63.3%	63.3%	75.9%	78.5%
Purely Syntactic	70.2%	66.7%	79.8%	79.8%
Purely Lexical	64.8%	66.3%	80.9%	81.4%
Overall	63.6%	66.2%	76.5%	77.5%

Table 2: Challenge set success rate for the systems under study, with "-1" and "-2" indicating respectively the versions of September 2017 and January 2018.

Category	Subcategory	#	GNMT-1	GNMT-2	DEEPL-1	DEEPL-2
M-Syn	Proclitic pronouns	21	81.0%	90.5%	95.2%	85.7%
	"Chez soi"	4	50.0%	50.0%	50.0%	50.0%
	Verb tense	11	81.8%	81.8%	72.7%	72.7%
	Verb tense concord	7	71.4%	85.7%	42.9%	42.9%
L-Syn	Argument switch	5	40%	20%	80%	100%
	Manner-of-movement verbs	6	33.3%	33.3%	50%	50%
	Verb-adverb transposition	3	66.7%	66.7%	100%	66.7%
	Non-finite \rightarrow finite clause	20	80%	80%	75%	80%
	"De/à ce que" \rightarrow "from the fact that"	2	0%	50%	50%	50%
	V1 V2 _{inf} \rightarrow V1 how to V2 _{inf}	2	100%	100%	100%	100%
	Middle voice	3	66.7%	100%	66.7%	100%
	Control verbs	5	40%	60%	60%	60%
	Count Vs mass nouns	6	83.3%	83.3%	66.7%	66.7%
	"Voilà [TIME] que"	3	66.7%	0%	100%	100%
	Factitives	13	38.5%	38.5%	76.9%	69.2%
	Two-position adjectives	9	88.9%	88.9%	88.9%	100%
	Genitives	2	100%	100%	100%	100%
Syn	Yes-no question syntax	4	50%	50%	100%	100%
	Tag questions	4	0%	50%	100%	100%
	WH movement, relatives	19	78.9%	68.4%	78.9%	78.9%
	WH movement, questions	10	80%	70%	90%	90%
	Negation	8	75%	87.5%	100%	100%
	Double negation	20	55%	50%	65%	65%
	Other doubled concepts	5	80%	40%	20%	20%
	Inalienable possession	6	100%	100%	100%	100%
	Subject inversion	8	87.5%	87.5%	87.5%	87.5%
Lex	Common idioms – fixed	25	32%	40%	52%	48%
	Common idioms – variable	24	33.3%	58.3%	66.7%	66.7%
	Support verbs	52	67.3%	71.2%	71.2%	80.8%
	Translation of grammatical words	199	64.8%	66.3%	80.9%	81.4%

Table 3: Summary of scores by fine-grained categories. “#” reports number of questions in each category, while the reported score is the percentage of questions for which the divergence was correctly bridged.

AI for Translation Quality

BtB-NRC Collaboration Agreement

04 October 2018

Context

It is now recognized that recent advances in Artificial Intelligence (AI) have enabled substantial improvements in Machine Translation (MT), making it possible for an organization such as the Translation Bureau (BtB) to consider using MT in a systematic fashion, to improve its internal processes. In fact, recent studies suggest that current MT has the potential not only to improve translator productivity, but also help translators produce better quality translations. It is in this context that the Translation Bureau is currently modernising its processes, gradually incorporating MT into its workflow.

However, it is also well-known that indiscriminate use of MT can have a catastrophic impact on translation quality. Thus, in an environment where AI is systematically used, effective Quality Control becomes a critical part of the operation. Interestingly, AI techniques can themselves play a central role in these quality control tasks.

This collaboration agreement between the Translation Bureau and the National Research Council centers around the idea of Using *Artificial intelligence for Translation Quality*. It involves not only the Bureau's Strategic Reengineering team and the NRC's Multilingual Text Processing team, but also the Université de Montréal's RALI laboratory. It is all about building AI applications to assist humans in controlling and optimizing translation quality. The work packages proposed below are organized around specific applications, that can be broadly separated in three groups: those application that focus on the quality of Machine Translation, those that focus on the quality of Human Translations, and those that aim at exploiting user feedback to improve translation processes.

Work Packages

Machine Translation Quality (MTQ)

Package MTQ1 - MT Quality Evaluation [NRC]

For an organization that crucially depends on Machine Translation for productivity, it is critical to continuously monitor the quality of its MT systems. This can be performed automatically, by means of *automatic MT quality evaluation metrics*. These metrics typically take as input a source text, its translation as produced by an MT system, and a reference human-quality

translation. By measuring the similarity between the MT and the reference translation, they effectively measure the quality of the MT. Over the past 15 years, dozens of different metrics have been proposed by MT researchers; the best known and most widely used is certainly BLEU. But the limitations of BLEU, especially when used with Neural machine translation (NMT) output, have been a sticking point since NMT research started to ramp up in the past couple of years, and many researchers and experts are actively advocating, looking for, and suggesting alternatives. NRC is at the forefront of this new research with the YiSi family of MTQ evaluation metrics, which rely on neural network-based models of word meanings.

This package is about deploying AI-based MTQ evaluation metrics at the Bureau, that produce reliable measures of translation quality, and improving these metrics to measure quality at fine-level resolutions (e.g. evaluating individual segments or phrases), while taking into account information from a wider context (e.g. paragraph or document).

Inputs

- System API [NRC & BtB]
- access to Megacorporus [BtB]
- MT outputs [NRC or BtB]
- Manual quality annotations of MT [BtB]

Deliverables:

- MTQ1.1: MTQ evaluation metric system
- MTQ1.2: Report on fine-resolution and contextually aware MTQ evaluation

Package MTQ2 - MT Quality Estimation [NRC]

MT Quality Estimation is a method used to automatically provide a quality indication for MT output without depending on human reference translations. In more simple terms, it tries to determine how good or bad the translations produced by an MT system, without human intervention. MTQ estimation is typically used to flag or block potentially problematic MT output, or to identify output for which the level of confidence is so high as to not require human intervention. It is akin to the “fuzziness level” in translation memory (TM): a segment with a “100%” level can be passed through with high confidence, one with 70% will require some degree of editing, and segments with scores lower than 70% may not even be worth looking at. MTQ Estimation has been an area of active research for over 15 years, but recent advances in neural network and AI technology are now making it possible to contemplate systems that reliably estimate quality at finer levels of resolution, better correlate with human judgement, take contextual information into consideration, have a unified view of TM and MT output and actively learn from user feedback.

The objective of this package is to develop and test an advanced MTQ estimation component, to be integrated into the Bureau’s document analysis framework.

Inputs

- System API [NRC & BtB]
- access to Megacorporus [BtB]
- MT and TM outputs [NRC & BtB]
- Translator interaction feedback [BtB]

Deliverables

- MTQ2.1: Preliminary report on advanced MTQ estimation
- MTQ2.2: MTQ Estimation system prototype
- MTQ2.3: Final report on advanced MTQ estimation

Human Translation Quality (HTQ)

Package HTQ1 - Computer-assisted HTQ Control [RALI]

With increased volumes and tight deadlines, not all translation handled by the Bureau goes through a formal revision process. In some situations, only sample-based quality evaluation can be performed. As a result, unrevised translations returned to clients sometimes contain incorrectly translated segments. Equally problematic is the fact that these incorrect segments also make their way into the *Megacorporus* (MC), the Bureau's corporate translation memory. There, they combine with *alignment* errors, owing to the segment alignment mechanism at the heart of the Megacorporus's automatic feeding system also occasionally misaligning segments. When similar documents are later submitted to the Bureau for translation, these misaligned and incorrectly translated segments can resurface, causing loss of productivity and further propagation of errors.

The focus of this package is on developing tools to assist with the quality control of human translations, and optimize the quality of the material in the Megacorporus. Recent research on parallel sentence detection in comparable and noisy parallel corpora has led to several innovative AI methods for modeling translation equivalence. The main objective of this package is to adapt these methods to the problem of identifying non-equivalent pairs of text segments, and from there develop an automatic translation error detection system. This system could then be used to assist in "cleaning-up" the Megacorporus and controlling the quality of new translations, by prioritizing the visual examination of sections of text that display atypical features: poor cross-lingual semantic similarity, major omissions and insertions, equivalence problems with numerical expressions (dates, amounts, etc.), named entities (names of people, organizations and places), terminological consistency issues, etc.

Because segment alignment is a critical part of the Megacorporus feeding mechanism, and possibly an important source of errors, special attention will be devoted within this package to examining the extent and causes of alignment errors in the Megacorporus, and proposing more advanced solutions.

Inputs

- System API [NRC/RALI & BtB]
- access to Megacorpus [BtB]
- Samples of problematic misaligned Megacorpus documents [BtB]
- Human annotations of Megacorpus errors [BtB]

Deliverables:

- HTQ1.1: Report on alignment errors in the Megacorpus
- HTQ1.2: Report on Automatic HT error detection
- HTQ1.2: Automatic HT error detection prototype

Package HTQ2 - Translation Readability Assessment [NRC]

Readability is the ease with which a reader can understand a written text. The readability of a text depends on its content (the complexity of its vocabulary, syntax and the topic it covers) and its presentation (such as typographic aspects like font size, line height, and line length). Higher readability eases reading effort and speed for any reader, but it is especially important for those who do not have high reading comprehension. Raising the readability level of a text from mediocre to good can make the difference between success and failure of its communication goals.

The level of readability of documents is an increasingly important factor in the Bureau's quality assessment of translations: it is a recurrent complain from clients that Bureau translations have a lower readability (are more difficult to understand) than their source language counterparts. Various methods have been proposed by researchers to automatically evaluate the readability of texts, e.g. the Flesch-Kincaid readability tests, and some of these are even available in word-processing environments such as MS Word. However, most of these metrics are targeted specifically for English, and are not directly applicable for French. Some metrics have been developed for French, but the measurements are not directly comparable across languages. Furthermore, all of these metrics are primarily designed to measure the readability of whole documents, and their reliability breaks down when they are used at finer levels of resolution (paragraph, sentences).

The main objective of this package is to develop an AI-based Translation Readability system, capable of measuring document-level readability for English and French, and reporting its results on a comparable scale. This package would also explore the feasibility of reliably detecting readability problems at finer resolutions, e.g. paragraphs or segments, thus assisting translators in producing text better adapted to the needs of their target audience.

Inputs

- System API [NRC & BtB]
- access to Megacorpus [BtB]
- Document- and segment-level readability annotations on Megacorpus [BtB]

Deliverables

- HTQ2.1: Document-level crosslingual Translation Readability system
- HTQ2.2: Report on automatic fine-grained crosslingual readability assessment
- HTQ2.3: fine-resolution crosslingual Translation Readability prototype

Human Translator Feedback (HF)

Package HF1 - Advanced Translation Memory Matching [RALI]

When a new document is submitted to the Bureau, its content is systematically analysed and compared to the Megacorpus, and similar segments are retrieved and proposed for re-use by translators. Promoting thematic, or contextual consistency in pairs of segments retrieved from the Megacorpus (MC) is likely to pay off, according to a preliminary study conducted at RALI (Gotti et al. 2006). This package is about studying which level of metadata/information can be used to improve the usefulness of the segments returned. The problem can be cast as finding a policy with which to query the MC, optimizing upon various types of information :

- document-level metadata : translator, client, thematic (as computed for instance by topic models)
- sentence-pair level information : frequency of a pair in the MC, date of last use
- document-level consistency : enforcing the use of similar term translations within a document

The ultimate objective is to produce MatchMC, an advanced translation memory matching system specifically adapted to the context of the Bureau.

Inputs

- System API [NRC/RALI & BtB]
- Access to Megacorpus (including metadata) [BtB]
- BtB matching algorithm description or sample IO [BtB]

Deliverables

- HF1.1: Report on Advanced TM Matching
- HF1.2: Advanced TM Matching system prototype

Package HF2 - Incremental Machine Translation Training [NRC]

In a computer-assisted translation setting such as considered by the Bureau, source text documents are first passed through translation memory and machine translation, to produce a *pre-translated* version. This version is then presented to the translator, who edits the result to produce his/her translation. In some settings, edited segments are fed back to the translation memory in real time, as they are produced by the translator, making them instantly available for reuse, either for similar segments further in the same document, or even to other translators working on related documents or sub-documents of the same project. A similar process may be applied to machine translation: by feeding back corrected MT (or TM) output to

the MT system, it is possible to effectively allow the system to learn from its errors, and gradually correct its behavior. In the long run, this minimizes the need for the translator to repeatedly correct the same mistakes.

Incremental training for MT is an active area of research, in which the main questions are 1) how to efficiently incorporate new information in MT systems? and 2) how to balance the contribution of this new (dynamic) information with the existing (static) knowledge within the system? The objective of this package is to experiment various incremental training MT strategies, and determine optimal settings to optimize MT quality.

This project presupposes a translation process that already incorporates MT, such as the translation of weather warnings for Environment Canada ("METEO").

Inputs

- System API [NRC & BtB]
- access to METEO system & data [BtB]
- Translator interaction feedback [BtB]

Deliverables

- HF2.1: Evaluation protocol for Incremental MT
- HF2.2: Incremental MT Training prototype (@METEO)
- HF2.3: Report on Incremental MT Training

Package HF3 - Adaptive NMT Training [RALI]

Studying the complementarity of a translation memory and neural machine translation is the focus of this package. The first scenario that will be tested is to train one neural MT system (making use of an existing package, such as OpenNMT or Sockeye) on the cleaned-up version of the Megacorpus (see package HTQ1 above). This system can be queried to provide translation of sentences unseen in the memory. Detectors implemented in package might be used for silencing bad translations, or simply marking translations that likely require particular attention. A number of challenges will have to be solved for an efficient integration, among which :

- the use of the Megacorpus (MC) structuration for designing a translation system aware of the domain of the text to translate
- the detection of wrong parts of a produced translations, in particular hallucinated ones (neural MT is known to produce translations that are fluent but sometimes not faithful to the source text)

A report measuring how neural MT can play a role and delineate these roles. In particular, a figure we want to report is the percentage of source sentences in the MC that can be translated into a translation sanctioned by the MC.

Inputs

- System API [NRC/RALI & BtB]
- access to Megacorporus [BtB]
- Access to existing/future MT used at BtB [BtB]

Deliverables

- HF3.1: NMT System Prototype
- HF3.2: Report on NMT Integration at the Bureau

Package HF4 - Assisted Translation Assignment [NRC]

Translators develop specialties and preferences. In many cases, the best guarantee of quality is to assign the work to the right person. Coincidentally, it is also often the best guarantee that the work will be done quickly. Just because a translator (or group of translators) is usually assigned work from a given client doesn't mean that this translator is the best fit for all documents or projects of that client. Each of the Bureau's clients (typically, government departments) produces a wide variety of documents, and although a majority of them may be of the same domain or genre, some may be of a very different nature. At the same time, some documents are very similar in genre or domain across many clients. Assigning the right document to the right translator requires a broad knowledge of who is best at doing what. In a large organization such as the Bureau, this knowledge may be distributed among many individuals. It is also knowledge that changes over time, as the Bureau's workforce and the requirements of its clients evolve.

The objective of this package is to develop an AI-based *translation assignment assistant*: a system that receives as input a source-language document or set of documents, and produces as output an ordered list of potential translators who are best suited to do the work. This list would primarily be computed based on a comparison between the linguistic profile of the document(s) and that of documents previously translated by individual translators. But we could also explore the potential if incorporating document metadata (client, author, document type, urgency, etc.), as well as quality evaluations of previously translated documents. These evaluations could be explicit (quality assessment performed by the Bureau, client feedback) or implicit (amount/nature of revision required after initial translation).

Inputs

- System API [NRC & BtB]
- Access to Megacorporus, including metadata (client, translator, etc.) [BtB]
- Access to quality assessment, revision history of Megacorporus documents [BtB]

Deliverables

- HF4.1: Preliminary report on Automating translation assignment
- HF4.2: Prototype Translation assignment assistant
- HF4.3: Final report on Automating translation assignment

Calendar

2018				2019				2020			
Apr.	Jul	Oct	Dec	Apr	Jul	Oct	Dec	Apr	Jul	Oct	Dec
HF2 (meteo)											



AI for Translation Quality

A Collaboration Project between the **BtB**, the **NRC** and the **RALI** laboratory
(Université de Montréal)



Context

- Recent advances in AI have enabled substantial improvements in MT, making it possible for an organization such as the BtB to consider using MT in a systematic fashion
- In an environment where MT is used systematically, effective Quality Control becomes a critical part of the operation
→ AI can help with this too



Project

Artificial intelligence for Translation Quality

Goal: Build AI applications to assist humans in controlling and optimizing translation quality.

Partners:

- BtB Strategic Reengineering team
- NRC's Multilingual Text Processing
- Université de Montréal's RALI laboratory.



Themes

Theme 1:	Theme 2:	Theme 3:
Machine Translation Quality (MTQ)	Human Translation Quality (HTQ)	Human Feedback (HF)
Using AI to measure and predict quality of MT output	Assist translators and language professionals in detecting quality issues in human translations	Optimize quality in the translation process by exploiting user feedback



MTQ1: Machine Translation Quality Evaluation

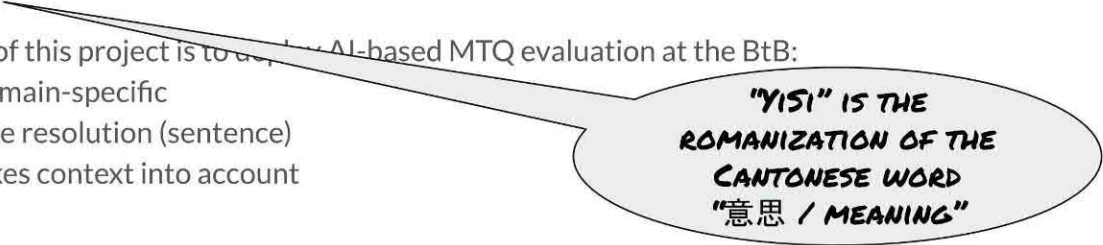
An organization that crucially depends on MT for productivity must continually monitor the quality of its MT systems.

Existing MT quality metrics such as BLEU have known limitations, especially with Neural MT that make them unreliable

NRC's **YiSi** family of MTQ evaluation metrics relies on neural network-based models of word meaning

The goal of this project is to develop AI-based MTQ evaluation at the BtB:

- Domain-specific
- Fine resolution (sentence)
- Takes context into account



**"YISI" IS THE
ROMANIZATION OF THE
CANTONESE WORD
"意思 / MEANING"**



MTQ2: Machine Translation Quality Estimation

MTQ *estimation* is about producing quality indications for MT output without depending on human reference translations. It can be used to flag or block bad quality MT, or to prioritize the work of translators.

The goal of this package is to develop and test an advanced MTQ estimation component, to be integrated into the BtB's document analysis framework.

Based on the YiSi technology, our aim is to create estimation methods that

- reliably estimate quality at fine levels of resolution
- better correlate with human judgement
- take contextual information into consideration
- have a unified view of TM and MT output
- actively learn from user feedback



HTQ1: Computer-assisted Translation Quality Control

Not all translation handled by the BtB goes through a formal revision process. As a result, unrevised translations containing incorrectly translated segments are sometimes returned to clients and archived into the Megacorporus (MC), the BtB's corporate translation memory, where they combine with alignment and segmentation errors, potentially causing loss of productivity and further propagation of errors.

The goal of this project is to develop tools to assist with the quality control of human translations, and optimize the quality of the material in the MC.

Recent research on parallel sentence detection in comparable and noisy parallel corpora has led to innovative AI methods for modeling translation equivalence. Our aim is to adapt these methods to the problem of identifying non-equivalent pairs of text segments, and from there develop an automatic translation error detection system.



HTQ2: Translation Readability Assessment

Readability is the ease with which a reader can understand a written text. The level of readability of documents is an increasingly important factor in the BtB's quality assessment of translations: it is a recurrent complain from clients that BtB translations have a lower readability (are more difficult to understand) than their source language counterparts.

Existing automatic readability metrics, such as Flesch-Kincaid, are targeted specifically to English, and are not directly applicable to French or other languages.

The main objective of this package is to develop AI-based Translation Readability metrics, capable of measuring document-level readability for English and French, and reporting its results on a comparable scale. We will also explore the feasibility of reliably detecting readability problems at finer resolutions, e.g. paragraphs or segments, thus assisting translators in producing text better adapted to the needs of their target audience.



HF1: Advanced Translation Memory Matching

When a new document is submitted to the BtB, its content is systematically analysed and compared to the MC, and similar segments are retrieved and proposed for re-use by translators. Currently, this retrieval is based on textual (surface) similarity only.

According to a preliminary study conducted at RALI (Gotti et al. 2006), it would be beneficial to promote thematic and/or contextual consistency in pairs of segments retrieved from the Megacorpus (MC)

Our goal is to produce an advanced TM matching system adapted to the context of the BtB. We will:

- Establish which metadata/information can be used to improve the usefulness of the segments returned.
- Determine a policy with which to query the MC, optimizing upon :
 - document-level metadata : translator, client, thematic (as computed for instance by topic models)
 - sentence-pair level information : frequency of a pair in the MC, date of last use
 - document-level consistency : enforcing the use of similar term translations within a document



HF2: Incremental Training for Machine Translation

Weather warnings from Environment Canada are translated by BtB translators, with the help of NRC's Portage MT technology. Within this context, we study the impact of **incremental training**, a process by which revised versions of MT output ('post-edits') are channeled back to the MT system to improve its behavior in real-time.

Our objective is to experiment various incremental training MT strategies, and determine optimal settings to optimize MT quality. More specifically, we examine:

- how to balance the contribution of this new (dynamic) information with the existing (static) knowledge within the system
- how to structure the data so as to maximize positive impacts on various types of weather warnings.



HF3: Adaptive Neural Machine Translation

In this package, we study the complementarity of a translation memory and neural machine translation, and integration with packages MTQ2 and HF1. Starting from neural MT systems (OpenNMT or Sockeye) trained with the Megacorporus, we explore:

- The integration of detectors for silencing bad translations, or simply marking translations that likely require particular attention. (See MTQ2) Special attention will be devoted to 'hallucinated' translations (neural MT is known to produce translations that are fluent but sometimes not faithful to the source text)
- the use of the Megacorporus (MC) structuration for designing a translation system aware of the domain of the text to translate (see HF1)



HF4: Assisted Translation Assignment

Translators develop **specialties** and **preferences**. Often, the best guarantee of quality is to assign the translation to the right person.

The objective of this package is to develop an AI-based **translation assignment assistant**: a system that receives as input a source-language document, and outputs a list of translators who are best suited to do the work, by comparing

- the linguistic profile of the document with those of documents previously translated by a translator
- document metadata (client, author, document type, urgency, etc.),
- quality evaluations of previously translated documents, either explicit (quality assessment performed by the Bureau, client feedback) or implicit (amount/nature of revision required after initial translation).



Calendar

FY 2018-19

Apr. Jul Oct Jan

HF2 (meteo)

FY 2019-20

Apr Jul Oct Jan

MTQ1 (YiSi)

HTQ1 (CleanMC)

HF4 (Assignment Assistant)

MTQ2 (QEstim)

HF3 (NMT)

FY 2020-21

Apr Jul Oct Jan

HF1 (MatchMC)

HTQ2 (Readability)

MT for Parliamentary Debates: Portage vs. Sockeye

With some observations about YiSi

Michel Simard

September 2019

This is a short report on our comparative evaluation of Portage and Sockeye systems trained on Hansard data. Overall, our human evaluation suggests that Sockeye does better in the English-to-French direction, while Portage is better at French-to-English. This result is in line with automatic evaluations with BLEU and WER. When measuring performance with YiSi however, Portage appears to come out as the best system, regardless of the language direction. Further analysis suggests that this is due to YiSi computing document-level scores as a simple average of sentence-level scores. Modifying YiSi to use a weighted average instead, taking sentence lengths into account, appears to lead to better agreement with human evaluations. When comparing systems, we also argue that *sentence-level preference rates*, denoting for what proportion of the test translations a given system obtains the best score, are more informative than absolute scores.

Introduction

During the summer of 2019, Samuel and Marc trained Portage and Sockeye systems on our Parliamentary data. Our automatic evaluation of these systems, using BLEU, WER and YiSi, produced conflicting results: BLEU and WER preferred Sockeye for English-to-French translation, and Portage for French-to-English. But YiSi systematically preferred Portage, regardless of the direction. To sort things out, we proceeded to do a QAFF-style human evaluation, in which human judges are asked their preferences about individual translations on a small sample, in a blind setting. This evaluation confirmed the verdict of BLEU and WER, which lead us to examine in more details why YiSi diverged with the other metrics.

This document reports on the results of this investigation. In Section 1, we present the main results of the automatic and human evaluations of the MT systems. In Section 2, we examine why the YiSi metric diverges from human judgement and other automatic

metrics. And in Section 3, we discuss using *per-sentence metric preferences* as an alternative, possibly more informative way of reporting the results of automatic evaluations when comparing systems.

1. Main results

The Portage and Sockeye systems were evaluated using a test set of 4863 sentences of recent Hansard parliamentary debates from the House of Commons. All metrics were computed on truecase texts, tokenized with the Portage tokenizer. BLEU and WER metrics were computed using the Portage programs `bleumain` and `wermain`. YiSi scores are actually YiSi-1, using `word2vec` monolingual embeddings trained on 100k randomly sampled documents from the BtB corpus: about 50k documents in each language, 59M English words for the English embeddings, 67M French words of French embeddings.

The results are presented in Table 1. As can be seen, for French-to-English (FR-EN) translation, all metrics indicate a preference for the Portage translations (recall that, for BLEU and YiSi, bigger is better, while for WER, smaller is better). For English-to-French (EN-FR) translation however, BLEU and WER suggest that Sockeye produces better translations, while YiSi still prefers the Portage translations, albeit by a small margin.

System:	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
BLEU	37.9	40.5	39.8	38.7
WER	53.0	51.5	49.8	51.5
YiSi	70.3	70.0	74.4	70.0

Table 1: BLEU, WER and YiSi-1 scores on Hansard test set. Best results for each system and metric are in bold.

To better assess the relative qualities of each system, we decided to perform a human evaluation of the translations. For this, we took inspiration from the QUAFF project from a few years back. In this sort of evaluation, each human judge receives a set of translations to compare, and some guidelines. In this evaluation, the translations were presented in an Excel sheet (one sheet per judge). The guidelines were sent by email, and went something like this (but in French, as all judges spoke French):

Each line in this file has a source sentence, two machine translations (mt1 and mt2) and a reference translation. Between columns mt1 and mt2, there's a column Pref: write down "1" or "2" in that column, depending which one you prefer.

- *There's no formal definition for what it means to "prefer": follow your intuition*
- *Read the source, often it's worth the effort!*
- *If both translations look equivalent or of comparable quality, write down "1" (this is the same as picking one at random)*
- *Some translations may look identical (I tried not to, but there seems to be a bug). Same thing: write down "1"*
- *There's a hidden column (col A) which gives away the order of translations. Please don't look at it, at least until you're done comparing!*
- *Even without looking at column A, it's sometimes easy to figure out which system produced which translation. Try to remain neutral :-)*

The sentences for that evaluation were sampled from the same test set of recent Hansard debates used for the automatic evaluation. As suggested in the guidelines, sampling was done so as to focus on sentences for which Portage and Sockeye produce different translations (that's about 90% of all sentences). But we also restricted the sampling to sentences between 8 and 24 words long (about 50% of all sentences). This was done in order to avoid very short sentences, which are often difficult to evaluate for lack of context, as well as long sentences, which are often tedious to compare.

Volunteers were recruited within the team to participate in the evaluation. Each volunteer received an Excel sheet with 100 pairs of translations to compare. In the end, annotations from seven volunteer judges were collected: three for FR-EN, four for EN-FR (Thanks to all who participated!). Figure 1 shows an example evaluation sheet such as those that annotators were working with.

Results of that evaluation are presented in Table 2. "Global preference" is just the percentage of test sentences for which judges preferred a given system. "Weighted preference" also takes into account the size of each example, measured as the number of word tokens in the source-language sentence.

In summary, annotators globally agreed with the BLEU and WER verdict: they preferred Sockeye for EN-FR translations, and Portage for FR-EN. Individual annotator preferences were mostly aligned with the global verdicts, except for Annotator 4, who preferred Portage translations for EN-FR translations. It should be emphasized, however, that all annotators were working on different subsamples of the testset, so individual assessments can not be directly compared. Also note that EN-FR and FR-EN annotators are not the same (EN-FR Annotator 1 is not the same person as FR-EN Annotator 1).

As specified in the guidelines, annotators were not given the option to label translations as "equivalent". This was done on purpose: in a previous evaluation, when an *equivalent* option was available, even though annotators were explicitly instructed to use it only as a last resort, some annotators ended up using that label for the majority of examples.¹ By

¹It should be noted that, in this previous study, the annotators were all professional translators. When asked to comment on his/her relatively high proportion of *equivalent* labels, one annotator just said: "Well, I just felt that, in the majority of cases, both machine translations were bad!"

	source	mt1	pref	mt2	ref
1	We talk about a national housing strategy, and they say we need to have more houses.	Nous parlons d'une stratégie nationale en matière de logement, et ils disent que nous avons besoin d'un plus grand nombre de maisons.		Nous parlons d'une stratégie nationale du logement, et ils disent que nous devons avoir plus de maisons.	Nous avons parlé d'une stratégie nationale sur le logement et ils ont réagi en disant qu'il fallait davantage de maisons.
2	It is important that we educate people who come to this country about those issues.	Il est important de sensibiliser les gens qui viennent au Canada au sujet de ces questions.	1	Il est important de sensibiliser les gens qui viennent au Canada à ces questions.	Il est important d'éduquer les gens qui viennent au pays à propos de ces questions.
3	The Government of Canada has been taking action through our healthy eating strategy.	Le gouvernement du Canada a pris des mesures dans le cadre de notre stratégie de saine alimentation.	2	Le gouvernement du Canada a pris des mesures dans le cadre de notre stratégie de saines habitudes alimentaires.	Le gouvernement du Canada a déjà pris des mesures dans le cadre de la Stratégie en matière de saine alimentation.
4	Therefore, we would not just take the department's word for it.	Par conséquent, nous ne nous contenterions pas de croire le ministère sur parole.	1	Par conséquent, nous n'aurions pas simplement prendre le ministère sur parole.	Par conséquent, nous n'aurions plus eu à accepter la décision du ministère sans poser de question.
5	Hon. David Lametti : Mr. Speaker, the hon. member's question is a good one.	L'hon. David Lametti : Monsieur le Président, le député pose une bonne question.		L'hon. David Lametti : Monsieur le Président, la question de la députée est excellente.	L'hon. David Lametti : Monsieur le Président, le député pose une bonne question.
6	It actually called for meaningful action, like taking our foot off the accelerator.	Il a réclamé des mesures concrètes, comme prendre notre pied à pied de l'accélérateur.		En fait, il a demandé des mesures concrètes, comme prendre le pied de l'accélérateur.	Notre motion exigeait en fait des mesures substantielles, notamment pour réduire la vitesse des changements.
7	I looked at some of the points	J'ai examiné certains des points qui ont été faites sur la réduction		J'ai examiné certains des points qui ont été soulevés au sujet de	Je me suis penchée sur certains des arguments présentés concernant

Figure 1: Example annotation spreadsheet.

forcing annotators to make a choice, we hope to minimize this tendency. We still provide annotators with an option for situations where they feel they can't decide (*"If both translations look equivalent or of comparable quality, write down 1"*). Because candidate machine translations are presented in random order, for the annotator, this comes down to assigning a random label. But because we would otherwise expect approximately equal numbers of "1" and "2" labels, we can estimate the number of "equivalent" cases. Globally, we observe 417 translations annotated "1" (59.6%) and 283 annotated "2" (40.4%). This suggests that only approximately 283 of the 417 translations annotated "1" denote a real preference for translation in column 1 (68%), and that the remaining 134 (32%) are actually "random" selections, denoting equivalent translations. The overall rate of "equivalent" translations would therefore be 19% (134 of 700 translation pairs).

2. The Trouble with YiSi

As can be seen in Table 1, for EN-FR translations, YiSi assigns a higher global score to Portage than to Sockeye. In this, it disagrees with BLEU, WER and human judges. Even for the FR-EN, it almost seems overly optimistic about Portage translations. Of course, these are global (document-level) evaluations, and YiSi is not known to perform any better than most other metrics at this level. Still, it is somewhat surprising to see YiSi disagreeing with all other metrics, especially on such a large test set.

As mentioned earlier, the sample on which the human evaluation is based is not com-

System:	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
Global preference:	45.5%	54.5%	54.3%	45.7%
Weighted preference:	45.0%	55.0%	53.1%	46.9%
Per annotator results:				
Annotator 1	41	59	53	47
Annotator 2	41	59	52	48
Annotator 3	45	55	58	42
Annotator 4	55	45	–	–

Table 2: Human preferences on subsamples of the Hansard test set.

pletely unbiased: the sampling was done so as to focus on sentences that are neither too short nor too long, and for which Portage and Sockeye produced different translations. Could it be that systems behave differently on such sentences than on short and long sentences? Computing metrics on the much smaller subsamples on which human evaluations were performed (300 translations into English, 400 translations into French) yields somewhat different results than on the complete test set: in Table 3, YiSi now agrees with BLEU and the human judges, but WER disagrees regarding EN-FR. This perhaps suggest that this subsample is too small to reliably compute automatic metric scores.

System:	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
BLEU	33.3	35.2	34.1	33.4
WER	58.2	56.1	51.9	51.2
YiSi	66.2	66.6	71.5	69.8

Table 3: Automatic metrics on the subsample of Hansard test set used for the human evaluation.

To get a better picture yet, we split the complete test set into three: *short* (source-language) sentences of 7 word tokens or less, *medium* sentences between 8 and 24 tokens long (from which the human evaluation subset was sampled), and *long* sentences, with 25 tokens or more. When the test set is split this way, all three metrics agree, and display a more revealing picture (Table 4). For FR-EN, all three metrics favor Portage over Sockeye, regardless of sentence size. However, for EN-FR, all metrics give better scores to Sockeye on medium and long sentences, but very clearly prefer Portage on

short sentences. This is likely explained by Portage’s “translation memory” effect on sentences that are effectively within the reach of single phrases in its phrase-table.

Sentence size	System:	EN-FR		FR-EN	
		Portage	Sockeye	Portage	Sockeye
Short:		(631 segments)		(520 segments)	
7 tokens or less	BLEU	43.4	35.7	50.0	37.6
	WER	39.3	50.0	34.0	50.0
	YiSi	78.4	69.2	86.2	69.4
Medium:		(1868 segments)		(2226 segments)	
8-24 tokens	BLEU	38.2	40.7	39.6	38.6
	WER	54.0	52.1	50.8	51.8
	YiSi	71.2	71.5	76.6	74.4
Long:		(2364 segments)		(2117 segments)	
25 tokens or more	BLEU	36.8	40.3	38.9	38.3
	WER	53.5	50.3	49.1	49.5
	YiSi	70.2	70.8	76.4	74.1

Table 4: Automatic metrics on the Hansard test set used for the human evaluation.

But this result also suggests an explanation for YiSi’s disagreement on this language direction: because it is computed as a straight average over sentence-level scores, it tends to over-value the effect of short sentences on the overall quality. In the EN-FR test set, short sentences account for 12.9% of all sentences, but only for 2.8% of the words. In contrast, BLEU and WER’s formulations are such that sentences naturally contribute to the overall score proportionally to their length. This suggests that it might make sense to compute document-level YiSi scores as a weighted average of sentence-level scores instead, in which weights are proportional to each sentence’s length. If the global (document-level) score for machine translations $M = m_1 \dots m_N$, given reference translations $R = r_1 \dots r_N$ is computed as the mean of segment-level scores:

$$Y(M, R) = \frac{1}{N} \sum_{i=1}^N Y(m_i, r_i),$$

then a weighted version of that score could be computed as:

$$Y_w(M, R) = \frac{\sum_{i=1}^N |r_i| Y(m_i, r_i)}{\sum_{j=1}^N |r_j|},$$

where $|r_j|$ denotes the number of tokens in segment r_j .

In Table 5, this weighted version is referred to as *w-YiSi*. As can be seen, its value differs only slightly from the unweighted YiSi, but enough that the new metric now globally agrees with BLEU, WER and human judgement.

System:	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
BLEU	37.9	40.5	39.8	38.7
WER	53.0	51.5	49.8	51.5
YiSi	70.3	70.0	74.4	70.0
w-YiSi	70.2	70.4	74.0	71.6

Table 5: BLEU, WER, YiSi and weighted YiSi (w-YiSi) scores on Hansard test set.

3. Per-sentence Preference

In a scenario where automatic metrics are used to compare MT systems, when using sentence-level metrics, it can make sense to look at per-sentence preference: rather than report the scores themselves, report the proportion of test sentences on which a given metric prefers one system over the other(s). In other words: report the results of automatic evaluations the same way we report those of human evaluations.

Table 6 shows these proportions on the small test subsets used for the human evaluation, for BLEU, WER and YiSi (sentence-level BLEU scores were computed with add-1 smoothing, à la Lin & Och 2004). Note that preference percentages don’t necessarily add to 100: that’s because ties are excluded from the counts. As can be seen, for all metrics, this way of comparing systems gives a global picture in agreement with the global human evaluation. It is particularly noteworthy that, while the global YiSi scores were mildly in favor of Portage for EN-FR translations (Table 1), in terms of individual per-sentence preferences, YiSi clearly prefers Sockeye.

Table 6 also shows the percentage of individual segments on which the given metric agrees with the human judge (under “Agreement”). It is interesting to note that, at the level of individual sentences, YiSi is substantially better than BLEU and WER at predicting which translation the human will prefer.²

²As discussed previously, human annotators were not given the possibility to explicitly label “equivalent” translations as such. This means that a fraction of human labels are actually randomly assigned, which we estimated in Section 1 to be approximately 19%. For example, if the human judge assigned the label “1” to some example, and if a given metric reports a tie for that example, then there is a possibility that the metric and the judge actually agree. The Agreement rates reported here do not account for this factor, but we examine this in Appendix A.

System:	EN-FR			FR-EN		
	Portage	Sockeye	Agreement	Portage	Sockeye	Agreement
BLEU	39.2%	50.5%	55.0%	52.0%	43.3%	63.3%
WER	35.0%	43.5%	49.2%	41.7%	40.0%	54.7%
YiSi	42.0%	53.5%	61.2%	55.7%	44.3%	67.3%

Table 6: Per-segment preferences of automatic metrics, on the subsample of Hansard test set used for the human evaluation. The “Agreement” column denotes the percentage of segments on which the given metric agrees with the human judge.

In Table 7, we report preferences for the complete test set, for all three metrics. Here again, we compute “weighted preferences”, in which each sentence’s contribution to the global preference rate is proportional to its length; these are reported as *w-pref* in the table. We can see that all metrics report preferences in line with global human preferences, whether weighted or not. One observation is that the difference between the two systems is more marked by preferences than by absolute score: For example, for the FR-EN systems, a difference of 1.1 BLEU point (39.8 vs. 38.7 – see Table 5) results in a 7% difference in preference rates (46.5% vs. 39.5%), and almost 9% with weighted preference rates (51.2% vs. 42.3%). For the same language direction, a difference of 2.4 points in weighted YiSi scores results in almost 20% difference in preference rates (53.9% vs. 34.5%), and close to 24% difference for weighted preference rates (59.4% vs. 35.6%).

System:	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
BLEU pref	38.8	50.2	46.5	39.5
WER pref	35.6	43.7	40.4	34.1
YiSi pref	42.6	48.1	53.9	34.5
BLEU w-pref	40.3	55.5	51.2	42.3
WER w-pref	37.5	48.5	45.1	36.6
YiSi w-pref	45.7	51.1	59.4	35.6

Table 7: Automatic metric preferences on the complete Hansard test set .

Conclusions

Overall, our automatic and manual evaluations of the Portage and Sockeye translations lead us to the conclusion that while Sockeye seems to perform better for English-to-

French, Portage is still the better option for French-to-English. We have not performed a qualitative evaluation of the translation errors, but a superficial examination of comments by judges suggests that Sockeye’s BPE-induced translations are often a problem, not only for OOV’s but even for mildly infrequent words. We suspect that this type of error would be completely unacceptable for professional translators, who would likely prefer Portage’s approach of copying OOV’s over to the target. For post-editing applications, this is a weakness of NMT that we will want to tackle.

This evaluation exercise was an opportunity to witness some strengths and weaknesses of YiSi. Among the most obvious weaknesses, we noted its macro-averaging of sentence-level scores to produce document-level evaluations, which results in over-valuing the quality of short sentences. Our experiments seem to suggest that using a weighted average instead might solve the problem, and lead to better correlation with human judgement.

We also noted that on individual sentence pairs, YiSi is substantially better than BLEU or WER at predicting which translation a human judge would prefer. We make the observation that, when comparing systems, it is possibly more informative for users to report preference rates (either weighted or not) than just absolute scores.

Appendix

A. Accounting for equivalent labels

As discussed earlier, annotators were not given the option to label translations as “equivalent”. For situations where they truly felt that both translations were equivalent, annotators were instructed to mark a preference for translation “1”. Because candidate machine translations were presented in random order, from a statistical point of view, this is effectively equivalent to assigning a random label. But it also allows us to estimate the proportion of “equivalent” translations. In the manual evaluation described here, we estimated that 32% of 1’s are actually “random” selections, denoting equivalent translations.

When measuring the agreement between per-sentence preferences of automatic metrics and human annotations (Table 6), we did not take this factor into account. To do so, we need to consider the probability that a label “1” actually denotes an equivalent translation rather than a preference for the translation from MT system in column 1. Based on the above estimate, every example labeled 1 has probability 0.32 of denoting an equivalent translation, and 0.68 of denoting an actual preference for MT system in column 1. This means that even when the automatic metric and the human label indicate a preference for the same system, if the annotator label is “1”, there is actually a 0.32 probability that the annotator really meant “equivalent”, in which case this should be counted as a disagreement. One way of accounting for this in our statistics is to count these situations as only “0.68 agreement”. Conversely, when the automatic metric produces a tie, if the annotator label is “1”, then there is actually a 0.32 probability that the metric and annotator actually agree: these situation can be counted as “0.32

agreement”.

In formal terms, given a series of automatic metric preferences $M = m_1 \dots m_n$ that can take values *system A*, *system B* or *equivalent*, corresponding annotator preferences $A = a_1 \dots a_n$ that take values *system A* or *system B* (but not *equivalent*), and the actual annotation labels $X = x_1 \dots x_n$ that are either “1” or “2”, we can measure the agreement rate between the metric and the annotator as the average of “agreement counts” $agree(m_i, h_i, x_i)$:

$$agree(m_i, h_i, x_i) = \begin{cases} m_i = h_i : & \begin{cases} x_i = 1 : 0.68 \\ x_i = 2 : 1.0 \end{cases} \\ m_i = \text{equivalent} : & \begin{cases} x_i = 1 : 0.32 \\ x_i = 2 : 0.0 \end{cases} \\ m_i \neq h_i : 0.0 \end{cases}$$

Table 8 compares these *adjusted agreement* rates to raw agreement rates, as reported in Table 6. In practice, this way of measuring produces markedly lower agreement rates. However, the general conclusion remains the same: YiSi is still better than BLEU and WER at predicting human preference on individual segments.

System:	EN-FR		FR-EN	
	Raw agreement	Adjusted agreement	Raw agreement	Adjusted agreement
BLEU	55.0%	47.1%	63.3%	53.1%
WER	49.2%	44.3%	54.7%	48.9%
YiSi	61.2%	51.3%	67.3%	55.5%

Table 8: Agreement between metrics and human judges on per-segment preference. “Raw agreement” is based on a simple count of agreements, “Adjusted agreement” takes into account the fact that some human preferences actually denote “equivalent” judgements.

COLLABORATION DE RECHERCHE BTB - CNRC

9 septembre 2019

Contexte

Les avancées récentes en IA ont permis des progrès substantiels en traduction automatique (TA). Il est maintenant réaliste pour une organisation comme le BtB d'envisager l'utilisation de produits de TA commerciaux dans ses processus.

Dans un environnement où la TA est utilisée de façon systématique, le contrôle de la qualité est plus critique que jamais, et doit également devenir systématique.

→ l'IA peut jouer ici aussi un rôle crucial

Collaboration BtB-CNRC

*« L'IA au service de la
qualité des traductions »*

Objectif

Créer des applications d'IA pour assister le contrôle et l'optimisation de la qualité des traductions au BtB.

Partenaires

- ☐ Réingénierie stratégique du BtB
- ☐ Traitement de textes multilingues du CNRC
- ☐ Laboratoire RALI de l'Université de Montréal

Calendrier

Années fiscales 2018-19, 2019-20 et 2020-21