A Challenge Set Approach to Evaluating Machine Translation

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

George Foster
Google*
fosterg@google.com

Abstract

Neural machine translation represents an exciting leap forward in translation quality. But what longstanding weaknesses does it resolve, and which remain? We address these questions with a challenge set approach to translation evaluation and error analysis. A challenge set consists of a small set of sentences, each hand-designed to probe a system's capacity to bridge a particular structural divergence between languages. To exemplify this approach, we present an English–French challenge set, and use it to analyze phrase-based and neural systems. The resulting analysis provides not only a more fine-grained picture of the strengths of neural systems, but also insight into which linguistic phenomena remain out of reach.

1 Introduction

The advent of neural techniques in machine translation (MT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) has led to profound improvements in MT quality. For "easy" language pairs such as English/French or English/Spanish in particular, neural (NMT) systems are much closer to human performance than previous statistical techniques (Wu et al., 2016). This puts pressure on automatic evaluation metrics such as BLEU (Papineni et al., 2002), which exploit surface-matching heuristics that are relatively insensitive to subtle differences. As NMT continues to improve, these metrics will inevitably lose their effectiveness. Another challenge posed by NMT systems is their opacity: while it was usually clear which phenomena were ill-handled by previous statistical systems—and why—these questions are more difficult to answer for NMT.

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Figure 1: Example challenge set question.

We propose a new evaluation methodology centered around a *challenge set* of difficult examples that are designed using expert linguistic knowledge to probe an MT system's capabilities. This methodology is complementary to the standard practice of randomly selecting a test set from "real text," which remains necessary in order to predict performance on new text. By concentrating on difficult examples, a challenge set is intended to provide a stronger signal to developers. Although we believe that the general approach is compatible with automatic metrics, we used manual evaluation for the work presented here. Our challenge set consists of short sentences that each focus on one particular phenomenon, which makes it easy to collect reliable manual assessments of MT output by asking direct yes-no questions. An example is shown in Figure 1.

We generated a challenge set for English to French translation by canvassing areas of linguistic divergence between the two language pairs, especially those where errors would be made visible by French morphology. Example choice was also partly motivated by extensive knowledge of the weaknesses of phrase-based MT (PBMT). Neither of these characteristics is essential to our method, however, which we envisage evolving as NMT progresses. We used our challenge set to evalu-

*Work performed while at NRC.

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 out-performed 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]

[1]With 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 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.

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.

[2]The 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*.

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 7.

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

[3] Sometimes 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/nnmt-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 7 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 7 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 burden on the annotators. That is, in figure 1, anno-

[5] https://github.com/rsennrich/nematus
[6] https://translate.google.com

tators 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 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 7. 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 agrement:

- *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).

[7] We 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%).

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 *inalienable 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* $\rightarrow$ *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 *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* $\rightarrow$ *couteau* ***en*** *acier/céramique/plastique* but *butter/meat/steak knife* $\rightarrow$ *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 misses NP_2 → NP_2 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 → 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.

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

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.

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 Merrienboer, 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.

Nadir 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. https://workshop2016.iwslt.org/downloads/qcri-machine-translation.pdf.

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. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

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.

Supplemental Materials

The supplemental materials comprise two separate files:

- `challenge-a.pdf`—instructions to authors, and rendered version of the challenge set (with annotator scores); and
- `Challenge_set-v2hA.json`—machine-readable version of the challenge set.