

A Challenge Set Approach to Evaluating Machine Translation

Paper Review

Isabelle, Pierre ¹, Cherry, Colin ¹ & Foster, George ²

February 8, 2024

¹National Research Council Canada (NRCC)

²Google

Table of Contents

Introduction

Related work

Evaluation

Machine translation systems

Conclusion

Introduction

The task: Compose a set of challenging translation problems for English - French translation designed for neural systems

The goal: To test machine translation competency with specific structural divergences between English and French

As Neural Machine Translation (NMT) continues to improve, MT systems for “easy” language pairs such as English/French or English/Spanish are much closer to human performance.

This puts pressure on automatic evaluation metrics such as BLEU (Papineni et al., 2002). Automatic evaluation is less reliable for highly competent models.

Furthermore, it was easier to tell which phenomena were ill-handled by previous statistical systems - and why.

Modern NMTs are kinda good enough that you need more competent assessors to say what its getting wrong.

What is a challenge set?

A challenge set is a set of sentences, each hand-crafted to test for a particular structural divergence between languages.

The intent of each challenge sentence is to test exactly one system capability at a time.

In this paper, the authors focused on English-French and all evaluations were conducted manually.

Related work

The **WMT 2016 News Translation Task** (Bojar et al., 2016) evaluated submissions according to BLEU and human judgments. NMT systems were submitted to 9 out of 12 translation directions, won 4 of these and tied for second in the other 5.

Controlled comparisons used BLEU to show that NMT out-performs strong PBMT systems on 30 translation directions from the **United Nations Parallel Corpus** (Junczys-Dowmunt et al., 2016) and the **IWSLT English-Arabic tasks** (Durrani et al., 2017)

Bentivogli et al. (2016) automatically detected a number of error categories by comparing machine outputs to professional post-edits on **IWSLT 2015 English-German** evaluation data.

NMT required less post-editing effort overall, with fewer lexical, morphological and word order errors.

However, NMT performance degraded faster as a function of sentence length.

Toral and Sánchez-Cartagena (2017) examined the outputs of competition-grade systems for the 9 WMT 2016 directions that included NMT competitors.

Their conclusions regarding morphological inflection and word order were similar to Bentivogli et al. (2016).

However they found an even greater degradation in NMT performance as sentence length increased.

Sennrich (2016) approached targeted evaluation of NMT using *contrastive translation pairs*.

They automatically introduced specific errors in reference sentences, and then checked whether the NMT prefers the original reference or the corrupted version.

They determined that a recently-proposed character-based model improves generalisation on unseen words, but at the cost of introducing new grammatical errors.

Evaluation

To build their challenge set, Isabelle et al. (2017) chose words that occurred at least 100 times in the training corpus.

They classify the structural divergence phenomena they hope to capture in the challenge according to three main types:
morpho-syntactic, **lexico-syntactic** and **purely syntactic**.

Morpho-Syntactic divergence

When translating a word into a morphologically rich language, there is a need to recover additional grammatically relevant information from the context of the target language word.

One important case of morpho-syntactic divergence is *subject-verb agreement*.

English verb forms strongly underspecify their French counterparts morpho-syntactically.

Subject-verb agreement

Extracting the missing information required to force agreement in person, number and gender with the grammatical subject of a verb can be challenging.

The agreement features of a co-ordinated noun phrase are a function of multiple elements:

1. The gender is feminine if all conjuncts are feminine, otherwise masculine wins
2. The conjunct with the smallest person ($p_1 < p_2 < p_3$) wins; and
3. The number is always plural when the coordination is “et” (“and”) but the case is more complex with “ou” (“or”).

Lexico-Syntactic divergence

Syntactically governing words such as verbs may impose specific requirements on their complements: They subcategorize for complements of a certain syntactic type.

However, a source and target language may have different requirements. One example is *argument switching*.

John misses Mary → Mary manque à John.

Lexico-Syntactic divergence

Another example of lexico-syntactic divergence is “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”, however a literal translation such as “Terry a nagé à travers la rivière” would be incorrect.

Syntactic divergence

Syntactic divergences occur not due to the presence of a particular lexical item, but because of differences in the set of available syntactic patterns.

Source language instances missing from the target language must then be mapped onto equivalent structures.

Syntactic divergence

To give an example, French and English are both SVO languages but in French pronouns must be phonetically attached to the verb on its left side when post-verbal complements are prenominalized.

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.

Summary

The actual test set includes several subcategories of each type (morpho-syntactic, lexico-syntactic and purely syntactic) of divergence.

Each subcategory was then tested using at least 3 different test sentences.

Test sentences were kept short in order to keep the targeted divergence in focus. There were 108 sentences included in the challenge set in total.

Machine translation systems

Machine translation systems

They also trained state-of-the-art (for the time) neural and phrase based systems for English-French translation on data from the WMT 2014 evaluation.

Below is a breakdown of the training data they used to train their models:

Table 1: Corpus Statistics

Corpus	Lines	En Words	Fr Words
Train	12.1M	304M	348M
Mono	15.9M	—	406M
Dev	6003	138k	155k
Test	3003	71k	81k

The authors trained a PBMT baseline model, and an NMT model.
Some thoughts:

1. The PBMT model was an Neural Network Joint Language Model (NNJM) which jointly models the target language and source language with target N-gram and source window.
2. They trained 2 PBMT models, PBMT-1 which is trained on identical data to the MNT system, and PBMT-2 which includes a French monolingual corpus.

3. The NMT model used a single-layer sequence to sequence architecture with attention
4. They also trained their own BPE model on the source and target corpora
5. They also evaluated Google's production system (GMNT), which at the time had recently moved to NMT.
6. Google's system uses at least 8 encoder + decoder layers and is trained on corpora 2-3 orders of magnitude greater than the WMT.

All four models (PBMT-1, PBMT-2, NMT and GMNT) were evaluated on the English-French challenge set by 3 native speakers of French, who rated each sentence as either a success or a failure.

The corresponding translations were judged successful if and only if the translated verb correctly agrees with the translated subject.

Table 2: Transposed Comparison of Translation Models

	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	—	—

The final column gives the proportion of system outputs on which all three annotators agreed.

Quantitative comparison

The PBMT models both perform better at the Lexico-syntactic examples, which the authors say reflects the fact that PBMT systems are naturally more attuned to lexical cues than to morphology or syntax.

WMT BLEU scores correlate poorly with challenge-set performance.

Inter-annotator agreement is close overall, but syntactic divergences appear to be harder to judge than other categories.

Qualitative assessment of NMT

Overall NMT beats PBMT in every category but particularly in the case of morpho-syntactic divergences.

The large gap between NMT and GNMT indicates neural MT systems are extremely data hungry.

However in spite of a significant data and compute gap, the GNMT model fails to reach 70% overall on the challenge set.

Table 3: Detailed Translation Model Comparison Across Subcategories

Category	Subcategory	N	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	17%	92%	75%
	With coordinated source	12	0%	100%	100%
	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%

Table 4: Detailed Translation Model Comparison Across Subcategories

Category	Subcategory	N	PBMT-1	NMT	Google NMT
Syntactic	Yes-no question syntax	3	33%	100%	100%
	Tag questions	3	0%	0%	100%
	Stranded preps	6	0%	0%	100%
	Adv-triggerred 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%

Conclusion

Conclusion

With the exception of idiom processing, in all cases where a clear difference was observed the result was in favour of neural MT.

The challenge set brings to light some important shortcomings of current neural MT, regardless of the massive amounts of data used for training.

NMT impressive generalisations still seem brittle - NMT may appear to have mastered subject-verb agreement or inalienable possession only to trip over a rather obvious instantiation of those rules.

Future work

Human judgments are still difficult to scale, and attempts to resolve would be worthwhile if achieved.

An automatic method for constructing a challenge set would be extremely useful.

Localising a divergence within a difficult sentence pair would be another useful subtask.

Lastly, how to train an MT system specifically to improve its performance on particular divergence phenomena.

References






Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). **Neural versus phrase-based machine translation quality: A case study.** *arXiv preprint arXiv:1608.04631*.

Bibliography ii



Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., N  v  l, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., ... Zampieri, M. (2016, August). **Findings of the 2016 conference on machine translation.** In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. J. Yepes, A. N  v  l, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, & M. Turchi (Eds.), *Proceedings of the first conference on machine translation: Volume 2, shared task papers* (pp. 131–198). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/W16-2301>

-  Durrani, N., Dalvi, F., Sajjad, H., & Vogel, S. (2017). **Qcri machine translation systems for iwslt 16.** *arXiv preprint arXiv:1701.03924*.
-  Isabelle, P., Cherry, C., & Foster, G. (2017). **A challenge set approach to evaluating machine translation.** *arXiv preprint arXiv:1704.07431*.
-  Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). **Is neural machine translation ready for deployment? a case study on 30 translation directions.** *arXiv preprint arXiv:1610.01108*.



Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). **Bleu: A method for automatic evaluation of machine translation.** *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.



Sennrich, R. (2016). **How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs.** *arXiv preprint arXiv:1612.04629*.



Toral, A., & Sánchez-Cartagena, V. M. (2017). **A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions.** *arXiv preprint arXiv:1701.02901*.