Multilingual Whispers: Generating Paraphrases with Translation

Christian Federmann, Oussama Elachqar, Chris Quirk
Microsoft
One Microsoft Way
Redmond, WA 98052 USA
{chrife,ouelachq,chrisq}@microsoft.com

Abstract

Naturally occurring paraphrase data, such as multiple news stories about the same event, is a useful but rare resource. This paper compares translation-based paraphrase gathering using human, automatic, or hybrid techniques to monolingual paraphrasing by experts and non-experts. We gather translations, paraphrases, and empirical human quality assessments of these approaches. Neural machine translation techniques, especially when pivoting through related languages, provide a relatively robust source of paraphrases with diversity comparable to expert human paraphrases. Surprisingly, human translators do not reliably outperform neural systems. The resulting data release will not only be a useful test set, but will also allow additional explorations in translation and paraphrase quality assessments and relationships.

1 Introduction

Humans naturally paraphrase. These paraphrases are often a byproduct: when we can't recall the exact words, we can often generate approximately the same meaning with a different surface realization. Recognizing and generating paraphrases are key challenges in many tasks, including translation, information retrieval, question answering, and semantic parsing. Large collections of sentential paraphrase corpora could benefit such systems.[1]

Yet when we ask humans to generate paraphrases of a given task, they are often a bit stuck. How much should be changed? Annotators tend to preserve the reference expression: a safe choice, as the only truly equivalent representation is to leave the text unchanged. Each time we replace a word with a synonym, some shades of meaning change, some connotations or even denotations shift.

[1]Expanding beyond the sentence boundary is also very important, though we do not explore cross-sentence phenomena in this paper.

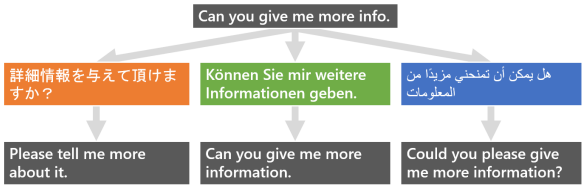

Figure 1: Generating broad-coverage paraphrases through pivot translation.

One path around the obstacle of reference bias is to provide a non-linguistic input, then ask humans to describe this input in language. For instance, crowd-sourced descriptions of videos provide a rich source of paraphrase data that is grounded in visual phenomena (Chen and Dolan, 2011). Such visual grounding helps users focus on a clear and specific activity without imparting a bias toward particular lexical realizations. Unfortunately, these paraphrases are limited to phenomena that can be realized visually. Another path is to find multiple news stories describing the same event (Dolan et al., 2004), or multiple commentaries about the same news story (Lan et al., 2017). Although this provides a rich and growing set of paraphrases, the language is again biased, this time toward events commonly reported in the news.

An alternative is to provide input in a foreign language. Nearly anything expressible in one human language can be written in another language. When users translate content, some variation in lexical realization occurs. To gather monolingual paraphrases, we can first translate a source sentence into a variety of target languages, then translate back into the source language, using either humans or machines. This provides naturalistic variation in language, centered around a common yet relatively unconstrained starting point. Although several research threads have explored this possibility (e.g., (Wieting and Gimpel, 2018)), we have seen few if any comparative evaluations of the quality of this approach.

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 17–26
Hong Kong, Nov 4, 2019.

Our primary contribution is to evaluate various methods of constructing paraphrase corpora, including monolingual methods with experts and non-experts as well as automated, semi-automated, and manual translation-based approaches. Each paraphrasing method is evaluated for fluency (*"does the resulting paraphrase sound not only grammatical but natural?"*) and adequacy (*"does the paraphrase accurately convey the original meaning of the source?"*) using human direct assessment, inspired by effective techniques in machine translation evaluation (Federmann, 2018).

In addition, we measure the degree of change between the original and rewritten sentence using both edit distance and BLEU (Papineni et al., 2002). Somewhat surprisingly, fully automatic neural machine translation actually outperforms manual human translation in terms of adequacy. The semi-automatic method of post-editing neural machine translation output with human editors leads to fluency improvements while retaining diversity and adequacy. Although none of the translation-based approaches outperform monolingual rewrites in terms of adequacy or fluency, they do produce greater diversity. Human editors, particularly non-experts, tend toward small edits rather than substantial rewrites. We conclude that round-tripping with neural machine translation is a cheap and effective means of gathering diverse paraphrases.

Our second contribution is a unique data release. As a byproduct of this evaluation, we have compiled a data set consisting of paraphrases gathered using monolingual rewrites and translation paraphrases generated through human translation, neural machine translation, and human post-edited neural machine translation. These 500 source sentences—together with all rewrites and intermediate translations—comprise a rare and interesting multilingual data set, useful for both monolingual and translation tasks. We include all human quality assessments for adequacy (semantic equivalence) and fluency of paraphrases, as well as translation adequacy assessments. Data is publicly available at `https://aka.ms/MultilingualWhispers`.

2 Related Work

Translation as a means of generating paraphrases has been explored for decades. Paraphrase corpora can be extracted from multiple translations of the same source material (Barzilay and McKeown, 2001). Sub-sentential paraphrases (mostly phrasal replacements) can be gathered from these multiple translations. Alternatively, one can create a large body of phrasal replacements from by pivoting on the phrase-tables used by phrase-based statistical machine translation (Bannard and Callison-Burch, 2005; Ganitkevitch et al., 2013; Pavlick et al., 2015).

Recent work has also explored using neural machine translation to generate paraphrases via pivoting (Prakash et al., 2016; Mallinson et al., 2017). One can also use neural MT systems to generate large monolingual paraphrase corpora. Another line of work has translated the Czech side of a Czech-English parallel corpus into English, thus producing 50 million words of English paraphrase data (Wieting and Gimpel, 2018). Not only can the system generate interesting paraphrases, but embeddings trained on the resulting data set prove useful in sentence similarity tasks. When added to a paraphrase system, constraints obtained from a semantic parser can reduce the semantic drift encountered during rewrites (Wang et al., 2018). Adding lexical constraints to the output can also increase diversity (Hu et al., 2019).

Past research has also explored effective methods for gathering paraphrases from the crowd (Jiang et al., 2017). However, to the best of our knowledge, no prior work has compared the efficacy of human experts, crowd-workers, human post-editing approaches and machine translation systems on gathering paraphrase quality.

3 Methodology

To run a comprehensive evaluation of paraphrase techniques, we create many paraphrases of a common data set using multiple methods, then evaluate using human direct assessment as well as automatic diversity measurements.

3.1 Data

Input data was sampled from two sources: Reddit provides volumes of casual online conversations; the Enron email corpus represents communication in the professional world.[2] Both are noisier than usual NMT training data; traditionally, such noise has been challenging for NMT systems (Michel and Neubig, 2018) and should provide a lower-bound on their performance. It would definitely be valuable, albeit expensive, to rerun our experiments on a cleaner data source.

[2]However, the Enron emails often contain conversations about casual and personal matters.

			Tokens per segment			
Segments	Types	Tokens	median	mean	min	max
500	2,370	9,835	19	19.67	4	46

Table 1: Key characteristics of the source sentences.

			Tokens per segment			
Segments	Types	Tokens	median	mean	min	max
14,500	7,196	285,833	19	19.72	1	68

Table 2: Key characteristics of collected paraphrases.

As an initial filtering step, we ran automatic grammar and spell-checking, in order to select sentences that exhibit some disfluency or clear error. Additionally, we asked crowd workers to discard sentences that contain any personally identifiable information, URLs, code, XML, Markdown, and non-English sentences. The crowd workers were also encouraged to select noisy sentences containing slang, run-ons, contractions, and other behavior observed in informal communications.

3.2 Paraphrase techniques

Expert human monolingual paraphrase. We hired trained linguists (who are native speakers of English) to provide paraphrases of the given source sentences, targeting highest quality rewrites. These linguists were also encouraged to fix any misspellings, grammatical errors, or disfluencies.

Crowd-worker monolingual paraphrase. As a less expensive and more realistic setting, we asked English native speaking crowd workers who passed a qualification test to perform the same task.

Human round-trip translation. For the first set of translation-based paraphrases, we employed human translators who translated the source text from English into some pivot language and back again. The translations were provided by a human translation service, potentially using multiple different translators (though the exact number was not visible to us). In our experiments we focused on a diverse set of pivot languages, namely: Arabic, Chinese, French, German, Japanese, and Russian.

While French and German seem like a better choice for translation from and back into English, due to the close proximity of English as part of the Germanic language family and its shared vocabulary with French, we hypothesize that the use of more distant pivot languages may result in a greater diversity of the back translation output.

We employed professional translators—native in the chosen target language—who were instructed to generate translations from scratch, without the use of any online translation tools. Translation from English into the pivot languages and back into English were conducted in separate phases, by different translators.

Post-edited round-trip translation. Second, we created round-trip translation output based on human post-editing of neural machine translation output. Given the much lower post-editing cost, we hypothesize that results contain only minimal edits, mostly improving fluency but not necessarily fixing problems with translation adequacy.

Neural machine translation. We kept the NMT output used to generate post-editing-based paraphrases, without further human modification. Given the unsupervized nature of machine translation, we hypothesize that resulting output may be closer to the source syntactically (and hopefully more diverse lexically), especially those source sentences which a human editor would consider incomplete or low quality.

Crowd-worker monolingual paraphrase grounded by translation. Finally, we also use a variant of the Crowd-worker monolingual paraphrase technique where the crowd worker is *grounded* by a translation-based paraphrase output. The crowd worker is then asked to modify the translation-based paraphrase to make it more fluent than the source, and as adequate.

Intuitively, one assumes that human translation output should achieve both highest adequacy and fluency scores, while post-editing should result in higher adequacy than raw neural machine translation output.

Considering translation fluency scores, NMT output should be closer to both post-editing and human translation output, as neural MT models usually achieve high levels of fluency (Bojar et al., 2016; Castilho et al., 2017; Läubli et al., 2018).

We hypothesize that translation helps to increase diversity of the resulting back translation output, irrespective of the specific method.

3.3 Assessments

We measure four dimensions of quality:

1. Paraphrase adequacy;
2. Paraphrase relative fluency;
3. Translation adequacy;
4. Paraphrase diversity.

Eval mode	Priming question used
Par_A	How accurately does candidate text B convey the original semantics of candidate text A? Slider ranges from *Not at all* (left) to *Perfectly* (right).
Par_F	Which of the two candidate texts is more fluent? Slider marks preference for *Candidate A* (left), no difference (middle) or preference for *Candidate B* (right).
NMT_A	How accurately does the above candidate text convey the original semantics of the source text? Slider ranges from *Not at all* (left) to *Perfectly* (right).

Table 3: Priming questions used for human evaluation of paraphrase adequacy (Par_A), paraphrase fluency (Par_F), and translation adequacy (NMT_A). Paraphrase evaluation campaigns referred to source and candidate text as "candidate A" and "B", respectively. Translation evaluation campaigns used "source" and "candidate text" instead.

Paraphrase adequacy For adequacy, we ask annotators to assess semantic similarity between source and candidate text, labeled as "candidate A" and "B", respectively. The annotation interface implements a slider widget to encode perceived similarity as a value $x \in [0, 100]$. Note that the exact value is hidden from the human, and can only be guessed based on the positioning of the slider. Candidates are displayed in random order, preventing bias.

Paraphrase fluency For fluency, we use a different priming question, implicitly asking the human annotators to assess fluency for candidate "B" relative to that of candidate "A". We collect scores $x \in [-50, 50]$, with -50 encoding that candidate "A" is much more fluent than "B", while a value of 50 denotes the polar opposite. Intuitively, the middle value 0 encodes that the annotator could not determine a meaningful difference in fluency between both candidates. Note that this may mean two things:

1. candidates are semantically equivalent but similarly fluent or non-fluent; or
2. candidates have different semantics.

We observe that annotators have a tendency to fall back to "neutral" $x = 0$ scoring whenever they are confused, e.g., when semantic similarity of both candidates is considered low.

Translation Adequacy We measure translation adequacy using our own implementation of source-based direct assessment. Annotators do not know that the source text shown might be translated content, and they do not know about the actual goal of using back-translated output for paraphrase generation. Except for the labels for source and candidate text, the priming question is identical to the one used for paraphrase adequacy evaluation. Notably, we have to employ bilingual annotators to collect these assessments. Scores for translation adequacy again are collected as $x \in [0, 100]$.

Paraphrase diversity Additionally, we measure diversity of all paraphrases (both monolingual and based on translation) by computing the average number of token edits between source and candidate texts. To focus our attention on meaningful changes as opposed to minor function word rewrites, we normalize both source and candidate by lower-casing and excluding any punctuation and stop words using NLTK (Bird et al., 2009).

We adopt *source-based direct assessment* (src-DA) for human evaluation of adequacy and fluency. The original DA approach (Graham et al., 2013, 2014) is reference-based and, thus, needs to be adapted for use in our paraphrase assessment and translation scoring scenarios. In both cases, we can use the source sentence to guide annotators in their assessment. Of course, this makes translation evaluation more difficult, as we require bilingual annotators. Src-DA has previously been used, e.g., in (Cettolo et al., 2017; Bojar et al., 2018).

Direct assessment initializes mental context for annotators by asking a priming question. The user interface shows two sentences:

- the source (src-DA, reference otherwise); and
- the candidate output.

Annotators read the priming question and both sentences and then assign a score $x \in [0, 100]$ to the candidate shown. The interpretation of this score considers the context defined by the priming question, effectively allowing us to use the same annotation method to collect human assessments with respect to the different dimensions of quality a defined above. Our priming questions are shown above in Table 3.

3.4 Profanity handling

Some source segments from Reddit contain profanities, which may have affected results reported in this paper. While a detailed investigation of such effects is outside the scope of this work, we want

Method	$\text{Par}_A \uparrow$	$\text{Par}_F \uparrow$	$\text{Par}_D \uparrow$	$\text{NMT}_A \uparrow$
Expert	83.20	**11.80**	3.48	–
HT	63.13	-7.13	5.98	88.8
NMT	64.62	-8.60	3.58	85.1
Non-Expert	**87.10**	9.40	1.11	–
Post-Edited NMT	67.57	-4.20	4.43	**90.0**
Multi-Hop NMT	42.05	-20.65	**6.18**	50.7

Table 4: Results by paraphrasing method. Adequacy (Par_A) and fluency (Par_F) are human assessments of paraphrases; paraphrase diversity (Par_D) is measured by the average string-edit-distance between source and paraphrase (higher means greater diversity); NMT_A is a human assessment of translation quality.

to highlight two potential issues which could be introduced by profanity in the source text:

1. Profanity may have caused additional monolingual rewrites (in an attempt to clean the resulting paraphrase), possibly inflating diversity scores;
2. Human translators may have performed similar cleanup, increasing the likelihood of back translations having a lower adequacy score.

4 Results

In total, we collect 14,500 paraphrases from 29 different systems, as described below:

- Expert paraphrase;
- Non-Expert paraphrase;
- Human translation (HT), for 6 languages;
- Human Post-editing (PE), for 6 languages;
- Neural MT (NMT), for 6 languages;
- Neural "multi-hop" NMT, for 2 languages;
- Grounded Non-Expert (GNE), with grounding from 7 translation methods.

All data collected in this work is publicly released. This includes paraphrases as well as assessments of adequacy, fluency, and translation adequacy. Human scores are based on two evaluation campaigns—one for adequacy, the other for fluency—with $t = 27$ annotation tasks, $a = 54$ human annotators, $r = 4$ redundancy, and $tpa = 2$ tasks per annotator, resulting in a total of $t * r = a * tpa = 108$ annotated tasks—equivalent to at least $9,504$ assessments per campaign (more in case of duplicates in the set of paraphrases to be evaluated), based on the alternate HIT structure with $88 : 12$ candidates-vs-controls setting as described in (Bojar et al., 2018).

Language	$\text{Par}_A \uparrow$	$\text{Par}_F \uparrow$	$\text{Par}_D \uparrow$	$\text{NMT}_A \uparrow$
Arabic	58.33	-12.57	4.96	81.6
Chinese	61.57	-7.67	5.70	71.3
Chinese-Japanese	40.60	-22.30	**6.42**	53.9
French	**71.50**	**-1.80**	3.68	84.2
German	70.90	-2.77	3.80	**87.5**
Japanese	59.67	-9.33	5.38	69.5
Japanese-Chinese	43.50	-19.00	5.95	47.4
Russian	68.67	-5.73	4.47	81.4

Table 5: Results by pivot language.

Table 4 presents empirical results organized by paraphrasing method, while Table 5 organizes by pivot languages used. "Multi-Hop NMT" refers to an experiment in which we created paraphrases translating via two non-English pivot languages, namely Chinese and Japanese. French and German perform best as pivot languages, while Chinese-Japanese achieves best diversity.

Table 6 shows results from our grounded paraphrasing experiment in which we compared how different translation methods affect monolingual rewriting quality. Based on results in Tables 5, we focus on French and German as our pivot languages. We also keep Chinese-Japanese "Two-Pivot NMT" to see how additional pivot languages may affect resulting paraphrase diversity.

Figure 2 shows convergence of adequacy scores for the grounded paraphrasing experiment, over time. Figure 3 shows convergence of relative fluency scores. Note how clustering reported in Table 6 appears after a few hundred annotations only. The clusters denote sets of systems that are not statistically significantly different.

4.1 Error Analysis

While neural machine translation based paraphrases achieve surprising results in terms of diversity compared to paraphrases generated by human Non-Experts, NMT does not reach the adequacy or fluency level provided by Expert paraphrases. The examples in Table 7 provides a flavor of the outputs from each method and demonstrates some of the error cases.

Partially paraphrasing entities and common expressions. NMT systems often mangle multi-word units, rewriting parts of non-compositional phrases that change meaning ("Material Design" $\rightarrow$ "hardware design") or decrease fluency.

Informal language. Inadequate or disfluent paraphrases are also caused by typos, slang and

Method	$\text{Par}_A \uparrow$	$\text{Par}_F \uparrow$	$\text{Par}_D \uparrow$	BLEU $\downarrow$	Labelling Time [seconds]						
					Min	P_{25}	Median	Mean	P_{75}	Max	StdDev
Non-Expert	91.7	13.3	1.106	78.8	7.47	21.52	30.84	40.35	48.07	120.0	28.34
GNE-PE French	88.2	11.9	2.222	59.9	4.73	10.26	18.64	33.16	43.39	120.0	32.30
Expert	88.2	14.6	3.482	39.0	–	–	–	–	–	–	–
GNE-PE German	88.1	11.7	2.214	60.5	4.50	9.58	15.05	35.36	52.05	120.0	35.91
GNE-NMT German	87.9	10.5	2.068	62.2	2.28	10.72	19.74	30.98	39.62	120.0	29.73
GNE-HT French	85.4	12.4	3.160	47.3	4.50	17.07	39.90	52.21	81.65	120.0	39.37
GNE-NMT French	83.1	5.1	2.374	54.9	1.75	2.80	7.29	22.48	28.64	120.0	30.92
GNE-HT German	82.8	9.9	3.914	36.8	6.02	14.48	41.47	50.53	76.67	120.0	38.66
GNE-NMT Chinese-Japanese	74.3	4.3	4.608	32.8	3.84	24.08	45.83	54.11	79.17	120.0	35.45

Table 6: Results for translation-based rewriting, ordered by decreasing average adequacy (Par_A). Horizontal lines between methods denote significance cluster boundaries. Edits measures average number of edits needed to create rewrite (higher means greater diversity). BLEU score measures overlap with original sentence (lower means greater diversity). Labelling time measured in seconds, with a maximum timeout set to two minutes. P_{25} and P_{75} refer to the 25th and 75th percentiles of observed labelling time, respectively; StdDev to standard deviation.

other informal patterns. As prior work has mentioned (Michel and Neubig, 2018), NMT models often corrupt these inputs, leading to bad paraphrases.

Negation handling. One classic struggle for machine translation approaches is negation – losing or adding negation is a common error type. Paraphrases generated through NMT are no exception.

4.2 Key findings

Given our experimental results, we formulate the following empirical conclusions:

"Monolingual is better" Human rewriting achieves higher adequacy and fluency scores compared to all tested translation methods. This comes at a relatively high cost, though.

"Non-experts more adequate..." Human experts appear worse than non-experts in adequacy. We have empirically identified a way to either save or produce more paraphrases for the same budget.

"...but less diverse" Non-expert paraphrases are not as diverse as those created by experts. Expert rewrites also fix source text issues such as profanity.

"MT is not bad" Neural machine translation performs surprisingly well, creating more diverse output than human experts.

"Post-editing is better" Paraphrase adequacy, paraphrase fluency and translation adequacy benefit from human post-editing. In our experiments, this method achieved best performance of all tested translation methods.

"Human translations are expensive and less adequate" While humans achieve high translation adequacy scores and good paraphrase diversity, the corresponding paraphrase adequacy values are worst among all tested methods (except two-pivot NMT, which solves a harder problem).

"Related languages are better..." Generating paraphrases by translation works better when pivot languages are closely related.

"...but less diverse" Unrelated pivot languages create more diverse paraphrases.

"Use neural MT for cheap, large data!" Seems good enough to work for constrained budgets, can be improved with post-editing as needed. Specifically, we have empirically proven that you can increase paraphrase diversity by using NMT pivot translation, combined with non-expert rewriting.

5 Conclusions

Somewhat surprisingly, strong neural machine translation is more effective at paraphrase generation than humans: it is cheap, adequate, and diverse. In contrast, crowd workers required more money, producing more adequate translations but with trivial edits. Although neural MT also produced less fluent outputs, post-editing could improve the quality with little additional expenditure. Expert linguists produced the highest quality paraphrases, but at substantially greater cost. Translation-based paraphrases are more diverse.

One limitation of this survey is the input data selection: generally all input sentences contained some kind of error. This may benefit some techniques – humans in particular can navigate these errors easily. Also, the casual data used often included profanity and idiomatic expressions. Translators often rewrote profane expressions, perhaps decreasing adequacy. Future work on different data sets could further quantify such data effects.

Method	Text
ORIGINAL	Rick, It was really great visiting with you the other day.
EXPERT	Rick, it was ~~really~~ great visiting ~~with~~ you the other day.
NMT CHINESE-JAPANESE	Rick, the visit with you a few days ago was great.
PE GERMAN	Rick, it was ~~really~~ great visiting with you the other day.
PE FRENCH	Rick, It was really fantastic visiting with you the other day.
HT FRENCH	Rick, it was really good to visit you the other day.
HT GERMAN	Rick, it was really great to visit you recently.
NON-EXPERT	Rick, it was really great visiting with you the other day.
NMT FRENCH	Rick, it was really great to visit with you the other day.
NMT GERMAN	Rick, It was really great to visit with you the other day.
ORIGINAL	Yeah exactly, btw how did u manage to update ur nvidia driver ?
EXPERT	Yes, exactly. How did you update your Nvidia driver?
NMT CHINESE-JAPANESE	Yes, exactly. By the way, how were you able to update your NVIDIA drivers?
PE GERMAN	Yes, exactly - how did you update your Nvidia driver?
PE FRENCH	Yes exactly, by the way how did you manage to update your Nvidia driver ?
HT FRENCH	Yeah, exactly, by the way, how did you manage to update your NVIDIA driver?
HT GERMAN	Yes exactly, moreover, how did you manage to update your NVIDIA driver?
NON-EXPERT	Yes, exactly. By the way, did you manage to update your Nvidia driver?
NMT FRENCH	Yes exactly, BTW How did you manage to update your NVIDIA driver?
NMT GERMAN	Yes, exactly, ~~btw~~ how did you manage to update your nvidia driver?
ORIGINAL	Is it actually more benefitial/safe to do this many exercises a day?
EXPERT	Is it ~~actually~~ more beneficial and safe to do so many exercises in a day?
NMT CHINESE-JAPANESE	Tell me if daily practice is good?
PE GERMAN	Is it actually more safe and important to do this many exercises a day?
PE FRENCH	Is it actually more benefitial/safe to do as many exercises a day?
HT FRENCH	Is it really more beneficial/safe to do so much exercise per day?
HT GERMAN	Is it really more beneficial / safer to do so many exercises per day?
NON-EXPERT	Is it actually more beneficial and safe to do this many exercises a day?
NMT FRENCH	Is it actually more benefitial/safe to do ~~this~~ many exercises per day?
NMT GERMAN	Is it actually beneditialat/sure to do these many exercises a day?
ORIGINAL	The cold and rain couldn't effect my enjoyment.
EXPERT	The cold and rain could not affect my enjoyment.
NMT CHINESE-JAPANESE	Cold and rain can not detract from my enjoyment.
PE GERMAN	The cold and rain will not affect my enjoyment.
PE FRENCH	The cold and rain could not effect my enjoyment.
HT FRENCH	Cold and rain dont satisfy me.
HT GERMAN	The cold and rain couldnt spoil my enjoyment.
NON-EXPERT	The cold and the rain couldn't affect my happiness.
NMT FRENCH	The cold and the rain could not affect my pleasure.
NMT GERMAN	The cold and rain couldn't affect my enjoyment.

Table 7: Example paraphrases generated by several monolingual and bilingual methods. Changed regions are highlighted – insertions are presented in green, and deleted phrases from the original sentence are highlighted in ~~red and strikethrough~~. Note how Non-Expert translations tend to be the most conservative, except when clearly informal language is rewritten or corrected.

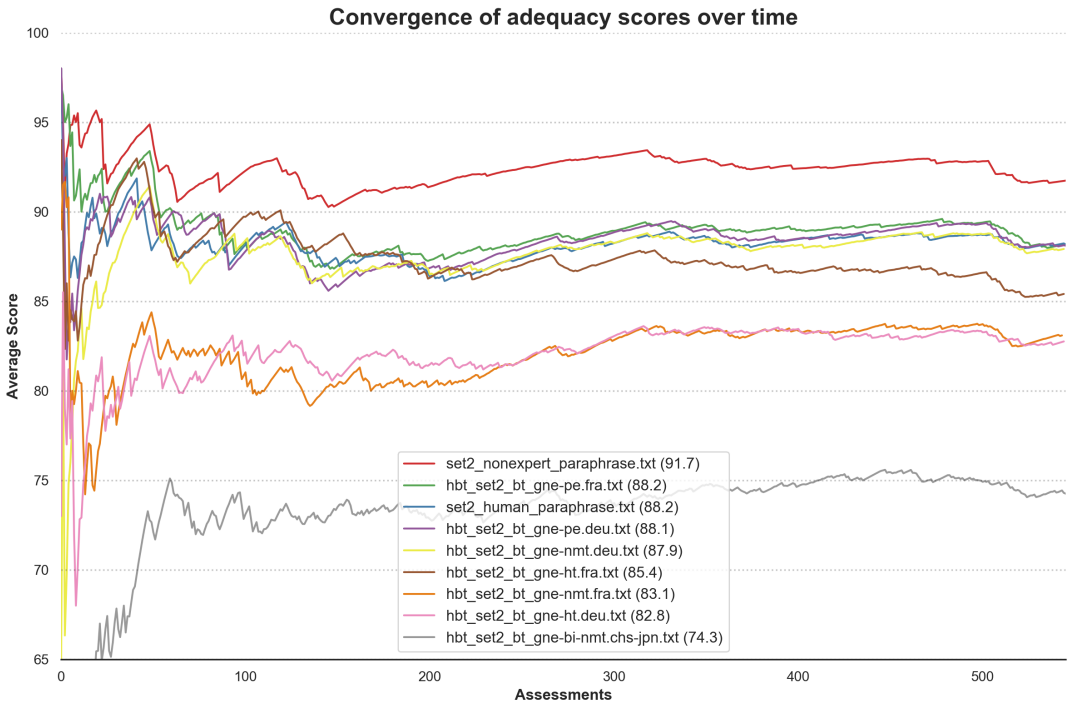

Figure 2: Convergence of adequacy scores over time. Despite the lack of an absolute standard of system assessment, a diverse set of judges rapidly converge to a consistent ranking of system quality. Within a 100 to 200 judgements, the rating has basically stabilized, though we continue to assess the whole set for greatest stability and confidence in ranking. We note, however, that readers should take caution in an absolute reading of these ratings – instead, it should reflect a relative quality assessment among the approaches under consideration.

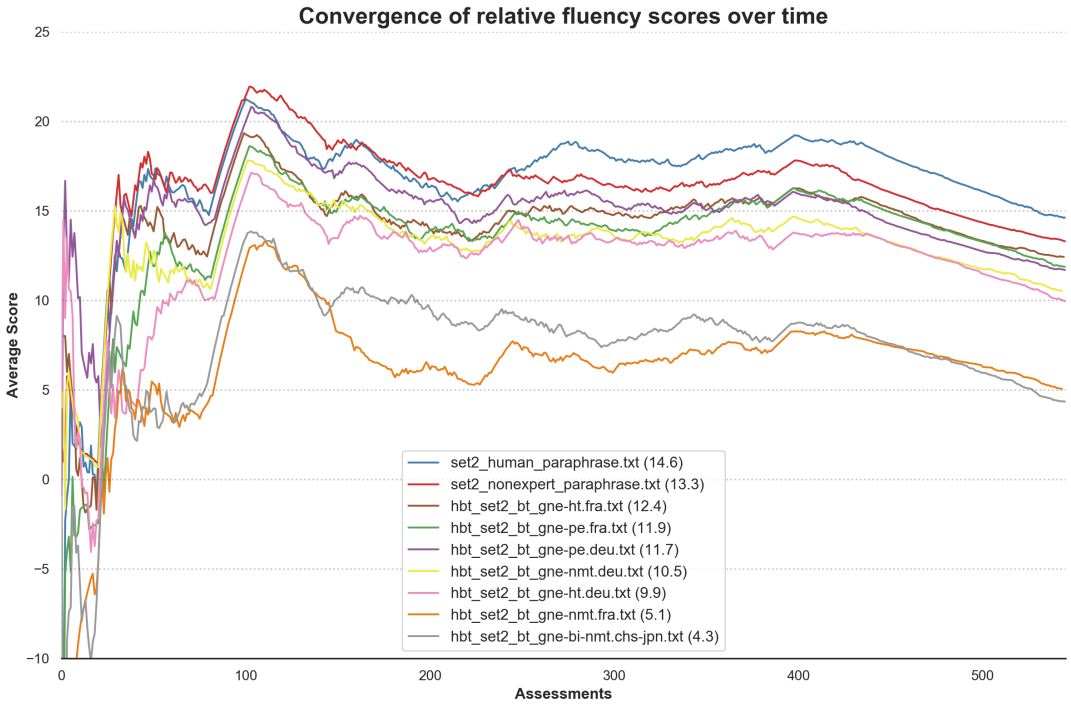

Figure 3: Convergence of relative fluency scores over time. These assessments reflect the same trends as adequacy – raters rapidly converge on a relative assessment of distinct systems.

Acknowledgments

The authors thank the three anonymous reviewers for their feedback and valuable suggestions, which we have addressed in the final version of this paper.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

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*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *International Workshop on Spoken Language Translation*, pages 2–14, Tokyo, Japan.

David Chen and William Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise Evaluation Framework for Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is Machine Translation Getting Better over Time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451. Association for Computational Linguistics.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI*.

Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 103–109, Vancouver, Canada. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A Continuously Growing Dataset of Sentential Paraphrases. In *Proceedings of The 2017 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing Revisited with Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better Paraphrase Ranking, fine-grained Entailment Relations, Word Embeddings, and Style Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural Paraphrase Generation with Stacked Residual LSTM Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2018. A task in a suit and a tie: paraphrase generation with semantic augmentation. *CoRR*, abs/1811.00119.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. Association for Computational Linguistics.