Machine Translation Believability

Marianna J. Martindale† **Kevin Duh**° **Marine Carpuat**‡
†iSchool, ‡Dept. of Computer Science, University of Maryland, College Park, USA
°HLTCOE, Johns Hopkins University, Baltimore, USA
mmartind@umd.edu, kevinduh@cs.jhu.edu, marine@cs.umd.edu

Abstract

Successful Machine Translation (MT) deployment requires understanding not only the intrinsic qualities of MT output, such as fluency and adequacy, but also user perceptions. Users who do not understand the source language respond to MT output based on their perception of the likelihood that the meaning of the MT output matches the meaning of the source text. We refer to this as *believability*. Output that is not believable may be off-putting to users, but believable MT output with incorrect meaning may mislead them. In this work, we study the relationship of believability to fluency and adequacy by applying traditional MT direct assessment protocols to annotate all three features on the output of neural MT systems. Quantitative analysis of these annotations shows that believability is closely related to but distinct from fluency, and initial qualitative analysis suggests that semantic features may account for the difference.

1 Introduction

Past work on evaluating Machine Translation (MT) has focused on the intrinsic quality of the translation product without taking into account how translations are perceived by their users. Yet, some translation errors are more obvious than others, and have different consequences depending on what the translations are used for.

In this work, we take a user-centered view of MT evaluation, exploring one aspect of users' perception of MT: *believability* of the output, defined as a monolingual user's perception of the likelihood that the meaning of the MT output matches the meaning of the input, without understanding the source. Assessing the degree to which MT is believable acknowledges that users play an active role in interpreting its output, informed by their linguistic competence, their common sense reasoning abilities, and their knowledge of the world. What we learn from assessing believability can complement traditional evaluation methods to inform the deployment and even development of MT systems, particularly for gisting and communication use cases.

We first define believability of MT and contextualize it within prior work on credibility and MT evaluation. We apply MT direct assessment (DA) protocols to obtain human judgments of believability, annotating the output of neural machine translation (NMT) systems for three challenging language pairs (Arabic-, Farsi-, and Korean-to-English) with varying translation quality. These annotations show that believability is closely related to, but distinct from fluency. Preliminary qualitative analysis suggests that in addition to fluency features, believability is also influenced by semantic features.

We define *believability* as a user's perception of the likelihood that the meaning of a given MT output matches the meaning of the input, *without understanding the input*. Whether the user accepts the output unquestioningly or finds it unbelievable, their judgment will affect how they act on it, regardless of the true accuracy of the translation. For example, a Facebook user might be dubious of a translation from Chinese with the phrase "blowing a little more cow" and ask the author for clarification and learn that it is a literal translation of an idiom, "吹牛" (meaning to brag). Users may take more consequential action, such as the Israeli police officers who chose not to consult an Arabic speaker before arresting a man based on a believable mistranslation of his "good morning" Facebook post as "attack them" (Berger, 2017).

To illustrate how believability can be independent of adequacy, Table 1 shows examples of different levels of believability for translations at different levels of adequacy, based on our annotations. The translations on the right (More Adequate) convey key information: a named entity (*sputnik*), its

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 88–95
April 20, 2021.

	Less Adequate	More Adequate
More Believable	"putnik" was an interactive film.	"Sputnik" was in the city center, the negative. It was not affected.
Less Believable	spaghetti, the nigerian was in the middle of the city, he didn't touch it.	sputnik was downtown, didn't look, never touched.

Table 1: Machine Translations of human translations of a line from a TED talk discussing the loss of film negatives in a fire (Hoffman, 2008). One film "Sputnik" was spared from the fire because it was not in the building. Original text: *"Sputnik" was downtown, the negative. It wasn't touched.*

location (*downtown*), and the fact that it was not affected, but a user might not accept the information in the bottom-right translation because it is not believable. The less adequate translations (left) are missing important information and also include incorrect information. The bottom-left translation is not believable so a monolingual user would not be misled by it. However, the more believable top-left translation might mislead a monolingual user.

Because these judgments are based on perception, they may be more subjective than traditional MT DA features. We control for some factors that may affect believability (Section 3), resulting in annotations that are similarly reliable to the DA features (Section 4).

2 Related Work

Although believability is an unexplored aspect of MT, there is prior work outside of MT on the general concept of *credibility*. Common elements of credibility include source, media, and message credibility (Rieh and Danielson, 2007). For MT, we can think of the MT provider as the source, the interface through which the MT is viewed as the medium, and the output itself as the message or content. All of these aspects likely affect the credibility of deployed MT systems, but our investigation of MT believability is focused on the content (MT output). Some intrinsic content features addressed in the credibility literature that may affect MT believability include reasonableness (Liu, 2004; Kim and Oh, 2009; Kim, 2010; John et al., 2011) and grammatical errors (Fogg et al., 2001; Everard and Galletta, 2005; Metzger et al., 2010; Chesney and Su, 2010; John et al., 2011).

Reasonableness is a semantic feature encompassing elements of plausibility, logic, and internal consistency. These elements are related to previously studied concepts in the Computational Linguistics literature: semantic plausibility, commonsense reasoning, and discourse coherence. Semantic plausibility can be thought of as, "whether in an ordinary real-life situation (not "fairy-tale" circumstances) the sentence could be reasonably uttered" (Kruszewski et al., 2016). If the source text is expected to reflect "ordinary real-life," the output should be plausible to be believable. MT output may also be unbelievable if it violates commonsense reasoning, a challenging element of Natural Language Understanding (Mostafazadeh et al., 2016). Lack of discourse coherence might likewise signal unbelievable translations. For document generation, improving the consistency of generated documents makes it harder for human subjects to distinguish automatically generated text from real text (Karuna et al., 2018).

Grammatical errors are related to *fluency*, a traditional MT quality evaluation feature. Fluency has been defined as a judgment of "whether the translation reads like good English...without knowing the accuracy of the content," and is typically combined with *adequacy*, an assessment of "the degree to which the information in a professional translation can be found in an MT (or control) output of the same text" (White et al., 1994). A user who cannot understand the source cannot judge adequacy, but may use expectations based on features like fluency and reasonableness to guess. Believability could thus be seen as *predicted adequacy* via a *human cognitive process* with inputs from surface features of the output such as fluency and semantic cues from context.

Two other Computational Linguistics concepts that relate to both reasonableness and grammatical errors may affect believability: acceptability and comprehensibility. In empirical linguistics, acceptability judgments measure users' linguistic competence (Schütze, 2016). While acceptability is primarily used to observe grammatical knowledge, judgments are not limited to grammaticality in practice: "semantic plausibility, various types of processing difficulties, and so on, can individually

or jointly cause grammaticality and acceptability to come apart" (Lau et al., 2017). These breakdowns can lead to issues with comprehensibility. Popović (2020) cites comprehensibility as a key factor in misleadingness of MT output: if a user cannot understand the text, they cannot be misled by it. Similarly, if they cannot understand it, the user is unlikely to believe that the translation is correct.

3 Annotating Believability

To understand the relationships between believability and traditional MT quality criteria (fluency and adequacy), we hired professionals to annotate MT output for these characteristics in tasks based on the fluency and adequacy DA methods of Graham et al. (2013) and Bojar et al. (2016). The annotated data sets are available at: https://github.com/mjmartindale/mt_believability .

Annotators Our annotators were salaried translators with proficiency levels of at least Advanced on the ACTFL scale (ACTFL, 2012) rather than MT researchers or crowd workers as in WMT (Barrault et al., 2019). Because they were not paid per item, they were willing to spend significant time on each item, averaging 15 items in 30 minutes. We believe this reflects more attention to detail as indicated by the correlation between annotators (see Section 4). We note that factors such as foreign language proficiency may affect believability judgments. Further work with a wider variety of annotators is needed to identify and quantify those effects.

Tasks We followed segment-level DA scoring best practices established by WMT (Barrault et al., 2019). The fluency and adequacy questions were taken directly from WMT16 (Bojar et al., 2016). The believability question uses our definition of believability with an introductory phrase to assure the annotator that we understand that it is not possible to truly evaluate the meaning without the source: "Even without having seen the source text, I believe the *meaning* of this translation is likely to match the *meaning* of the original." Annotations were performed using the Turkle[1] annotation platform. Screenshots of the annotation interface are provided in Appendix B.

Long documents were broken up into salient chunks and segments were annotated in their original order to provide discourse context, as in WMT19 "Segment Rating + Document Context" (Barrault et al., 2019). For each chunk, annotators first scored fluency and believability based only on the MT output. They then scored the same segments for adequacy given both source and output.

Annotations For each segment, we calculate a z-score and a label for each feature. We calculate scores following Bojar et al. (2018). Each annotator's raw scores are converted to z-scores based on their own mean and standard deviation, and the z-scores for each segment are averaged across annotators. Segments with positive z-scores are labeled TRUE and negative z-scores are labeled FALSE.

Test Data We chose a test set that is comparable across three typologically different languages with different amounts of training data. Our test data comes from The Multi-Target TED Talks Task (MTTT)–a collection of bitexts across 20 languages (Duh, 2018). The test set is fully sentence parallel with original talk transcripts as the English and human translations for the other languages. We use the non-English translations in MTTT as "source" and machine translate into English. In the test set, there are 29 talks totalling 1,982 segments, however, we exclude one talk ("Nellie McKay sings 'Clonie"') that is too poetic for MT. The final set is 1,976 segments.

MT Systems Because our goal is to examine *segments* annotated for believability, fluency, and adequacy judgments rather than to *compare* systems, we need MT that will produce outputs across a range of quality. Output that is inadequate but believable is of particular interest, so we rely on estimates of the distribution of "fluently inadequate" translations on MTTT from Martindale et al. (2019) to inform our choice of models. They estimated that fluently inadequate translations were most frequent in the "general" NMT models, trained on out-of-domain data. We use their Arabic, Farsi, and Korean "general" models to capture the range of training data sizes and output quality we believe will provide interesting examples for our analysis. The training data is 49M, 6.2M, and 1.4M segments in Arabic, Farsi, and Korean, respectively. The systems are built in Sockeye (Hieber et al., 2017) using the 'SockeyeNMT rm1' settings from the MTTT leaderboard[2]. The resulting systems achieved BLEU (Papineni et al., 2002) scores of 26.6 for Arabic, 22.2 for Farsi and 11.6 for Korean.

[1] https://github.com/hltcoe/turkle

[2] www.cs.jhu.edu/ kevinduh/a/multitarget-tedtalks/

	Arabic			Farsi			Korean		
	FL	BL	AD	FL	BL	AD	FL	BL	AD
Mean Corr.	0.698	0.689	0.728	0.809	0.793	0.830	0.773	0.754	0.793
Std dev	0.137	0.122	0.129	0.087	0.099	0.083	0.112	0.120	0.076

Table 2: Average correlation with the mean for fluency (FL), believability (BL), and adequacy (AD)

	Arabic	**Farsi**	**Korean**	**All**
FL-BL	0.89	0.96	0.97	0.94
BL-AD	0.71	0.74	0.75	0.73
FL-AD	0.62	0.73	0.72	0.69

Table 3: Pearson correlation between fluency (FL), believability (BL), and adequacy (AD).

	%Arabic	**%Farsi**	**%Korean**	**%All**
Fluent	57.9	50.4	49.4	52.5
Believable	61.0	51.2	49.0	53.7
Adequate	59.4	52.4	44.1	52.0
BL+/AD-	25.9	19.4	21.8	22.2

Table 4: Percent of segments with each label (rows 1-3) and percent believable but inadequate (BL+/AD-) segments (row 4).

	Arabic	**Farsi**	**Korean**	**All**
BL+/FL+	92.1	93.8	93.8	93.1
BL-/FL+	8.0	6.2	6.3	6.9
BL+/FL-	18.3	8.0	5.4	10.1
BL-/FL-	81.7	92.1	94.6	89.9

Table 5: Percent of fluent (FL+) and disfluent (FL-) segments that are believable (BL+) or unbelievable (BL-)

4 Quantitative Analysis

Annotation Statistics 63 translators participated in the annotation process. Korean-to-English and Arabic-to-English had the most annotators (26 and 27) and highest number of annotators per segment (10 for Korean and at least 7 for Arabic). There were three annotators whose annotations were deemed unreliable due to low correlation with the mean (< 0.5). Only 10 annotators were available for Farsi, resulting in fewer annotators per segment (median: 4) but more segments per annotator (median: 802), making the z-score process more reliable. After excluding the questionable annotators, we see strong correlation of individual annotator scores with the mean as shown in Table 2. We see similar average correlation with the mean between fluency and believability and small variance across features. This suggests that among our annotators believability is no more subjective than fluency.

Label Distribution Although the distribution of labels is specific to this set of output, it provides context for the other results. The first three rows of Table 4 show the percent of segments with each label. We see that the percent positive examples for each label roughly relates to the system BLEU score, with Arabic having the highest and Korean the lowest.

Feature Relationships Table 3 shows the Pearson correlations between the scores for fluency (FL), believability (BL), and adequacy (AD). The BL-AD relationship is important because inadequate believable translations may mislead monolingual users. The BL-AD correlation is higher than the FL-AD correlation across all languages. This may reflect the influence of context: adequate segments may fit the context well enough to be believable even if not fluent. The same trend is reflected in the fourth row of Table 4, BL+/AD-. Most inadequate translations are not believable, but 19-25% are potentially misleading. Arabic has the highest BL+/AD-. The larger training data may improve both translation and generation, improving overall quality but enabling more believable errors. However, the lower quality Korean also has a higher percent potentially misleading than Farsi. This could mean that all results are idiosyncratic to this data or perhaps the relationship is bimodal.

As expected, FL-BL has the strongest correlation. This indicates that the two features are closely linked, but they are not identical. Table 5 illustrates this distinction. The top two rows show the percent of fluent (FL+) translations that are believable (BL+) and unbelievable (BL-) while the bottom two rows show the percent of disfluent (FL-) transla-

FL	BL	Adequate	Inadequate
✓	✓	I mean, I'm nervous enough.	this is my lady.
✓	✗	I am the emperor of orange. I am emperor of orange.	we need a human body. we should eat it.
✗	✓	there was a fire on 9 days ago.	three days later, this disappeared, and a week later, there was no complaints.
✗	✗	hoping to attract all peoples' minds and be the first to overcome space.	you can see how hard it is to carry kumin ververbert with a bible in 1455.

Table 6: Example output for different segments for each combination of fluency, believability, and adequacy.

tions that are (un)believable. If one were to attempt to identify potentially misleading translations using fluency as a proxy for believability, as in Martindale et al. (2019), 6-8% of translations that are fluent but not believable would be incorrectly labeled as misleading, while nearly 20% of segments from the Arabic system that are believable but not fluent would be missed.

5 Qualitative Analysis

Based on informal examination of a random sample of segments, we find that those labeled as unbelievable often fail semantically, with strange phrases (e.g., "kidney steel", "iron code"), illogical clauses (e.g., "the only natural thing is that my son is the vending machine"), or unlikely argument structure (e.g., "to prove that it seems impossible", "...a state of treason for a raven, which explains that he's cute"). Unbelievable translations may also be grammatical but unintelligible (e.g., "if you want a long time, I'm actually doing something about it."). By contrast, segments labeled as disfluent may include grammatical errors and/or awkward, non-idiomatic phrases such as "he set me a date" or "in a direct time". These observations support our intuition that believability is more influenced by semantic features than fluency is, but further analysis is needed. Additional examples for each combination of fluency, believability, and adequacy are shown in Table 6.

6 Conclusions and Future Work

This work used traditional NLP annotation methods to measure users' perceptions of *believability* of MT output. These methods allow us to identify broad relationships between believability and traditional MT quality metrics, fluency and adequacy, showing that believability is strongly correlated with fluency and somewhat correlated with adequacy. Preliminary qualitative analysis of examples where believability and fluency judgments disagreed suggests that semantic features can overwhelm grammatical features in believability judgments.

A full qualitative analysis of the believability-annotated examples would suggest features that may have influenced annotator's judgments and could indicate approaches that may be effective in automatically predicting believability. Believability used alone could enable an adversarial MT system to deliberately mask errors and produce misleading output, but believability predictions combined with MT quality estimation (Specia et al., 2009) could be used to flag potentially misleading output.

Because believability is a user-centric metric, gaining a complete understanding would require more user-centric methods. The annotator agreement in our results may indicate that believability is less subjective than one might expect, or it may simply indicate that our annotators were fairly homogeneous. A user study could not only tell us exactly what features were most salient in which contexts, but could indicate whether demographic features such as age or education affect perceptions of believability.

References

ACTFL. 2012. ACTFL Proficiency Guidelines 2012.

Loïc Barrault, Ondrej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation*, volume 2, pages 1–61.

Yotam Berger. 2017. Israel Arrests Palestinian Be-

cause Facebook Translated 'Good Morning' to 'Attack Them'. *Haaretz*. [Online; accessed 6-Dec-2017].

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*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Thomas Chesney and Daniel K. S. Su. 2010. The impact of anonymity on weblog credibility. *International Journal of Human-Computer Studies*, 68(10):710–718.

Kevin Duh. 2018. The multitarget ted talks task. http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/.

Andrea Everard and Dennis F. Galletta. 2005. How Presentation Flaws Affect Perceived Site Quality, Trust, and Intention to Purchase from an Online Store. *Journal of Management Information Systems*, 22(3):56–95.

B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and Marissa Treinen. 2001. What makes Web sites credible? a report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 61–68, New York, NY, USA. Association for Computing Machinery.

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 & Interoperability with Discourse*, pages 33–41.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

David Hoffman. 2008. David Hoffman: What happens when you lose everything?

Blooma Mohan John, Alton Yeow-Kuan Chua, and Dion Hoe-Lian Goh. 2011. What Makes a High-Quality User-Generated Answer? *IEEE Internet Computing*, 15(1):66–71.

Prakruthi Karuna, Hemant Purohit, Ozlem Uzuner, Sushil Jajodia, and Rajesh Ganesan. 2018. Enhancing cohesion and coherence of fake text to improve believability for deceiving cyber attackers. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 31–40.

Soojung Kim. 2010. Questioners' credibility judgments of answers in a social question and answer site. *Information Research*, 15(2):15–2.

Soojung Kim and Sanghee Oh. 2009. Users' relevance criteria for evaluating answers in a social Q&A site. *Journal of the American Society for Information Science and Technology*, 60(4):716–727.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There Is No Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation with Distributional Semantics. *Computational Linguistics*, 42(4):637–660.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Ziming Liu. 2004. Perceptions of credibility of scholarly information on the web. *Information Processing & Management*, 40(6):1027–1038.

Marianna J Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243.

Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, 60(3):413–439.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. 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*, Philadelphia, PA.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264.

Soo Young Rieh and David R. Danielson. 2007. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364.

Carson Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Classics in Linguistics. Language Science Press, Berlin.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of Machine Translation systems. In *In EAMT*, pages 28–35.

John S. White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, Maryland, USA.

A Additional Annotation Statistics

Statistics on the number of segments per annotator and annotations per segment are provided in tables 7 and 8.

	Arabic	**Farsi**	**Korean**
Annotators	27	10	26
Min segs	9	47	28
Mean segments	822	760	
Median segs	525	802	400
Max segs	1925	1976	1976

Table 7: Segments annotated per annotator

Annotators	**Arabic**	**Farsi**	**Korean**
2+	1976	1976	1976
4+	1976	1439	1976
8+	1976	48	1976
10	1713	0	1976
Mean Annotations	3	4	10
Median	9	4	10
Max	10	8	10

Table 8: Annotations per segment

B Annotation Interface

Figures 1 and 2 are screenshots of the annotation interface for the monolingual fluency and believability task and the bilingual adequacy task.

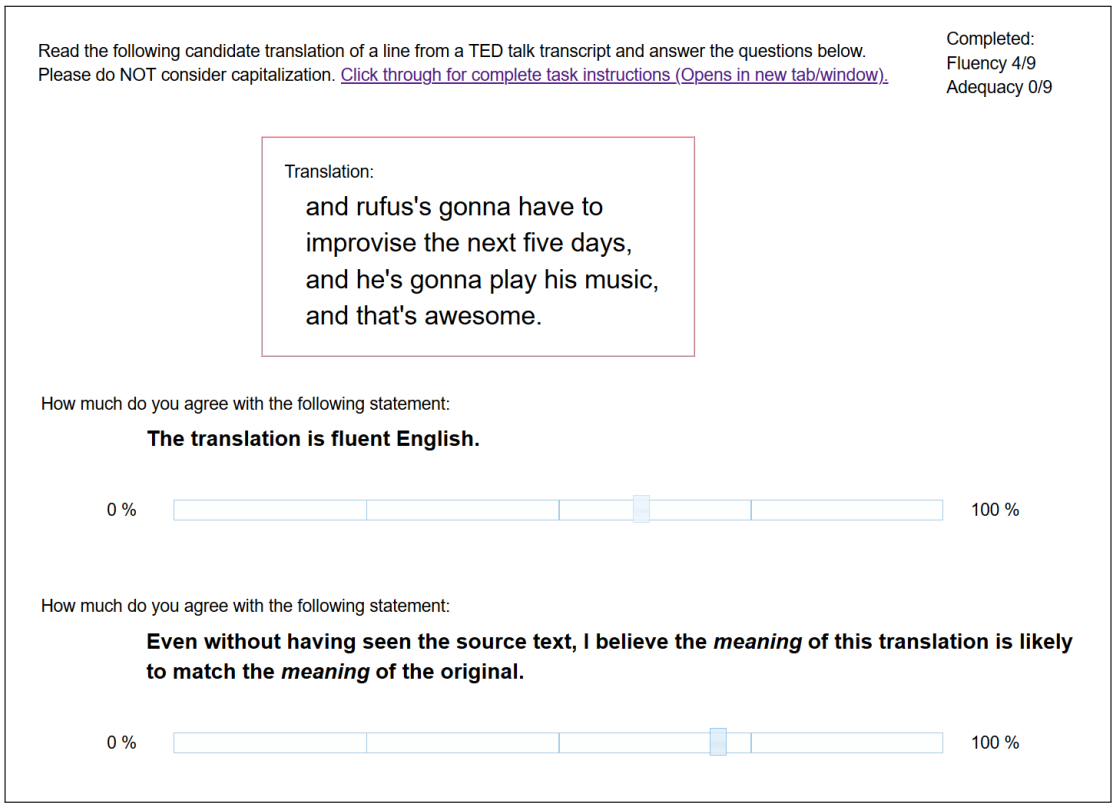

Figure 1: Example question from monolingual annotation phase, fluency and believability

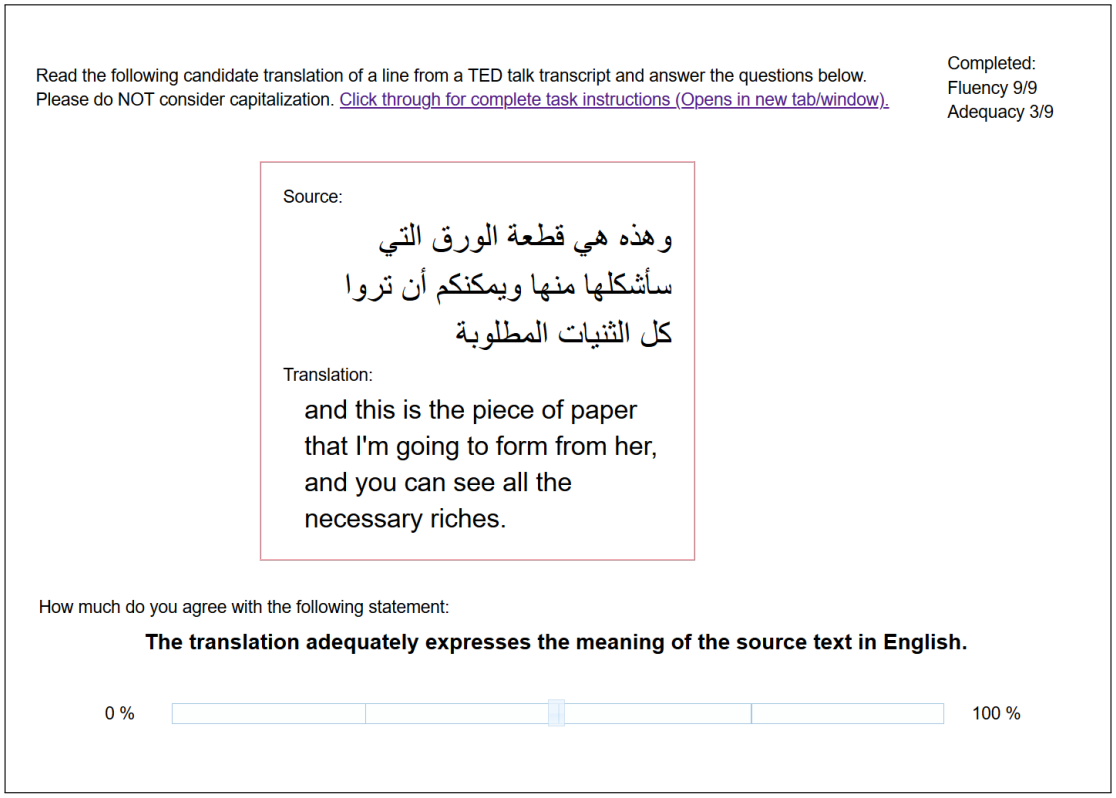

Figure 2: Example question from bilingual annotation phase with adequacy question