It's not a Non-Issue:
Negation as a Source of Error in Machine Translation

Md Mosharaf Hossain,[ꝺ] **Antonios Anastasopoulos,**[ӡ] **Eduardo Blanco,**[ꝺ] and **Alexis Palmer**[?]

[ꝺ]Department of Computer Science and Engineering, University of North Texas
[ӡ]Department of Computer Science, George Mason University
[?]Department of Linguistics, University of North Texas

mdmosharafhossain@my.unt.edu   {eduardo.blanco,alexis.palmer}@unt.edu   antonis@gmu.edu

Abstract

As machine translation (MT) systems progress at a rapid pace, questions of their adequacy linger. In this study we focus on negation, a universal, core property of human language that significantly affects the semantics of an utterance. We investigate whether translating negation is an issue for modern MT systems using 17 translation directions as test bed. Through thorough analysis, we find that indeed the presence of negation can significantly impact downstream quality, in some cases resulting in quality reductions of more than 60%. We also provide a linguistically motivated analysis that directly explains the majority of our findings. We release our annotations and code to replicate our analysis here: https://github.com/mosharafhossain/negation-mt.

1 Introduction

Machine translation (MT) has come a long way, steadily improving the quality of automatic systems, relying mostly on the advent of neural techniques (Goldberg, 2017) and availability of large data collections. As neural MT (NMT) systems start to become ubiquitous, however, one should take note of whether the improvements, as measured by evaluation campaigns such as WMT (Bojar et al., 2017, 2018a, 2019, *et alia*) or IWSLT shared tasks (Cettolo et al., 2017; Niehues et al., 2018, *et alia*) or other established benchmarks, do in fact mean better quality, especially with regards to the semantic adequacy of translations.

We focus specifically on negation, a single yet quite complex phenomenon, which can severely impact the semantic content of an utterance. As NMT systems increasingly improve, to the point of claiming human parity in some high-resource language pairs and in-domain settings (Hassan et al., 2018),[1] fine-grained semantic differences become

increasingly important. Negation in particular, with its property of logical reversal, has the potential to cause loss of (or mis-)information if mistranslated.

Other linguistic phenomena and analysis axes have gathered significant attention in NMT evaluation studies, including anaphora resolution (Hardmeier et al., 2014; Voita et al., 2018) and pronoun translation (Guillou and Hardmeier, 2016), modality (Baker et al., 2012), ellipsis and deixis (Voita et al., 2019), word sense disambiguation (Tang et al., 2018), and morphological competence (Burlot and Yvon, 2017). Nevertheless, the last comprehensive study of the effect of negation in MT pertains to older, phrase-based models (Fancellu and Webber, 2015). In this work, we set out to study the effect of negation in modern NMT systems. Specifically, we explore:

1. Whether negation affects the quality of the produced translations (it does);
2. Whether the typically-used evaluation datasets include a significant amount of negated examples (they don't);
3. Whether different systems quantifiably handle negation differently across different settings (they do); and
4. Whether there is a linguistics-grounded explanation of our findings (there is).

Our conclusion is that indeed negation still poses an issue for NMT systems in several language pairs, an issue which should be tackled in future NMT systems development. Negation should be taken into consideration especially when deploying real-world systems which might produce incredibly fluent but inadequate output (Martindale et al., 2019).

2 Negation 101

Negation at its most straightforward—simple negation of declarative sentences—involves reversing

---

[1]We direct the reader to (Läubli et al., 2018) and (Toral et al., 2018) for further examination of such claims.

the truth value of a sentence. *Skies are blue*, for example, becomes *Skies are not blue*. **Clausal negation** of an existing utterance, defined roughly as negation of the entire clause rather than a subpart, produces a second utterance whose meaning contradicts that of the first.[2] Without stipulations about special contexts (e.g. referring to *skies* in two different places), it is not possible for both utterances to be true. **Constituent negation** involves negation of a non-clausal constituent, such as *happy* in *The universe is not happy with us*. Finally, **lexical negation** refers to two different phenomena. The first are words formed by adding negative affixes, such as *non-starter* or *unhappy*. The second are words with no negative affixes whose meanings nonetheless convey negation, such as the verb *lack* in *This engine lacks power*.

Negation is a core property of human language; every language provides mechanisms for expressing negation, and, cross-linguistically, there are a small number of strategies used for its expression (Dahl, 1979; Miestamo, 2007; Horn, 2010). Following Dryer (2013a), languages may use three different mechanisms for clausal negation: a) **negative particles** (e.g. English *not* or German *nicht*); b) **negative morphemes or affixes** (e.g. Lithuanian *ne-*); and/or c) **negative auxiliary verbs**, (e.g. Finnish *ei*), a strategy in which (generally) inflection moves from the main verb to the auxiliary verb. Depending on the language, constituent and lexical negation may be coded using the same mechanisms as clausal negation, or with entirely different mechanisms. Negative pronouns (e.g. English *nothing* or *nowhere*), negative adverbials (*never*), and negative polarity items (*any*) add additional complexity to the expression of negation.

Even within negation, cross-linguistic variability is immense, ranging from different linguistic strategies, to different structural relationships between linguistic elements, even to differences in interpretation. Szabolcsi and Haddican (2004) show that the interpretation of conjunction, disjunction, and negation can differ cross-linguistically. In English "not X and Y" can be interpreted as either $\neg X \wedge \neg Y$ ("neither X nor Y") or $\neg X \vee \neg Y$ ("not X or not Y") while in Hungarian (and Russian, Italian, etc.) only the first interpretation is allowed. Such differences are potential error sources for translating negation.

Negation-related translation errors. Here we show examples of different types of errors that NMT systems make when translating negation.

1. Negation Omission: The system output completely omits the translation of the negation cue. The following reference translation from the Turkish-English WMT18 shared task reads: "[...] Don't run for public office, if you can't take heat from voters." The best performing system, though, outputs "[...] if you can't take criticism from voters, you're a candidate for state duty," contradicting the reference translation.[3]

2. Negation Reversal: The semantic meaning is reversed, so that the sentence ends up meaning the opposite of the intended meaning, often in fine-grained ways. Consider the following reference translation from WMT19 Lithuanian-English: "The family lawyer of the deceased biker also asks for reversal of the verdict of not guilty." Here the "verdict of not guilty" entails that there was no conviction. The output of the translation system, however, implies there was a conviction to be reversed: "The family lawyer of the dead rider also asks for the conviction to be lifted."[4]

3. Incorrect Negation Scope: The system output makes errors in argument mapping, such that the wrong constituent is negated. Here we look at an example from Finnish-English WMT18. Differences between the reference translation ("The reason is not the Last Judgment") and the best performing system output ("The last judgment is not the reason") lead to differences in interpretation.[5]

4. Mistranslation of Negated Object: When the negated element in the sentence is wrongly translated, meaning is disturbed. This example comes from WMT18 German-English. The reference translation begins "No exchange of personal data occurred [...]," and the system output for the best system begins "There was no exchange of personnel [...]."[6] Aside from the mistranslated object noun phrase, the meaning is intact.

These error types vary in their severity, but each has the potential to completely change the meaning conveyed by the translation.

---

[2]See Penka (2015) and Horn (2010), among others, for more on the semantics, syntax, and interpretation of negation.

[3]The source sentence in Turkish is: "Eğer seçmenlerin eleştirilerini kaldıramıyorsan devlet görevine aday olma."

[4]The Lithuanian source is: "Panaikinti išteisinamąjį nuosprendį prašo ir žuvusio motociklininko šeimos advokatė."

[5]Finnish source sentence: "Viimeinen tuomio ei ole syy."

[6]The source sentence in German is: "Zu einem Personalienaustausch kam es aber nicht, da der 75-Jährige die Dame auf dem Parkplatz nicht mehr finden konnte."

3 Experimental Setup

We follow the setup of the Conference on Machine Translation (WMT), in particular, the 2018 and 2019 Shared Tasks (Bojar et al., 2018b; Barrault et al., 2019). We compare reference and system translations using normalized direct assessments. Direct assessments are scores from 0 to 100 provided by researchers or crowd workers. In order to account for the hundreds of human annotators with potentially different criteria, raw direct assessments are normalized using the mean and standard deviation of each annotator. The normalized direct assessment is the average of the sentence-level direct assessments. We refer to normalized direct assessment with Z-score or simply Z.

Z-scores are the official ranking criterion in WMT competitions, and are preferred to automated metrics to assess the quality of MT. Nevertheless, most of the MT community still relies on automated metrics for development and system comparisons. Thus, we also work with three automated metrics, in particular BLEU (Post, 2018), chrF++ (Popović, 2017), and METEOR (Denkowski and Lavie, 2011).

In the remainder of the paper we present two complementary analyses. First, we investigate the role of negation in machine translation with an emphasis on numeric evaluation (Section 4). Second, we investigate from a linguistic perspective what makes translating negation difficult (Section 5).

Datasets. We work with all submissions to the news translation tasks in the WMT18 and WMT19 competitions. Table 2 shows the language directions in each competition along with the number of sentences with Z-scores in the corresponding test set. Specifically, we investigate systems for both translation directions between English and Russian (RU), Estonian (ET), German (DE), Turkish (TR), Finnish (FI), Czech (CS), and Chinese (ZH). We also use Lithuanian (LT), Gujarati (GU), and Kazakh (KK) to English systems from WMT 19.

Negation Detection. Due to the lack of reliable negation detection systems in most languages, our study is limited to focusing on cases where negation is present in English. This creates slightly different settings for translation into and out of English, hence we distinguish them in our analyses and presentation of the results. For translation out of English, we detect negation cues in the source sentence. For translation into English, we detect negation cues in the reference translations.

In order to detect negation automatically, we train a negation cue detector for English using a Bi-LSTM neural network with a CRF layer, as described in (Hossain et al., 2020a). Trained and evaluated with a publicly available corpus (Morante and Blanco, 2012), it obtains 0.92 F1. The cue detector recognizes single-token cues (*not*, *n't*, *never*, *no*, *nothing*, *nobody*, etc.) as well as affixal cues (*im*possible, *dis*agree, fear*less*, etc.). The supplemental materials provide details about the architecture and input representation of the cue detector.

Important note. We make the strong assumption in our analyses that presence of a negation cue in the English reference translation indicates presence of negation in the source sentence. We acknowledge that investigating the role of negation in machine translation by only looking at English negations likely misses valid insights. For example, a Spanish sentence containing negation (e.g., *"El ladrón no estaba preocupado hasta que vino la policía"*) can be translated into English either with negation (*"The thief was not worried until the police arrived"*), or without negation (*"The thief only worried when the police arrived"*). We reserve for future work a more thorough analysis of correspondences between negation in source sentences and negation in English reference translations.

4 Quantitative Analysis

We conduct a thorough quantitative analysis of the WMT18 and WMT19 submissions around six questions investigating the role of negation in MT.

Q0: Is Negation Present in Evaluation Datasets? We found that 9.6–20% of source sentences contain negation depending on the target language. These percentages are lower than what we observe in online reviews (23–29%), books (29%) and conversation transcripts (27–30%) (Hossain et al., 2020b), and are also lower than previous reports (Morante and Sporleder, 2012).

Q1: Is Translating Sentences with and without Negation Equally Hard? Table 1 shows the Z-scores obtained by the best submission in each translation direction using all sentences, sentences with negation, and sentences without negation. Many language pairs obtain substantially worse Z-scores for sentences containing negation: Turkish, Russian, and Estonian in WMT18 (20.0–68.9% lower), and Lithuanian, Gujarati, and Russian in

| Translation *into* English | | | | |
| WMT18 | all | w/ neg. | | w/o neg. | |
	Z	Z	%Δ	Z	%Δ
ru→en	0.215	0.132	(−38.6)	0.230	(+7.0)
de→en	0.413	0.376	(−9.0)	0.421	(+1.9)
et→en	0.326	0.232	(−28.8)	0.343	(+5.2)
tr→en	0.045	0.014	(−68.9)	0.051	(+13.3)
fi→en	0.153	0.223	(+45.8)	0.139	(−9.2)
cs→en	0.298	0.333	(+11.7)	0.290	(−2.7)
zh→en	0.140	0.162	(+15.7)	0.137	(−2.1)
WMT19					
ru→en	0.156	0.123	(−21.2)	0.161	(+3.2)
de→en	0.146	0.136	(−6.9)	0.147	(+0.7)
fi→en	0.285	0.306	(+7.4)	0.283	(−0.7)
lt→en	0.234	0.093	(−60.3)	0.262	(+12.0)
gu→en	0.210	0.112	(−46.7)	0.221	(+5.2)
kk→en	0.270	0.326	(+20.7)	0.264	(−2.2)

| Translation *from* English | | | | | |
| WMT18 | all | w/ neg. | | w/o neg. | |
	Z	Z	%Δ	Z	%Δ
en→ru	0.352	0.245	(−30.4)	0.371	(+5.4)
en→de	0.653	0.689	(+5.5)	0.646	(−1.1)
en→et	0.549	0.439	(−20.0)	0.569	(+3.6)
en→tr	0.277	0.094	(−66.1)	0.308	(+11.2)
en→fi	0.521	0.569	(+9.2)	0.512	(−1.7)
en→cs	0.594	0.574	(−3.4)	0.599	(+0.8)
en→zh	0.219	0.229	(+4.6)	0.218	(−0.5)

Table 1: Evaluation of the best WMT18 and WMT19 submissions (normalized direct assessments, Z) for each language direction using all sentences, sentences with negation (w/ neg.), and sentences without negation (w/o neg.). Translating sentences with negation obtains worse results in many languages (e.g., Turkish, Russian, Estonian, Lithuanian, Gujarati).

WMT19 (21.2–60.3% lower). When both translation directions are available for a language pair (WMT18), translating negation into English from these languages consistently receives lower scores than translating from English.

Interestingly, sentences with negation receive better Z-scores in Finnish, Chinese, and Kazakh (4.6–45.8% higher). Finally, only two languages show opposite trends translating from and into English negations: German and Czech, although the differences in Z-scores are smaller (-3.4–11.7%).

Naturally, other factors beyond the presence of negation can affect the results. Since sentence length tends to negatively correlate with translation quality, the length difference between sentences with and without negation could be an important factor. Sentences with negation are on average

longer than sentences without negation,[7] but we do not consider the differences (typically within 2-6 words) to be significant. Nevertheless, we replicated our analysis from Table 1 using only the sentences that fall within a standard deviation of the mean sentence length $[\mu \pm \sigma]$ for each dataset. The results, shown in Table 6 in the Appendix, do not differ significantly from our Table 1 analysis. We identify one major inconsistency, in the case of English-Finnish translation direction; we attribute it to issues in identifying negation in that dataset, and we leave further analysis as future work. In any case, performing the same analysis on sentences that fall outside that bucket (that is, sentences shorter or longer than one standard deviation of the average $(0, \mu - \sigma)$ and $(\mu + \sigma, +\infty)$, respectively, available in Tables 7 and 8 in the Appendix) does not yield conclusive results. We attribute this to the fact that these buckets include fewer data samples and too many outliers (of very easy or very hard sentences).

In any case, the consistency of the results in the "average" case leads us to conclude: a) that **negation affects Z-scores in all language pairs and directions**, and b) **that translating from or into English sentences containing negation in most language directions is harder than sentences without negation.**

Q2: Does Negation Affect Rankings? Just because the best system in each language direction obtains better or worse Z-scores (Q1), it is not necessarily the case that *all* systems do better or worse. In order to check if rankings are affected, we calculate the correlation between rankings obtained with Z-scores and (a) sentences with negation and (b) sentences without negation (outlined in Table 2). We use Kendall's τ coefficient (Kendall, 1938), which only considers the ranking—not the differences in Z-scores. τ coefficients range from -1 to 1 (absolute negative and positive correlation), and $\tau = 0$ indicates no correlation at all.

We observe that the Z-score rankings for all sentences and sentences without negation are very close in all language directions: $\tau \in (0.8, 1.0)$, and all but two (tr→en and lt→en) are above 0.9. The rankings obtained based on sentences with negation, on the other hand, show lower τ correlations, except language pairs involving German and Czech.

---

[7]This is expected if one considers the additional tokens required, such as the negation cue, or other potential syntactic changes (e.g. adding an auxiliary verb).

Translation *into* English			
WMT18	w/ neg.	w/o neg.	w/ v. w/o
ru→en (1,820)	0.786	1.000	0.786
de→en (2,121)	0.933	0.983	0.917
et→en (1,593)	0.729	0.972	0.692
tr→en (2,678)	<u>0.738</u>	<u>0.800</u>	<u>0.527</u>
fi→en (1,769)	0.833	0.944	0.778
cs→en (1,849)	1.000	1.000	1.000
zh→en (2,039)	0.495	0.906	0.398
WMT19			
ru→en (1,692)	0.575	0.912	0.486
de→en (2,000)	0.723	0.941	0.650
fi→en (1,548)	0.879	0.970	0.848
lt→en (1,000)	0.855	0.871	0.722
gu→en (1,016)	0.745	1.000	0.745
kk→en (1,000)	0.855	1.000	0.855
Translation *from* English			
WMT18	w/ neg.	w/o neg.	w/ v. w/o
en→ru (2,076)	0.944	1.000	0.944
en→de (759)	0.917	0.983	0.900
en→et (1,019)	0.978	1.000	0.978
en→tr (420)	0.643	0.929	<u>0.571</u>
en→fi (760)	0.848	0.970	0.818
en→cs (1,594)	1.000	1.000	1.000
en→zh (1,876)	0.758	0.956	0.714

Table 2: Comparison of rankings (Kendall's τ) of all submissions to WMT18 and WMT19 using (a) all sentences and those with negation (w/ neg.), (b) all sentences and those without negation (w/o neg.), and (c) sentences with and without negation (w/ v. w/o). The numbers between parentheses indicate the number of sentences with human scores for the best system. All τ coefficients are statistically significant ($p < 0.05$) except those that are underlined (null hypothesis: there is no association between the rankings). We note, however, that the τ coefficients are substantially lower in most language directions when negation is present.

Except those involving German and Chinese, all language pairs have at least one translation direction with $\tau \leq 0.85$. We also observe that the rankings change much more (lower τ coefficients) translating from Chinese, Estonian, and Russian into English sentences containing negation than translating from English sentences containing negation into those languages.

It is worth noting that the relative drop in Z-score of the best system when negation is present (%Δ, Table 1) is not a good predictor of ranking changes. For example, all submissions translating from English to Russian obtain proportionally worse Z-scores when negation is present in the source text, thus the ranking barely changes (τ=0.944) despite
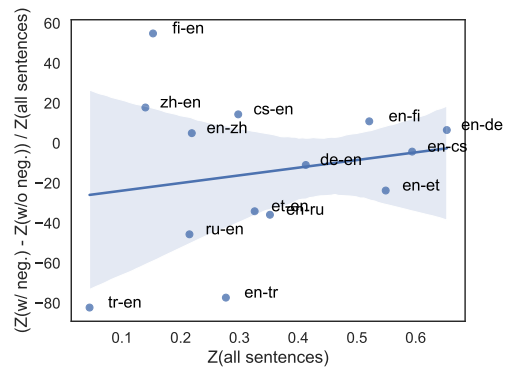

Figure 1: Correlation between the Z-score with all sentences and the *relative* drop for the sentences with negation (best system per language pair).

%Δ=−30.4. On the other hand, we observe many ranking changes translating from Chinese to English (τ=0.495) and vice versa (τ=0.758), despite %Δ of only 15.7 and 4.6.

In Table 3, we show the ranking of submissions translating from Russian to English obtained with all sentences (official WMT ranking), and the rankings obtained with sentences containing and not containing negation. The ranking changes in WMT18 (↑ and ↓ arrows) illustrate the τ correlation coefficients obtained with sentences with negation (many changes, τ=0.575) and without negation (few changes, τ=0.912). The supplementary materials contain similar tables for selected language pairs in WMT18 and WMT 19.

We conclude that **rankings based on sentences containing negation are substantially different for most translation directions. Thus, different systems behave differently translating from and into English sentences containing negation.**

Q3: Is Translating Negation Harder with Harder Language Pairs and Directions? Not all translation directions are equally easy to model, as evidenced by the wide variance of direct assessment scores in the WMT competitions (Table 1). Figure 1 shows that there is is a weak positive correlation between the *relative* differences in Z-score with negation and the overall Z-score. There are however many exceptions/outliers, e.g., translations from Finnish or Russian into English receive roughly the same Z-scores, but negation is much harder from Russian than from Finnish. We conclude that **the difficulty of translating between two languages is only a weak indicator of how difficult it is to translate negation, most notable with overall Z-scores below 0.3.**

		All sentences				Sentences w/ negation				Sentences w/o negation	
		System	Z			System	Z			System	Z
WMT18	1	Alibaba	0.215	1	↑1	online-B	0.172	1	–	Alibaba	0.230
	2	online-B	0.192	2	↑1	online-G	0.155	2	–	online-B	0.196
	3	online-G	0.170	3	↓2	Alibaba	0.132	3	–	online-G	0.173
	4	uedin	0.110	4	–	uedin	0.084	4	–	uedin	0.115
	5	online-A	0.034	5	–	online-A	0.009	5	–	online-A	0.039
	6	afrl	-0.014	6	↑1	JHU	-0.044	6	–	afrl	0.012
	7	JHU	-0.027	7	↓1	afrl	-0.154	7	–	JHU	-0.024
WMT19	1	FB-FAIR	0.156	1	–	FB-FAIR	0.123	1	–	FB-FAIR	0.161
	2	online-G	0.134	2	↑2	online-B	0.091	2	–	online-G	0.143
	3	eTranslation	0.122	3	↑6	online-A	0.077	3	–	eTranslation	0.134
	4	online-B	0.121	4	↓2	online-G	0.073	4	↑2	MSRA.SCA	0.133
	5	NEU	0.115	5	↓2	eTranslation	0.044	5	–	NEU	0.126
	6	MSRA.SCA	0.102	6	↓1	NEU	0.040	6	↓2	online-B	0.125
	7	rerank-re	0.084	7	↑1	online-Y	-0.023	7	–	rerank-re	0.108
	8	online-Y	0.076	8	↑4	TartuNLP-u	-0.046	8	–	online-Y	0.091
	9	online-A	0.029	9	↑1	afrl-sys	-0.066	9	↑1	afrl-sys	0.024
	10	afrl-sys	0.012	10	↓3	rerank-re	-0.072	10	↓1	online-A	0.022
	11	afrl-ewc	-0.039	11	–	afrl-ewc	-0.101	11	–	afrl-ewc	-0.030
	12	TartuNLP-u	-0.040	12	↑1	online-X	-0.101	12	–	TartuNLP-u	-0.039
	13	online-X	-0.097	13	↓7	MSRA.SCA	-0.102	13	–	online-X.0	-0.096

Table 3: Rankings of all submissions translating from Russian to English using all sentences (official WMT ranking), sentences with negation, and sentences without negation. FB-FAIR, afrl, and afrl-sys refer to Facebook-FAIR, afrl-ruen-syscomb, and afrl-syscomb19. We use ↑k (↓k) to indicate the gains (losses) in absolute ranking with respect to the ranking obtained with all sentences. For example, in WMT19, MSRA.SCA is 7 positions lower in the ranking obtained with sentences with negation (from 6th to 13th), and online.A is 6 positions higher (from 9th to 3rd). As the τ coefficients indicate (Table 2), there are barely any changes in the ranking obtained with sentences without negation, but many changes in the ranking obtained with sentences with negation.

Q4: Is Translating Negation between Similar Languages Easier? Intuition may lead us to believe that it is easier to translate negation between similar languages. We show the correlation between language similarity and relative differences in Z-scores with and without negation in Figure 2. To calculate similarity between two languages, we follow Zhang and Toral (2019) and Berzak et al. (2017). Briefly, we obtain feature vectors for each language from lang2vec (Littell et al., 2017), and define the similarity between two languages as the cosine similarity between their feature vectors. More specifically, we concatenate 103 morphosyntactic features and 87 language family features (only those relevant to the languages we work with) from the URIEL typological database.

We conclude that **similarity between languages is only a weak indicator of how difficult it is to translate negation**. We revisit this question in Section 5 with an in-depth linguistic discussion.
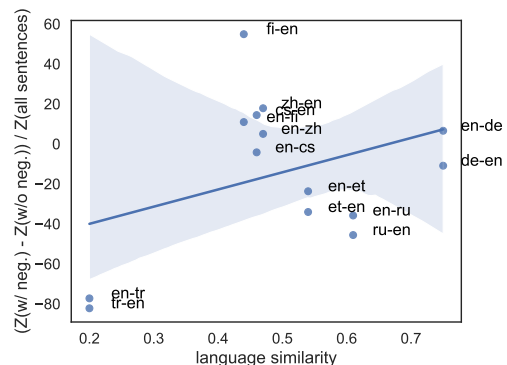

Figure 2: Correlation between the language similarity and the *relative* drop for the sentences with negation.

Q5: Are Automatic Metrics Worse with Negation? Machine translation evaluation is an active field of research, and new automatic metrics are proposed yearly (Fomicheva and Specia, 2019, *et alia*). The ideal metric would correlate perfectly with human judgments, and increased correlation is often a justification for new metrics (Lavie and Agarwal, 2007). We investigate whether three popular metrics (BLEU, chrF++ and METEOR) are

	Z-score v.							Z-score v.					
	BLEU		chrF++		METEOR			BLEU		chrF++		METEOR	
	w/	w/o	w/	w/o	w/	w/o		w/	w/o	w/	w/o	w/	w/o
ru→en	0.64	0.79	0.78	0.86	0.79	0.79	en→ru	0.83	0.94	0.83	0.94	0.83	0.94
de→en	0.80	0.85	0.83	0.92	0.78	0.88	en→de	0.78	0.75	0.78	0.77	0.75	0.77
et→en	0.78	0.83	0.71	0.85	0.71	0.85	en→et	0.76	0.85	0.89	0.89	0.82	0.93
tr→en	<u>-0.32</u>	<u>-0.60</u>	<u>-0.74</u>	<u>-0.20</u>	<u>-0.53</u>	<u>0.00</u>	en→tr	0.79	<u>0.57</u>	0.64	0.86	0.64	0.79
fi→en	0.78	0.78	0.78	0.94	0.72	0.94	en→fi	0.85	0.82	0.67	0.94	0.79	0.91
cs→en	<u>0.60</u>	<u>0.80</u>	<u>0.80</u>	<u>0.80</u>	1.00	<u>0.80</u>	en→cs	1.00	1.00	1.00	1.00	1.00	1.00
zh→en	0.47	0.80	0.52	0.73	0.47	0.82	en→zh	0.41	<u>0.39</u>	0.69	0.65	<u>-0.16</u>	<u>0.32</u>

Table 4: Comparison of WMT18 rankings (Kendall's τ) using several metrics, and sentences with and without negation. We compare the official scoring (normalized direct assessments, Z-score) and three automated metrics: BLEU, chrF++ and METEOR. We provide the same analysis for WMT19 in the supplementary materials. All τ coefficients are statistically significant ($p < 0.05$) except those that are underlined (null hypothesis: there is no association between the Z-score and the automatic metric). The differences in the rankings obtained with the Z-score and the three metrics are more substantial when negation is present (lower τ coefficients).

equally suitable when applied to sentences with and without negation. Note that the smallest change when negation is present (e.g., dropping *never* or *n't*) is likely to result in a (very) bad translation. Table 4 shows the correlations (Kendall's τ) between the rankings on WMT 18, obtained with the Z-scores and the three metrics using (a) sentences with negation, and (b) sentences without negation.

When negation is present, the three metrics obtain worse correlation coefficients or just slightly better (within 5%). This is true with all translation directions except those involving Finnish or Turkish. The drops in correlation coefficients are substantial with the three metrics in many translation directions, e.g., from Russian, Chinese or Gujarati into English. While there is no winner across all language directions (e.g., chrF++ is better with Russian, and METEOR with Czech), the correlation coefficients show, unsurprisingly, that BLEU is the least suited to evaluate translation quality when negation is present.

We conclude that **automatic metrics are bad estimators of machine translation quality when negation is present, and that chrF++ and METEOR are better suited than BLEU.**

5 So what's going wrong?

We have shown that translations with negation receive lower scores on average than translations without negation, for most languages. We have also shown that this is true regardless of a system's overall performance. To better understand *why* negation is challenging for NMT systems, we look into the properties of negation in individual languages, as well as the particular errors made by the systems. For this analysis, we restrict ourselves to 5 language pairs, mostly selecting languages with substantial negation-related performance differences for one or more of the quantitative analyses (Turkish, Lithuanian, Russian, and Finnish). We also consider German, as English-German/German-English is one of the most commonly used translation benchmarks.

5.1 Typological perspective on negation

The World Atlas of Linguistic Structures (WALS Dryer and Haspelmath, 2013) is a typological database of the world's languages. The core of WALS is a partially-filled grid of typological features vs. languages. For example, WALS identifies 5 broad categories of features related to the realization of negation. The left side of table 5 shows the relevant feature values for our languages.[8]

(i.) Clausal: Form of simple clausal negation. This feature captures the morphological form of clausal negation for declarative sentences (Dryer, 2013a) and has three possible values. **Particle** languages use a negation word (e.g. *not* in EN or *nicht* in DE), **Affix** languages use a negative affix (e.g. the prefix *ne-* in LT), and **Aux** languages use a negative auxiliary verb. Aux languages are the rarest; only 47 out of 1157 languages with a value for this property in WALS use a negative auxiliary verb. Finnish is one of these. In FI the negative

---

[8]We leave out prohibitives (van der Auwera et al., 2013), as this construction is infrequent and does not affect our analysis.

	(a) Typological properties					(b) Negation-related translation errors		
Language	Clausal	PredNeg	Symm	Order	Sim-EN	Direction	neg.sents	%neg.errors
Turkish	Affix	Yes	Both	[V-Neg]	1.0	tr → en	460	17.2%
Lithuanian	Affix	Yes	–	[Neg-V]	1.0	lt → en	164	11.0%
Russian	Particle	Yes	Symm	NegV	2.5	ru → en	218	8.3%
German	Particle	No	Symm	NegV/VNeg	3.0	de → en	281	7.5%
Finnish	Aux.	Yes	Asymm	NegV	1.5	fi → en	185	2.7%
English	Particle	No	Both	NegV				

Table 5: (a) Typological properties related to negation (left), and (b) Error analysis of negation-containing test set sentences (right) for five non-English WMT languages. With the exception of Finnish, languages more similar to English with respect to negation show fewer negation-related errors in translation. We discuss Finnish below.

verb (underlined) inflects to agree with the subject, as for example: *Linja-auton kuljettajaa ei epäillä rikoksesta*, "The bus driver is not suspected of a crime." This means that clausal negation in Finnish declarative sentences is always expressed with the same verb, and the form of the verb relies only on the person and number of the subject; other aspects of the sentence do not change the form of the verb.

(ii.) PredNeg: Presence of predicate negation with negative indefinites. In some languages, negative indefinite pronouns (e.g. *nothing, nowhere*) require an additional particle negating the predicate (Haspelmath, 2013). Russian is an example of a language with the predicate negation requirement: Сергей Сироткин: Я ничего не мог сделать [...] , "Sergey Sirotkin: There was nothing I could do [...]." ничего translates as *nothing* in English, and не is the predicate negation particle. Most varieties of English do not use a predicate negation particle, though some do, leading to a value of 'No/Mixed' for this feature in WALS. Because our systems all translate from and into a mainstream variety of English, we use 'No' for this analysis.

(iii.) Symm: Symmetricity of negation. Miestamo (2013) characterizes languages as symmetric or asymmetric with respect to simple clausal negation. In symmetric languages (e.g. DE and RU), negated clauses are the same as their non-negated counterparts, with the exception of the negation word. In asymmetric languages (e.g. Finnish), presence of the negation word triggers other grammatical changes in the sentence. For FI, the negative auxiliary is inflected and the main verb changes to an uninflected form. Some languages (like English) show both behaviors depending on context.

(iv.) Order: Order of negation cues and other constituents. Various aspects of the ordering of negation cues in the clause are addressed in two sections of WALS (Dryer, 2013b,c). Here we look only at the relative ordering of the negation cue and the verb. NegV (negation cue before the verb), the ordering seen in EN, RU, and FI, is the most common (525/1325 languages sampled). LT also shows NegV ordering, with the difference that the negation cue is an affix rather than a particle (indicated by square brackets and hyphen: [Neg-V]). TR shows the opposite ordering, and German allows both NegV and VNeg.

Similarity to English. For this analysis, we assign each language a score for its similarity to English with respect to typological properties of negation, based on WALS data. English is the comparison language because it is the common language in all translation directions. For each of the four WALS properties, a language scores one point for a feature with the same value as English, and a half point for a partial match. For example, Russian gets a half-point for the Symm feature.

5.2 Are errors due to negation?

In a next step, we look closely at negation-containing sentences to determine how often systems get the negation wrong. For these five language pairs, we extract all test-set sentences with negation cues detected in the reference translations; the number of sentences per language pair is shown in Table 5. After sorting sentences according to Z-score, we compare reference translation and system output and annotate (Yes/No) whether the system gets negation wrong, compared to the reference. In contrast to Fancellu and Webber (2015), who do fine-grained annotation of translation errors related to negation (focusing on deletion/insertion/substitution of cues, focus, and scope), we ask a broader question designed to capture semantic adequacy focused on handling of

negation. We only choose **Yes** if the system output shows a glaring error in meaning *related to negation* (see Section 2 for discussion of some typical error types).

The percentage of sentences that contain negation with negation-related errors (Table 5, right side) ranges from only 2.7% in Finnish up to more than 17% in Turkish. The core finding from this analysis is that (except for Finnish) **the languages with the fewest negation-related errors are most similar to English with respect to the typology of negation.**

Turkish differs from English on three of four WALS features. In addition, clausal negation in Turkish occurs through a negative affix attached to the verb root. The complexity of Turkish verbal morphology means that: a) the negative morpheme undergoes changes in form depending on its context; and b) the negative morpheme is tucked away in the interior of the verb word, between the stem and both tense-aspect-mood markers and person agreement morphology (Emeksiz, 2010). In *Ben seni unut-ma-di-m,* ("I have not forgotten you") (Emeksiz, 2010), the negative morpheme appears as *-ma*. Affixal clausal negation is not unusual from a typological perspective, but the morphological richness of Turkish makes it particularly difficult to recognize negated clauses.

Lithuanian, Russian, and German each differ from English in their values for 1-2 features. Interestingly, these languages occupy a sort of middle ground for the percentage of negation errors seen.

Finnish seems to be a special case. Though it differs from English on 2-3 feature values (2.5 in our scoring system), we see a very low proportion of negation-related errors in system output translations. We attribute this to the negative auxiliary, a way of expressing negation that is easy to identify, even for non-speakers of the language (and presumably also for NMT systems). 175 out of 185 source sentences contain at least one easily-identifiable negation cue: a conjugated form of the negative auxiliary, prefixal negation on adjectives, a negative conjunction ($\sim$ EN *lest, neither*), or a negative preposition ($\sim$ EN *without*). We hypothesize that the clarity and detectibility of source-side negation cues improve the quality of NMT systems when translating negation.

Other observations. In addition to the main finding above, we notice that, for some languages, certain negation cues are either more or less likely to appear in sentences with negation errors. For example, in German, negation errors are most likely when the cue is *nicht*-V (negation particle modifying the verb); this is also the most frequent cue overall. Of 46 source sentences containing the negative indefinite article *kein*, only one triggers a negation error. We have performed this analysis only for languages where we can reliably (manually) identify negation cues on the source side; we hope to extend to more languages in the future.

Figurative expressions, identified either on the source side or judging by the reference translation, often contain negation errors. We also see recurring problems with the interaction of negation and the translation of certain temporal expressions, and occasional problems with negation errors in the reference translations (Freitag et al., 2020). Encouragingly, errors of outright contradiction between the reference translation and system output are rare.

6 Conclusions

We show, through both quantitative and qualitative analysis, that negation remains problematic for modern NMT systems. Though negation is ubiquitous and universal in its semantic effect, its realization varies tremendously from language to language. Typological similarity with respect to negation seems to correspond to better translation of negation, at least for the language pairs we investigate. Looking forward, we propose to harness linguistic insights about particular languages to better translate negation, and to devise fine-grained evaluation metrics to capture the adequacy of negation-involving translations.

References

Johan van der Auwera, Ludo Lejeune, and Valentin Goussev. 2013. The prohibitive. In Matthew S. Dryer

and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kathryn Baker, Michael Bloodgood, Bonnie J Dorr, Chris Callison-Burch, Nathaniel W Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Modality and negation in simt use of modality and negation in semantically-informed syntactic mt. *Computational Linguistics*, 38(2):411–438.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017. *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors. 2018a. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018b. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Östen Dahl. 1979. Typology of sentence negation. *Linguistics*, 17(1-2):79–106.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Matthew S. Dryer. 2013a. Negative morphemes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Order of negative morpheme and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013c. Position of negative morpheme with respect to subject, object, and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Zeynep Erk Emeksiz. 2010. Negation in turkish. *Dilbilim Araştırmaları*, 2:1–16.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado. Association for Computational Linguistics.

Marina Fomicheva and Lucia Specia. 2019. Taking mt evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. arXiv:2004.06063.

Yoav Goldberg. 2017. *Neural network methods in natural language processing*. Morgan & Claypool Publishers.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2014. Translating pronouns with latent anaphora resolution. In *Workshop on Modern Machine Learning and Natural Language Processing (NIPS 2014), Montreal, Canada, 8-12 Dec 2014*.

Martin Haspelmath. 2013. Negative indefinite pronouns and predicate negation. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7.

Laurence R Horn. 2010. *The expression of negation*, volume 4. Walter de Gruyter.

Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. 2020a. Predicting the focus of negation: Model and error analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8389–8401, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020b. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Marianna 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, Dublin, Ireland. European Association for Machine Translation.

Matti Miestamo. 2007. Negation–an overview of typological research. *Language and linguistics compass*, 1(5):552–570.

Matti Miestamo. 2013. Symmetric and asymmetric standard negation. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Jan Niehues, Ronaldo Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *Proceedings of IWSLT*.

Doris Penka. 2015. Negation and polarity. *The Routledge Handbook of Semantics*, pages 303–319.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*.

Anna Szabolcsi and Bill Haddican. 2004. Conjunction meets negation: A study in cross-linguistic variation. *Journal of Semantics*, 21(3):219–249.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Identifying Negations

In order to identify negations in English sentences (in source sentences when the translation direction is English to foreign, otherwise in reference translations and system outputs), we develop a negation cue detector that consists of a two-layer Bidirectional Long Short-Term Memory network with a Conditional Random Field layer on top (BiLSTM-CRF). This architecture (Figure 3) is similar to the one proposed by Reimers and Gurevych (2017). We train and evaluate the model with CD-SCO, a corpus of Conan Doyle stories with negation annotations (Morante and Blanco, 2012). CD-SCO includes common negation cues (e.g., never, no, n't), as well as prefixal (e.g., impossible, unbelievable) and suffixal negation (e.g., motionless).

We map each token in the input sentence to its 300-dimensional pre-trained GloVe embedding (Pennington et al., 2014). In addition, we extract token-level universal POS tags using spaCy (Honnibal and Montani, 2017) and leverage another embedding (300-dimensional) to encode them. Embedding weights for universal POS are learned from scratch as part of the training of the network. We concatenate the word and POS embeddings, and feed them to the BILSTM-CRF architecture (size of cell state: 200 units). The representations learnt by the 2-layer BiLSTM are fed to a fully connected layer with ReLU activation function (Nair and Hinton, 2010). Finally, the CRF layer yields the final output.

We use the following labels to indicate whether a token is a negation cue: S_C (single-token negation cue, e.g., never, not), P_C (prefixal negation, e.g., inconsistent), SF_C (suffixal negation, e.g., emotionless), and N_C (not a cue).

Training details. We merge the train and development instances from CD-SCO, and use 85% of the result as training and the remaining 15% as development. We evaluate our cue detector with the original test split from CD-SCO. We use the stochastic gradient descent algorithm with RMSProp optimizer (Tieleman and Hinton, 2012) for tuning weights. We set the batch size to 32, and the dropout and recurrent dropout are set to 30% for the LSTM layers. We stop the training process after the accuracy in the development split does not increase for 20 epochs, and the final model is the one which yields the highest accuracy in the development accuracy during the training process (not necessarily the model from the last epoch). Evaluating with
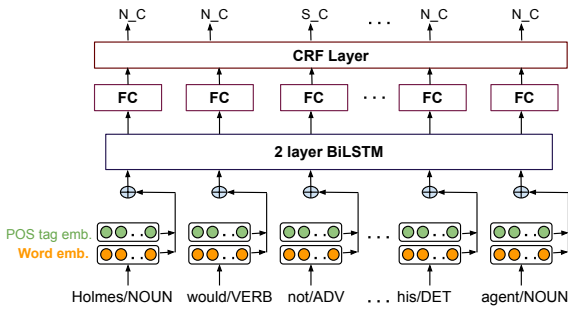
Figure 3: The BiLSTM-CRF model architecture to identify negation cues. The input is a sentence, where each token is the actual word and its universal POS tag. The model outputs a sequence of labels indication of negation presence. The example input sentence is "Holmes/NOUN would/VERB not/ADV listen/VERB to/ADP such/ADJ fancies/NOUN ,/PUNCT and/CCONJ I/PRON am/VERB his/DET agent/NOUN."

the test set yields the following results: 92.75 Precision, 92.05. Recall, and 92.40 F1. While not perfect, the output of the cue detector is reliable, and an automatic detector is the only way to count negations in large corpora. The code is available at https://github.com/mosharafhossain/negation-cue.

Note that our cue detector model has nearly 4.3 million parameters and takes 30 minutes on average to train on a CPU machine (Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz) with 64 GB of RAM.

B Impact of Sentence Length

We present results on the sentence length bucket analysis (discussed in Section 4) in Table 6 (sentences that fall within a standard deviation of the mean sentence length), Table 7 (sentences shorter than one standard deviation of the mean), and Table 8 (sentences longer than one standard deviation of the mean).

C Z-score vs. Metrics: WMT19

In Section 4, we discuss correlations between Z-scores and three widely-used automated metrics for assessing the quality of machine translation outputs using the data from WMT18. Table 9 shows the same analysis using the data from WMT19. The conclusions are the same.

| | Translation *into* English | | | | |
| | all | w/ neg. | | w/o neg. | |
WMT18	Z	Z	%Δ	Z	%Δ
ru→en	0.231	0.087	(−62.3)	0.257	(+11.3)
de→en	0.401	0.413	(+3.0)	0.399	(−0.5)
et→en	0.327	0.236	(−27.8)	0.343	(+4.9)
tr→en	0.066	0.038	(−42.4)	0.072	(+9.1)
fi→en	0.149	0.172	(+15.4)	0.144	(−3.4)
cs→en	0.303	0.378	(+24.8)	0.287	(−5.3)
zh→en	0.169	0.175	(+3.6)	0.168	(−0.6)
WMT19					
ru→en	0.142	0.098	(−31.0)	0.148	(+4.2)
de→en	0.185	0.189	(+2.2)	0.185	(+0.0)
fi→en	0.294	0.272	(−7.5)	0.297	(+1.0)
lt→en	0.212	0.065	(−69.3)	0.243	(+14.6)
gu→en	0.216	0.119	(−44.9)	0.227	(+5.1)
kk→en	0.286	0.384	(+34.3)	0.276	(−3.5)

| | Translation *from* English | | | | |
| | all | w/ neg. | | w/o neg. | |
WMT18	Z	Z	%Δ	Z	%Δ
en→ru	0.348	0.249	(−28.4)	0.366	(+5.2)
en→de	0.651	0.723	(+11.1)	0.639	(−1.8)
en→et	0.535	0.478	(−10.7)	0.545	(+1.9)
en→tr	0.269	0.116	(−56.9)	0.294	(+9.3)
en→fi	0.568	0.708	(+24.6)	0.541	(−4.8)
en→cs	0.616	0.612	(−0.6)	0.617	(+0.2)
en→zh	0.241	0.244	(+1.2)	0.241	(+0.0)

Table 6: Evaluation of the best WMT18 and WMT19 submissions (same format as Table 1) using only the sentences that fall within a standard deviation of the mean sentence length $[\mu \pm \sigma]$ for each dataset.

D Ranking of Submissions

In Section 4, we show the ranking of systems obtained with all sentences, sentences with negation, and sentences without negation translating from Russian to English. In this section, we show the rankings of a few other language directions (Estonian to English in Table 10, Chinese to English in Table 11, Lithuanian to English in Table 12, Gujarati to English in Table 13 and English to Turkish in Table 14). We observe again many changes in the rankings calculated with sentences containing negation.

Translation *into* English					
	all	w/ neg.		w/o neg.	
WMT18	Z	Z	%Δ	Z	%Δ
ru→en	0.25	0.65	(+160.0)	0.21	(−16.0)
de→en	0.49	0.31	(−36.7)	0.51	(+4.1)
et→en	0.34	0.27	(−20.6)	0.35	(+2.9)
tr→en	0.12	0.33	(+175.0)	0.09	(−25.0)
fi→en	0.18	0.53	(+194.4)	0.13	(−27.7)
cs→en	0.27	0.20	(−25.9)	0.28	(+3.7)
zh→en	0.08	0.44	(+450.0)	0.06	(−25.0)
WMT19					
ru→en	0.26	0.09	(−65.3)	0.28	(+7.7)
de→en	0.11	-0.02	(−118.2)	0.12	(+9.09)
fi→en	0.33	0.61	(+84.8)	0.31	(−6.1)
lt→en	0.48	0.40	(−16.7)	0.48	(+0.0)
gu→en	0.21	-0.41	(−295.2)	0.22	(+4.8)
kk→en	0.25	0.05	(−80.0)	0.27	(+8.0)

Translation *from* English					
	all	w/ neg.		w/o neg.	
WMT18	Z	Z	%Δ	Z	%Δ
en→ru	0.44	0.42	(−4.5)	0.44	(+0.0)
en→de	0.71	0.36	(−49.3)	0.76	(+7.0)
en→et	0.76	0.53	(−30.3)	0.78	(+2.6)
en→tr	0.66	0.57	(−13.6)	0.68	(+3.0)
en→fi	0.54	0.50	(−7.4)	0.55	(+1.9)
en→cs	0.56	0.46	(−17.9)	0.58	(+3.6)
en→zh	0.01	0.25	(+2400.0)	0.00	(−100.0)

Table 7: Evaluation of the best WMT18 and WMT19 submissions (same format as Table 1) using only the sentences shorter than one standard deviation of the average $(0, \mu - \sigma)$ for each dataset.

Translation *into* English					
	all	w/ neg.		w/o neg.	
WMT18	Z	Z	%Δ	Z	%Δ
ru→en	0.08	0.03	(−62.5)	0.09	(+12.5)
de→en	0.40	0.29	(−27.5)	0.44	(+10.0)
et→en	0.31	0.20	(−35.5)	0.34	(+9.6)
tr→en	-0.14	-0.18	(+28.6)	-0.13	(−7.1)
fi→en	0.15	0.29	(+93.3)	0.12	(−20.0)
cs→en	0.31	0.26	(−16.1)	0.32	(+3.2)
zh→en	0.05	0.04	(−20.0)	0.05	(+0.0)
WMT19					
ru→en	0.15	0.21	(+40.0)	0.13	(−13.3)
de→en	-0.07	-0.05	(−28.6)	-0.07	(+0.0)
fi→en	0.21	0.38	(+81.0)	0.18	(−14.3)
lt→en	0.14	0.13	(−7.1)	0.14	(+0.0)
gu→en	0.18	0.15	(−16.7)	0.18	(+0.0)
kk→en	0.16	-0.13	(−181.3)	0.18	(+12.5)

Translation *from* English					
	all	w/ neg.		w/o neg.	
WMT18	Z	Z	%Δ	Z	%Δ
en→ru	0.29	0.14	(−51.7)	0.33	(+13.8)
en→de	0.62	0.72	(+16.1)	0.57	(−8.1)
en→et	0.42	0.28	(−33.3)	0.47	(+11.9)
en→tr	-0.06	-0.34	(+466.7)	0.00	(−100.0)
en→fi	0.26	0.02	(−92.3)	0.32	(+23.1)
en→cs	0.52	0.50	(−3.8)	0.53	(+1.9)
en→zh	0.28	0.17	(−39.3)	0.31	(+10.7)

Table 8: Evaluation of the best WMT18 and WMT19 submissions (same format as Table 1) using only the sentences longer than one standard deviation of the average $(\mu + \sigma, +\infty)$ for each dataset.

	Z vs. BLEU		Z vs. chrF++		Z vs. METEOR	
	w/ neg.	w/o neg.	w/ neg.	w/o neg.	w/ neg.	w/o neg.
ru → en	0.420	0.714	0.508	0.736	0.464	0.604
de → en	0.483	0.633	0.517	0.600	0.517	0.567
fi → en	0.788	0.818	0.848	0.909	0.879	0.758
lt → en	0.709	0.796	0.673	0.796	0.709	0.796
gu → en	0.564	0.745	0.782	0.964	0.527	0.891
kk → en	0.855	0.818	0.673	0.855	0.745	0.782

Table 9: Comparison of WMT19 rankings (Kendall's τ) using several metrics, and sentences with and without negation. We compare the official scoring (normalized direct assessments, Z-score) and three automated metrics: BLEU, chrF++ and METEOR. All τ coefficients are statistically significant ($p < 0.05$) (null hypothesis: there is no association between the Z-score and the automatic metric). The differences in the rankings obtained with the Z-score and the three metrics are more substantial when negation is present (lower τ coefficients).

		All sentences				Sentences w/ negation				Sentences w/o negation	
		System	Z			System	Z			System	Z
	1	tilde-nc-nmt	0.326	1	–	tilde-nc-nmt	0.232	1	–	tilde-nc-nmt	0.343
	2	NICT	0.238	2	↑3	uedin	0.179	2	–	NICT	0.255
	3	tilde-c-nmt	0.215	3	–	tilde-c-nmt	0.167	3	–	tilde-c-nmt	0.224
	4	M4t1ss	0.187	4	↓2	NICT	0.146	4	–	M4t1ss	0.207
WMT18	5	uedin	0.186	5	↑1	tilde-comb	0.129	5	–	uedin	0.187
	6	tilde-comb	0.171	6	↑4	online-A	0.116	6	–	tilde-comb	0.179
	7	online-B	0.117	7	↓3	M4t1ss	0.075	7	↑2	talp-upc	0.134
	8	HY-NMT-et-en	0.106	8	–	HY-NMT-et-en	0.065	8	↓1	online-B	0.13
	9	talp-upc	0.106	9	↑2	CUNI-Kocmi	0.061	9	↓1	HY-NMT-et-en	0.114
	10	online-A	0.063	10	↓3	online-B	0.043	10	–	online-A	0.053
	11	CUNI-Kocmi	0.007	11	↓2	talp-upc	-0.039	11	–	CUNI-Kocmi	-0.003
	12	neurotolge.ee	-0.117	12	–	neurotolge.ee	-0.126	12	–	neurotolge.ee	-0.116
	13	online-G	-0.341	13	–	online-G	-0.351	13	–	online-G	-0.339
	14	UnsupTartu	-0.950	14	–	UnsupTartu	-0.964	14	–	UnsupTartu	-0.948

Table 10: Rankings of all submissions translating from Estonian to English using all sentences (official WMT 2018 ranking), sentences with negation, and sentences without negation. tilde-comb refers to tilde-c-nmt-comb.

			All sentences				Sentences w/ negation				Sentences w/o negation	
			System	Z			System	Z			System	Z
WMT18	1		NiuTrans	0.14	1	↑8	UMD	0.197	1	–	NiuTrans	0.137
	2		online-B	0.111	2	–	online-B.0	0.18	2	↑2	Unisound	0.108
	3		UCAM	0.109	3	↑4	Li-Muze	0.179	3	–	UCAM	0.104
	4		Unisound	0.108	4	↑2	Unisound	0.17	4	↓2	online-B	0.103
	5		Tencent-ens	0.099	5	↓4	NiuTrans	0.162	5	–	Tencent-ens	0.092
	6		Unisound	0.094	6	↓1	Tencent-ens	0.156	6	–	Unisound	0.083
	7		Li-Muze	0.091	7	↓4	UCAM	0.149	7	↑1	NICT	0.083
	8		NICT	0.089	8	–	NICT	0.13	8	↓1	Li-Muze	0.08
	9		UMD	0.078	9	↓5	Unisound	0.105	9	–	UMD	0.063
	10		online-Y	-0.005	10	–	online-Y	0.027	10	–	online-Y	-0.01
	11		uedin	-0.017	11	↑1	online-A	-0.003	11	–	uedin	-0.018
	12		online-A	-0.061	12	↓1	uedin	-0.012	12	–	online-A	-0.069
	13		online-G	-0.327	13	↑1	online-F	-0.329	13	–	online-G	-0.323
	14		online-F	-0.377	14	↓1	online-G	-0.354	14	–	online-F	-0.383

Table 11: Rankings of all submissions translating from Chinese to English using all sentences (official WMT 2018 ranking), sentences with negation, and sentences without negation. Tencent-ens refers to the Tencent-ensemble-system.

			All sentences				Sentences w/ negation				Sentences w/o negation	
			System	Z			System	Z			System	Z
WMT19	1		GTCOM	0.234	1	↑1	tilde-nc	0.186	1	–	GTCOM	0.262
	2		tilde-nc	0.216	2	↑1	NEU	0.147	2	↑1	NEU	0.226
	3		NEU	0.213	3	↑1	MSRA.MASS	0.134	3	↑2	tilde-c-nmt	0.226
	4		MSRA.MASS	0.206	4	↓3	GTCOM	0.093	4	↓2	tilde-nc	0.222
	5		tilde-c-nmt	0.202	5	–	tilde-c-nmt	0.084	5	↓1	MSRA.MASS	0.22
	6		online-B	0.107	6	–	online-B	0.013	6	–	online-B	0.126
	7		online-A	-0.056	7	–	online-A	-0.126	7	–	online-A	-0.043
	8		TartuNLP-c	-0.059	8	–	TartuNLP-c	-0.141	8	–	TartuNLP-c	-0.043
	9		online-G	-0.284	9	↑1	JUMT	-0.37	9	–	online-G	-0.261
	10		JUMT.6616	-0.337	10	↓1	online-G	-0.402	10	–	JUMT	-0.33
	11		online-X	-0.396	11	–	online-X	-0.531	11	–	online-X	-0.369

Table 12: Rankings of all submissions translating from Lithuanian to English using all sentences (official WMT 2019 ranking), sentences with negation, and sentences without negation. GTCOM and tilde-nc refer to the systems GTCOM-Primary and tilde-nc-nmt, respectively.

		All sentences				Sentences w/ negation				Sentences w/o negation	
		System	Z			System	Z			System	Z
	1	NEU	0.210	1	–	NEU	0.112	1	–	NEU	0.221
	2	UEDIN	0.126	2	↑1	GTCOM	0.069	2	–	UEDIN	0.141
	3	GTCOM	0.100	3	↑3	NICT	0.037	3	–	GTCOM	0.104
WMT19	4	CUNI-T2T	0.090	4	–	CUNI-T2T	0.018	4	–	CUNI-T2T	0.098
	5	aylien-mt	0.066	5	↓3	UEDIN	-0.014	5	–	aylien-mt	0.086
	6	NICT	0.044	6	↑2	IITP-MT	-0.018	6	–	NICT	0.045
	7	online-G	-0.189	7	↓2	aylien-mt	-0.128	7	–	online-G	-0.189
	8	IITP-MT	-0.192	8	↓1	online-G	-0.192	8	–	IITP-MT	-0.211
	9	UdS-DFKI	-0.277	9	–	UdS-DFKI	-0.277	9	–	UdS-DFKI	-0.277
	10	IIITH-MT	-0.296	10	–	IIITH-MT	-0.355	10	–	IIITH-MT	-0.29
	11	Ju-Saarland	-0.598	11	–	Ju-Saarland	-0.566	11	–	Ju-Saarland	-0.601

Table 13: Rankings of all submissions translating from Gujarati to English using all sentences (official WMT 2019 ranking), sentences with negation, and sentences without negation. CUNI-T2T, aylien-mt, and GTCOM refer to the systems CUNI-T2T-transfer-guen, aylien-mt-gu-en-multilingual, and GTCOM-Primary, respectively

		All sentences				Sentences w/ negation				Sentences w/o negation	
		System	Z			System	Z			System	Z
	1	online-B	0.277	1	↑1	uedin	0.231	1	–	online-B	0.308
	2	uedin	0.222	2	↑1	alibaba5732	0.138	2	↑1	alibaba5732	0.232
WMT18	3	alibaba5732	0.216	3	↑2	alibaba5744	0.127	3	↓1	uedin	0.221
	4	NICT	0.128	4	–	NICT	0.1	4	–	NICT	0.135
	5	alibaba5744	0.111	5	↓4	online-B	0.094	5	–	alibaba5744	0.107
	6	online-G	0.058	6	–	online-G	-0.004	6	–	online-G	0.071
	7	RWTH	-0.06	7	–	RWTH	-0.097	7	–	RWTH	-0.052
	8	online-A	-0.254	8	–	online-A	-0.481	8	–	online-A	-0.206

Table 14: Rankings of all submissions translating from English to Turkish using all sentences (official WMT 2018 ranking), sentences with negation, and sentences without negation. alibaba5732 and alibaba5744 refer to the systems alibaba-ensemble-model.5732 and alibaba-ensemble-model.5744, respectively.