Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks

Cui, Ruixiang; Hershcovich, Daniel; Søgaard, Anders

Published in:
NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics

DOI:
10.18653/v1/2022.naacl-main.359

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks

Ruixiang Cui, Daniel Hershcovich, Anders Søgaard
University of Copenhagen
{rc, dh, soegaard}@di.ku.dk

Abstract

Logical approaches to representing language have developed and evaluated computational models of quantifier words since the 19th century, but today’s NLU models still struggle to capture their semantics. We rely on Generalized Quantifier Theory for language-independent representations of the semantics of quantifier words, to quantify their contribution to the errors of NLU models. We find that quantifiers are pervasive in NLU benchmarks, and their occurrence at test time is associated with performance drops. Multilingual models also exhibit unsatisfying quantifier reasoning abilities, but not necessarily worse for non-English languages. To facilitate directly-targeted probing, we present an adversarial generalized quantifier NLI task (GQNLI) and show that pre-trained language models have a clear lack of robustness in generalized quantifier reasoning.

1 Introduction

Quantifier words—such as each or most or more than three—have been extensively studied, both in logic and in linguistics (Westerståhl, 1989; Peters and Westerståhl, 2006), going all the way back to Frege (1879). In this paper, we examine the extent to which they present a challenge to modern NLU systems. Our analysis is motivated by three observations:

Quantifier words are abstract Unlike nouns, verbs and adjectives, quantifier words do not have referents out in the world. Rather, quantifier words specify relationships between sets of entities, events and properties. To provide intuitions about the semantics of quantifier words, and to be able to refer to quantifiers in a language-independent way, we rely on the notion of generalized quantifiers (Mostowski, 1957), as described in §2.

Quantifier words vary across languages Quantifier word inventories differ across languages.

Table 1: Examples of quantifiers (marked in bold texts) in NLP tasks, with RoBERTa’s prediction for QA and XLM-R’s prediction for NLI after fine-tuning.

| CONTEXT: A piece of paper was later found on which he had written his last statements in two languages, Latin and German. Only one statement was in Latin and the rest in German. | QA English |
| QUESTION: In what language were most statements written? | ANSWER: German | PREDICTED ANSWER: Latin and German |
| PREMISE: Más de tres personas resultaron heridas en un accidente de dos vehículos el lunes por la noche. (translation: More than three people were injured in a two-vehicle crash Monday evening.) | NLI Spanish |
| HYPOTHESIS: Había 4 personas involucradas. (translation: There were 4 people involved.) | LABEL: Neutral | PREDICTED LABEL: Entailment |

Often what is considered rough translation equivalents also differ in syntax, fine-grained semantics or pragmatics. Stateva et al. (2019) show, e.g., that perceptions of the numerical bounds of existential quantifiers differ across speakers of English, French, Slovenian, and German. Other papers showing discrepancies between quantifier systems include comparisons of Salish to English (Matthewson, 2001), Adyghe to English (Nikolaeva, 2012), or of Dutch, Hebrew and Bengali (Gil, 1982). The cross-linguistic differences in how generalized quantifiers are expressed motivates a cross-lingual error analysis, since quantifiers may contribute more to error when processing some languages rather than others.

Quantifier words are important Quantifier words are extremely important for tasks that require inference, including natural language inference, question answering, fact-checking, etc. Datasets have, for example, been developed for numerical reasoning in English (Dua et al., 2019). Several researchers have identified quantifier words as important sources of errors for natural language processing systems (Joshi et al., 2020); see Table 1 for examples of such errors. Unfortunately, most
Table 2: The categorization set of quantifiers for task analysis. The first six are Aristotelian/counting quantifiers and the following four are proportional quantifiers. The last one is a Ramsey quantifier (Schmerl and Simpson, 1982). For each quantifier, its logical denotation is listed in the second column. The third column contains English examples with quantifiers taken from XNLI.

<table>
<thead>
<tr>
<th>Generalized Quantifiers</th>
<th>Logical Denotation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>some(A)(B) = 1</td>
<td>$A \cap B \neq \emptyset$</td>
<td>This process is known to increase security in several ways.</td>
</tr>
<tr>
<td>all(A)(B) = 1</td>
<td>$A \subseteq B$</td>
<td>Everyone agreed the food was terrible.</td>
</tr>
<tr>
<td>more than k</td>
<td>$A \cap B &gt; k$</td>
<td>They do let them go more than twice a week.</td>
</tr>
<tr>
<td>less than k</td>
<td>$A \cap B &lt; k$</td>
<td>San Augustin Acólman has less than 1,000 residents.</td>
</tr>
<tr>
<td>k (A)(B) = 1</td>
<td>$A \cap B = k$</td>
<td>Please donate 100 million to the School of Nursing.</td>
</tr>
<tr>
<td>between p and k (A)(B) = 1</td>
<td>$p &lt;</td>
<td>A \cap B</td>
</tr>
<tr>
<td>the p/k (A)(B) = 1</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>the few (A)(B) = 1</td>
<td>$</td>
<td>A \cap B</td>
</tr>
<tr>
<td>each other (A)(B) = 1</td>
<td>$\forall a \in (A \cap B) \exists b \in (A \cap B)(a \neq b)$</td>
<td>All of these trails are located within a one hour drive of each other.</td>
</tr>
</tbody>
</table>

Contributions: We analyze how quantifiers are represented in NLU benchmarks, and how their occurrence at test time contributes to errors by neural language models (LMs). We derive a linguistically motivated 11-way categorization set for generalized quantifiers and look into their distribution in three steps: (a) monolingual NLI; (b) cross-lingual NLI; (c) cross-lingual question answering. We also propose GQNLI, an adversarial generalized quantifier NLI challenge dataset. Our work shows that (i) generalized quantifiers are pervasive and cause overall performance drops in NLU benchmarks; (ii) the contribution of quantifier words to system error varies across languages; and (iii) generalized quantifiers are particularly difficult for LMs in interaction with negation and subsumption.

2 Background

Generalized quantifiers (GQs) are developed upon first-order predicate logic, denoting relations between sets (Mostowski, 1957). Given a universe $E$, a quantifier $Q$ would be treated as a mapping $Q_E$ from the Cartesian product of powersets $\mathcal{P}(E) \times \mathcal{P}(E)$ to the set $\{false, true\}$ or, as a binary relation on subsets of $E$ (Dvôr&žák and Hol&lácek, 2015). GQs are generalizations of the $\forall, \exists$ quantifiers from first-order predicate logic (Mostowski, 1957; Lindström, 1966; Montague, 1973; Bach et al., 1995; Keenan and Paperno, 2012). A generalized quantifier is, abstractly, a relation between sets. Generalized quantifier theory, while developed by logicians, is used by formal linguists to analyze the meaning of quantifier words in combination with referential expressions (Barwise and Cooper, 1981; Higginbotham and May, 1981).

Most human languages contain ways of expressing generalized quantifiers, and their semantics exhibit striking similarities across languages (Matthewson, 2004; Fintel and Matthewson, 2008; Steinert-Threlkeld, 2019). At the same time, generalized quantifiers can be instantiated very differently across languages due to pragmatic considerations (Grice, 1989) or cognitive economy and cost-benefit optimisation in the exchange of information (Levinson et al., 2000; Steinert-Threlkeld, 2021; Uegaki, 2022). Quantifier words also exhibit syntactic differences, e.g., with some languages having specialized words to express quantity, while others rely on metaphorical usage of common nouns (Katsos et al., 2012). In English, most is a determiner, but Spanish and French express the same concept through common nouns, la mayoría and la majorité. The relative stability of the core semantics of quantifiers makes a cross-linguistic comparison possible, but the syntactic and pragmatic variation associated with the expression of generalized quantifiers poses a challenge for multilingual NLU. We consult quantifier taxonomy studies (Keenan and Westersthl, 1997; Peters and Westersthl, 2006; Szymanik and Thorne, 2015; Szymanik, 2016) and derive a categorization set for quantifier analysis in NLU benchmarks. In Table 2, we list the 11-way quantifier categorization set and their logical denotation based on set theory.

While other foci of formal linguistics have attracted the attention of NLP researchers—including coreference (Ogrodniczuk et al., 2019, 2020), nega-
Table 3: Quantifier distribution in four NLI tasks, among which three are monolingual English and one is cross-lingual. The table show statistics of the test set, if not available, dev set, of the target task. All but the last rows show the occurrence time of the type of quantifier in the first column. The last row represents the distribution rate of any quantifier in the dataset.

Figure 1: Relative distribution of quantifiers in NLI and QA tasks ranked by semantic complexity. The bars show the relative frequency of such quantifier and the lines indicate the cumulative frequency for a task.

**3 NLU Benchmarks**

We conduct an error analysis focusing on the role of generalized quantifiers in two NLU tasks, Natural Language Inference (NLI) and Question Answering (QA), which generally require understanding of quantifiers. For each type of task, both monolingual and cross-lingual evaluation are conducted. We focus on generalized quantifiers in the hypotheses in NLI examples—and on generalized quantifiers in the *question* fields in question answering. To this end, we identify quantifiers by the lemma and the universal dependency relation (Nivre et al., 2020) of a quantifier after preprocessing the sentences using *Stanza* (Qi et al., 2020). Take the sentence “The Yiddish culture has survived for more than a thousand years.”, we annotate it as

**3 Note that generalized quantifiers are not always explicit in discourse. The sentence inadequate sleep causes obesity should be interpreted as *Most of those who do not sleep adequately, gain weight* (Zadeh, 1983). Such implicit quantifiers related to pragmatic variation are important for language understanding, but will be ignored in this work.

**4 Quantifiers in English NLI Benchmarks**

NLI is commonly framed as a three-way classification task with labels *entailment*, *contradiction* and *neutral* (Bowman et al., 2015a). While SOTA models exhibit low error rates on NLI benchmarks, it is unclear when they succeed or fail in their underlying reasoning. We are interested in whether generalized quantifiers challenge modern NLI models. In our error analysis, we initially focus on three English NLI datasets, MultiNLI (MNLI; Williams et al., 2018), SNLI (Bowman et al., 2015a) and ANLI (Nie et al., 2020) as testbeds.
across, about 10% of all hypotheses contain quantifier words, indicating the pervasiveness of quantification. We also plot the frequency of quantifiers in NLI in Figure 1 and find the quantifier word distribution follows Zipf’s law (Zipf, 1949). Note the top three most common quantifiers account for more than 90% of all.

**Experiments and Results**  In order to investigate whether NLU systems can solve quantifiers in NLI, we experiment with two pretrained LMs: BERT³ (Devlin et al., 2019) and RoBERTa⁴ (Liu et al., 2019). We use the codebase by Nie et al. (2020). The training data combines SNLI, MNLI, FEVER-NLI (Nie et al., 2019) and ANLI.

In Table 4, we report the test set performance on SNLI and ANLI, and the dev set performance on MLNI matched and mismatched sections. We can observe that SOTA models suffer from performance drops across almost all quantification phenomena in every task. When it comes to performance over all quantifiers, the improvement from RoBERTa to BERT (2.2%) is less prominent than that over full datasets (2.9%), suggesting RoBERTa is particularly challenged.

Taking a closer look at error by category, proportional quantifiers seem harder to solve than Aristotelian/counting quantifiers. Except for k%, all proportional quantifiers—pk, most, and few—are about 10% lower than the five counting quantifiers (except less than k) with BERT; and about 5% lower with RoBERTa. RoBERTa is not generally superior to BERT; e.g., for k%, BERT outperforms it by 22%. We show a pairwise analysis of how GQs affect performance when they appear in both the premises and hypotheses in the Appendix B. Generally, our results attest to the difficulty of resolving GQs in NLI benchmarks.

### Table 4: BERT and RoBERTa performance on NLI tasks.

<table>
<thead>
<tr>
<th>Quantifier</th>
<th>M_m</th>
<th>M_mm</th>
<th>SNLI</th>
<th>A_R1</th>
<th>A_R2</th>
<th>A_R3</th>
<th>weig.</th>
<th>M_m</th>
<th>M_mm</th>
<th>SNLI</th>
<th>A_R1</th>
<th>A_R2</th>
<th>A_R3</th>
<th>weig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>some</td>
<td>82.5</td>
<td>84.1</td>
<td>86.9</td>
<td>100</td>
<td>0</td>
<td>47.1</td>
<td>83.4</td>
<td>83</td>
<td>84.8</td>
<td>86.9</td>
<td>100</td>
<td>100</td>
<td>41.1</td>
<td>83.7</td>
</tr>
<tr>
<td>all</td>
<td>85.9</td>
<td>88.3</td>
<td>89.2</td>
<td>46.7</td>
<td>37.5</td>
<td>34.5</td>
<td>83.2</td>
<td>85.9</td>
<td>92.1</td>
<td>90.8</td>
<td>66.7</td>
<td>37.5</td>
<td>34.5</td>
<td>85.3</td>
</tr>
<tr>
<td>&gt; k</td>
<td>85.7</td>
<td>100</td>
<td>87.5</td>
<td>70</td>
<td>43.8</td>
<td>42.9</td>
<td>73</td>
<td>100</td>
<td>100</td>
<td>83.3</td>
<td>100</td>
<td>91.7</td>
<td>100</td>
<td>91.7</td>
</tr>
<tr>
<td>&lt; k</td>
<td>87.2</td>
<td>81.8</td>
<td>92.4</td>
<td>43.6</td>
<td>43.5</td>
<td>33.3</td>
<td>84.8</td>
<td>88.3</td>
<td>88.8</td>
<td>92.9</td>
<td>56.3</td>
<td>61.3</td>
<td>43.8</td>
<td>87.8</td>
</tr>
<tr>
<td>k%</td>
<td>74.3</td>
<td>79.5</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>74.4</td>
<td>77</td>
<td>87.2</td>
<td>100</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>72.2</td>
<td></td>
</tr>
<tr>
<td>most</td>
<td>78.6</td>
<td>73.3</td>
<td>90.9</td>
<td>33.3</td>
<td>33.3</td>
<td>73.9</td>
<td>85.7</td>
<td>50</td>
<td>100</td>
<td>88.6</td>
<td>50</td>
<td>84.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>few</td>
<td>75</td>
<td>100</td>
<td>85.7</td>
<td>50</td>
<td>84.1</td>
<td>50</td>
<td>100</td>
<td>85.4</td>
<td>88.8</td>
<td>91.7</td>
<td>65.3</td>
<td>56.5</td>
<td>40.3</td>
<td>85.5</td>
</tr>
<tr>
<td>all GQs</td>
<td>85</td>
<td>84.8</td>
<td>91.2</td>
<td>50.5</td>
<td>44.4</td>
<td>39</td>
<td>83.3</td>
<td>89.5</td>
<td>89.4</td>
<td>92.3</td>
<td>71.7</td>
<td>49.6</td>
<td>49</td>
<td>87.3</td>
</tr>
<tr>
<td>full</td>
<td>86.5</td>
<td>86.1</td>
<td>91.3</td>
<td>58.6</td>
<td>48</td>
<td>43.2</td>
<td>84.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The penultimate row stands for the overall performance when quantifiers exist in a dataset. The last row reports the overall performance in a dataset. Number marked in bold signifies a lower score than the overall performance.

5 Quantifiers in Cross-lingual NLU

**Benchmarks**

Quantifiers are acquired in similar orders across languages (Katsos et al., 2016), although languages express quantifiers in different ways. For example, there are eight different universal quantifiers with different level of distributivity in Malagasy (Matthewson, 2008). This poses challenges to training multilingual LMs and transfer learning. We are interested in whether quantifiers are universally and evenly challenging for all languages.

5.1 Quantifiers in Cross-lingual NLI

We choose XNLI (Conneau et al., 2018), a manual translation of the development and test set of MNLI into 15 languages, for this multilingual error analysis. We should clarify that for XNLI, the authors annotate entailment labels for the English data only and apply them to the other languages. We do not assume label changes due to translation in this study, but it is worth investigate in the future. We choose five languages belonging to different language families, namely Arabic, Chinese, German, Spanish and Vietnamese as targets. The last column in Table 3 shows the numbers of quantifiers in XNLI. The distribution rate is 10%. Note that the universal quantifier is the most common quantifier in XNLI.
We fine-tune mBERT\(^5\) (Devlin et al., 2019) and XLM\(^6\) (Lample and Conneau, 2019) on the MNLI training set and evaluate them on XNLI. We report the results in Table 5. We find that performance varies across languages. For Chinese and Vietnamese, we see significant drops in performance for examples with GQs, whereas for Arabic and German, we see improvements. The results per quantifier are more homogeneous, however.

Similar to our results for English, we can see that the lowest accuracies in XNLI are with proportional quantifiers, such as most and few. But the gap in non-English languages is wider for these two categories, especially for Chinese, the difference reaches 30%. Other hard quantifiers include all, > k, < k, and each other.

**Quantifiers in Cross-lingual QA** Cross-lingual question answering (XQA) is another important NLU task that evaluates the cross-lingual transferability of LMs. We evaluate the effect of quantifiers on system errors across two XQA datasets, namely XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020). As demonstrated in Figure 1, quantifier word distributions in XQA tasks also follow Zipf’s law, as in NLI tasks, but \(k\) is more frequent (perhaps because of a traditional emphasis on numerical reasoning), and we see less variance across languages. This is probably because question answering is targeting quantification less directly. To evaluate cross-lingual QA performance on GQs, we fine-tune mBERT and XLM-R\(^7\) (Conneau et al., 2020) using Hu et al. (2020)’s architecture. We present results for mBERT in Table 7; for XLM-R results, please refer to Appendix D.

Just as with XNLI, LMs suffer from performance drops across all languages for almost all GQ phenomena with significant, cross-lingual variation. The most distinguished is that Exact Match (EM) suffers from a greater deterioration than F1 scores for all languages. For example, the weighted EM difference for mBERT on MLQA is 2.9% while the weighted F1 is 1%. As one example in Table 1, we observe that the plausible answers selected by models, while being incorrect, result in a sharper decrease of EMs comparing to F1s. Questions containing GQs also tend to have less verbal answers comparing to those without GQs, and therefore require higher precision.

Regarding cross-lingual comparisons, Chinese and Arabic are the two languages that do not have

---

\(^5\)multi_cased_L-12_H-768_A-12  
\(^6\)xlm-mlm-100-1280  
\(^7\)xlm-roberta-large
lower performance over GQs compared to the performance over the complete dataset. Despite the overall trends, subtle differences from XNLI performance still exist. For example, XLM-R is worse than mBERT on quantifier reasoning on XQuAD Chinese, especially at proportional quantifiers, but this is not the case on MLQA Chinese.

### 6 GQNLI

We have seen how quantifiers present challenges to NLI and QA models. Using an approach similar to ANLI (Nie et al., 2020) and DynaBench (Kiela et al., 2021), we use model difficulty (RoBERTa’s) as a heuristic to select hard examples for a challenge dataset that can hopefully be used to evaluate any future progress on this. We propose GQNLI, a generalized quantifier NLI challenge dataset, consisting of 30 premises and 300 hypotheses. The average sentence lengths of hypothesis and premises are 15.97 and 7.35, respectively. Both numbers are comparable to those of MNLI, but lower than ANLI’s (Williams et al., 2020). It should be noted that GQNLI is designed for evaluating future models; obviously not for benchmarking RoBERTa.

#### Dataset Creation

Firstly, we manually create 100 premise-hypothesis pairs, in which various types of GQs appear. For each premise and hypothesis, the number of GQs varies from one to three. To choose the premises, we randomly sampled 100 premises with GQs from SNLI and ANLI test sets, respectively, and selected 10 premises in total, that we consider are semantically adequate for adding GQs and making simple hypotheses.

To construct the hypotheses, we rely on RoBERTa fine-tuned on MNLI and manually select examples about which the model is unsure or incorrect. To focus on GQs, we keep the challenge examples otherwise simple (Ribeiro et al., 2020), and avoid lexical variations in the hypotheses. Hard examples were found to be characterized by (i) mixing generalized quantifiers with other logical operators, such as subsumption or negation, and (ii) combining multiple different generalized quantifiers. We discuss these observations in Section 7.

Two of the authors annotated the examples. The inter-annotator agreement (Fleiss’ kappa) was 0.895, substantially higher than ANLI’s (0.672–0.740). It is worth noting that the level of semantic or pragmatic interpretation difference of GQs is reflected in the measurement.

We augmented the examples by substituting non-quantifier words (e.g., replacing “dogs” with “cats”) while keeping the labels, to exclude the effect of specific lexical items. The resulting labels are uniformly distributed. Table 8 presents GQNLI statistics. Since the dataset is curated to probe the ability to reason with quantifiers, the distribution of generalized quantifiers does not follow Zipf’s law; see §4. A list of GQNLI examples per category is shown in Appendix E.

### Experiments and Results

We evaluate seven types of models on GQNLI, fine-tuned with different combinations of NLI datasets. As data creation only relied on RoBERTa and MNLI, nothing prevents that models with different architectures and training data will perform well. They do not, however. The results are shown in Table 8.

We see that all models have great difficulty with GQNLI. With more training data, models improve, but the best performance is 48%, less than 15 points above chance level. In general, the counting quantifiers, especially the existential and universal quantifiers, are easier than proportional quantifiers. Particularly, most models struggle with less than k and between. This is in some contrast with the NLU tasks studied above, where these quantifiers were among the easiest.

We also observe unstable GQ reasoning ability reflected in the measurement.
Quantifier & some & all & $k < k$ & $k % most & few & each other & Overall \\
# Occurrence & 27 & 51 & 51 & 33 & 170 & 21 & 24 & 45 & 18 & 9 & 36 \\

DeBERTa fine-tuned with DocNLI, which unifies

to higher accuracies. Especially,

## Discussion

### Negation

The interaction between negation words and quantifiers increases semantic complexity (Partee, 1970; Horn, 2010). We investigate whether this holds for NLI tasks, using negation cue detection to find all cases where a negation word and a quantifier appear in the hypotheses.

We break down the performances on negation of the seven models in Appendix F. As indicated, LMs overall have polarized results for negation cases comparing to the entire dataset. We can see a majority of the models even predicted opposite labels for some GQ categories, with 0% accuracy. BART is no longer the second best model, replaced by RoBERTa. The improvement by training with more data is overall consistent for reasoning over GQs with negation.

For a cross-lingual investigation of the interaction of GQs and negation, we find that in XNLI, the number of cases combining both phenomena is insufficient: we identified four such cases, involving only the quantifiers “all” and “more than.” For English, mBERT predicted two cases successfully. For Chinese, German, Vietnamese and Arabic, one is correct. For Spanish, all are wrongly predicted.

It is evident that NLU models suffer from reasoning difficulties in certain cases when negation interacts with GQs, especially in cross-lingual evaluation. In future work, we are interested in expanding GQNLI to more instances and more languages to facilitate qualitative investigations.

### Subsumption

In generalized term subsumption languages (TSLs; Yen, 1991; Ali and Shapiro, 1993), a term $a$ subsumes another term $b$ if and only if the extension of $a$ is a superset of the extension of $b$. Rather than surface number comparison, subsumption reasoning requires knowledge of the relations between supersets and subsets. For example, to decide whether “There are six dogs. Three brown dogs, a black dog and a white dog run along the green grass” entails “One dog sits”, LMs should be aware that “six dogs” is a superset of the extension of the “brown dogs”, “black dog” and “white dog”. Another example in GQNLI is to infer whether “There are twelve singers on a stage, less than half from Argentina and one from Cape Verde” entails “Several singers do not come from Chile”.

We annotate 63 cases out of the first 100 in GQNLI requiring subsumption reasoning. We show the statistics and results regarding subsumption in Appendix G. It can be seen that more training data leads to higher accuracies. Especially, DeBERTa fine-tuned with DocNLI, which unifies

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th>% Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>S.M,F,ANLI</td>
<td>40.7</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>S.M,F,ANLI</td>
<td>37.0</td>
</tr>
<tr>
<td>ALBERT</td>
<td>S.M,F,ANLI</td>
<td>66.7</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>MNLI</td>
<td>48.1</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>M,F,ANLI</td>
<td>81.5</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>M,F,Ling,DocNLI</td>
<td>77.8</td>
</tr>
<tr>
<td>SBERT</td>
<td>MNLI</td>
<td>59.3</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>MNLI</td>
<td>55.6</td>
</tr>
<tr>
<td>BART</td>
<td>MNLI</td>
<td>40.7</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>MNLI</td>
<td>50.6</td>
</tr>
<tr>
<td>SBERT</td>
<td>S,M,F,ANLI</td>
<td>66.7</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>S,M,F,ANLI</td>
<td>54.9</td>
</tr>
<tr>
<td>ALBERT</td>
<td>S,M,F,ANLI</td>
<td>70.4</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>MNLI</td>
<td>48.1</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>M,F,ANLI</td>
<td>81.5</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>M,F,Ling,DocNLI</td>
<td>77.8</td>
</tr>
</tbody>
</table>

Table 8: GQNLI statistics and seven types of models’ performance with different combinations of training data. The second row shows the occurrence time of the type of GQ in GQNLI. The following rows show models’ performance on the dataset. We tested most competitive models fine-tuned for NLI available on Hugging Face. All numbers in bold represent the highest accuracy in one category. Due to space limitation we provide the link to each model in the Appendix H.
the two classes “neutral” and “contradict” into a new class “not entail”, has a significant improvement on subsumption cases with neutral label. The training bias give an advantage to the model on the subsumption subset, half cases of which are labelled neutral. But such bias has a negative effect on non-subsumption cases; the accuracy drops by 20.2% comparing to the model without training with DocNLI. It is worth investigating whether DocNLI is truly helping subsumption reasoning in future work. Subsumption is a key concept in the study of knowledge representation (Woods, 1991), but is neglected in current NLP research. The fact that LMs struggle to perform subsumption reasoning asserts the necessity to explicit tackle the problem.

8 Related Work

We examine the sensitivity of NLU models to generalized quantifiers. These models are designed to induce correlations from large volumes of data, not to reason symbolically with logical quantifiers. Such models have, nevertheless, been probed for logical knowledge.

Mul and Zuidema (2019), for example, show neural networks encode fragments of first-order logic and exhibit zero-shot generalization ability. Evans et al. (2018) present a neural architecture that improves performance on propositional logical inference. Bowman et al. (2015b) also suggest neural networks learn semantic representations for logical inference in natural languages. However, on the same task, Veldhoen and Zuidema (2017) find neural networks fail to do so on a more stringent test. Geiger et al. (2019) also show that neural networks fail to exhibit robust logical inference. Srivastava et al. (2018) use semantic parsers to encode quantifiers and improve zero-shot learning in classification tasks. Haruta et al. (2020) present a system that computes logical inference over GQs and see improvements on two specialized datasets, FraCaS (Cooper et al., 1994) and MED (Yanaka et al., 2019). None of these papers explicitly discussed generalized quantifiers, and all were limited to studying the ability of neural networks to capture the logical semantics of English.

Many studies have instead focused on LMs’ ability to capture negation (Gururangan et al., 2018; Naik et al., 2018; Hossain et al., 2020; Ettinger, 2020; Hartmann et al., 2021) or coreference (Ye et al., 2020; Varkel and Globerson, 2020; Abdou et al., 2020). Others have focused on LMs’ ability to reason with numbers (Johnson et al., 2020). DROP (Dua et al., 2019), for example, is a question answering dataset designed specifically to probe LMs’ ability to count, add and subtract for answering factoid questions. Models have also been tailored for numerical reasoning (Geva et al., 2020; Zhang et al., 2020). Cobbe et al. (2021) proposes to use a verification task during pretraining of LMs to improve their ability to solve math word problems. Others have studied monotonicity inference (Hu et al., 2019; Yanaka et al., 2019, 2020), and Fang and Lou (2021) recently focused on the two quantifier words part and whole in an error analysis for named entity recognition.

Many NLU benchmarks contain quantifier words, but their influence on performance has not been studied systematically. One exception to this is that generalized quantifiers have been used to generate adversarial examples in the context of numerical reasoning (Naik et al., 2018; Nie et al., 2020). TaxiNLI (Joshi et al., 2020), which categorizes 15 types of reasoning abilities, is a dataset drawn from MNLI. In their taxonomy, the Quantifier category only refers to universal and existential quantifiers, not to generalized quantifiers, and ditto for Kim et al. (2019). All of the above focused on English, but in an extension to TaxiNLI, K et al. (2021) incorporated quantifiers into the Logic class and found a large cross-lingual transfer gap on LMs.

9 Conclusion

Quantifiers lie in the intersection of logic, linguistics and NLP research. It is essential for NLU systems to learn quantifier reasoning. We examined generalized quantifiers in multilingual NLU tasks with regards to their expressiveness and logical reasoning requirement. Our survey and experiments indicate quantifiers are neglected to a degree and cause significant performance drops for neural LMs. To better understand LMs’ reasoning abilities, we release GQNLI, a novel generalized quantifier NLI challenge dataset. With the pervasiveness of generalized quantifiers, we stress that more efforts are necessary to investigate: (1) when and why models systematically fail when quantifiers interact with other operators; (2) how to improve cross-lingual transferability of quantifiers; (3) how we can exploit the theoretical results about generalized quantifiers from logic and linguistic studies,
so as to improve the logical inference ability of neural LMs.

Acknowledgements

We would like to thank Miryam de Lhoneux, Constanza Fierro, Desmond Elliott and the anonymous reviewers for their valuable feedback.

References


Antonín Dvořák and Michal Holčapek. 2015. Type 1 , 1 fuzzy quantifiers determined by fuzzy measures on residuated lattices. part iii. extension, conservativity and extensionality. Fuzzy Sets Syst., 271(C):133–155.


Gottlob Frege. 1879. Begriffsschrift. eine der arithmetischen nachgebildete formelsprache des reinen denkens.


Laurence R. Horn. 2010. The expression of negation.


Napoleon Katsos, Maria-Jose Ezeizabarrena, Anna Gavarro, Jelena Kuvac Kraljevic, Gordana Hrzica, Kleaithes K Grohmann, Athina Skordi, Kristine Jensen de Lopez, Lone Sundahl, Angeliek Van Hout,


Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. Anlizing the adversarial natural language inference dataset.


Table 9: Regular Expressions for generalized quantifiers.

### Appendices

#### A Regular Expressions for Generalized Quantifiers

Table 9 lists the regex we use to parse generalized quantifiers in sentences augmented with universal dependency tags. The approach does not find all the generalized quantifiers exhaustively but rather approximates the common distributions.

#### B Pairwise Observation

While the analysis in Section 4 is based on quantifiers in hypotheses, next we consider the interaction of quantifiers in hypotheses and quantifiers in premises. To this end, we calculate the difference between overall performance and performance for premise-hypothesis pairs of GQs. In Figure 2, we visualize the results as heatmaps (see Table 10 for exact numbers of occurrences and accuracies). Surprisingly, whenever quantifiers appear in both the premise and the hypothesis, LMs largely fail to predict the entailment. Percentage quantifiers, supposed to be semantically more complex than counting quantifiers, are not de facto harder in NLI. We studied all 27 cases of percentage quantifiers in the English NLI datasets, and found that in most cases, percentage quantifiers occurrences are identical across premises and hypotheses, i.e., triggering little or no inference. The other two proportional quantifiers, most and few, are hard for...
Figure 2: Fine-grained analysis of RoBERTa performance on 6 English NLI subtasks. Each heatmap represents hypotheses with a type of quantifier. The rows stand for premises with the quantifier of that label. The numbers are calculated as the accuracy over the whole dataset minus the fine-grained accuracy given a specific premise and hypothesis (the higher the number, the worse the performance). For each heatmap, the last column represents the accuracy gap weighted by all 6 tasks. “UN” stands for an entry where no explicit quantifier is identified.

LMs to resolve, e.g., in some quantifier pairs, models yield 0% accuracy. Although *each other* is supposed to be hardest to resolve due to the complex semantics of reciprocals (Szymanik and Thorne, 2015), it is not reflected in NLI tasks as such. The reason is similar to percentage quantifiers, while annotators intend to alter counting quantifiers when writing hypotheses, reciprocality is seldomly considered a linguistic ability that needs testing for NLU systems. And the annotation for Ramsey quantifier is simply a knockoff, making reciprocal relation identification unwarranted through shallow correlations.

C Fine-grained NLI Analysis

D XQA Result: mBERT and XLM-R

Table 11 compares the results of mBERT and XLM-R on two XQA tasks, XQuAD and MLQA.

E GQNLI Examples

Table 12 list one example per category in GQNLI.

F GQNLI Negation Cases

We present the results of seven models’ performance on cases with negation cues in GQNLI in Table 13.

G GQNLI Subsumption Cases

See Table 14 for models’ performance on cases requiring subsumption reasoning in GQNLI. We also break down subsumption results by entailment labels into two categories: neutral and non-neutral.
H GQNLI Experiment Details

We reused the fine-tuned BERT and RobERTa in Section 4. The other fine-tuned LMs are from Hugging Face. We list the models and their links in Table 15.
<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Premise</th>
<th>MNLI_m_dev</th>
<th>MNLI_mm_dev</th>
<th>SNLI_test</th>
<th>ANLI_R1_test</th>
<th>ANLI_R2_test</th>
<th>ANLI_R3_test</th>
<th>Total #occurrence</th>
<th>%Acc</th>
<th>רכש</th>
<th>יכwifi=pk</th>
<th>יכwifi=pk</th>
<th>יכwifi=pk</th>
<th>יכwifi=pk</th>
</tr>
</thead>
<tbody>
<tr>
<td>most</td>
<td>most</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
</tr>
<tr>
<td>few</td>
<td>few</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
<td>4891</td>
</tr>
</tbody>
</table>

Table 10: Statistics of pairwise analysis in Monolingual NLI Benchmarks
Table 13: Models’ performance on instances with negation cues in QGNLI.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th># Occurrence with negation cues</th>
<th>some</th>
<th>all</th>
<th>k &gt; 6</th>
<th>k &lt; 6</th>
<th>k = 6</th>
<th>between 3</th>
<th>p/k</th>
<th>k 6</th>
<th>most 9</th>
<th>few 9</th>
<th>each 9</th>
<th>Overall 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>S,M,F,ANLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>ELECTRA</td>
<td>S,M,F,ANLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>SBERT</td>
<td>S,M,F,ANLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>MNLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>ALBERT</td>
<td>S,M,F,ANLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>BART</td>
<td>MNLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>MNLI</td>
<td></td>
<td>0.67</td>
<td>100</td>
<td>33.3</td>
<td>50</td>
<td>0</td>
<td>30</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>22.2</td>
<td>33.3</td>
<td>39.2</td>
</tr>
<tr>
<td>Model</td>
<td>Training Data</td>
<td># Occurrence</td>
<td>Subsumption (neutral)</td>
<td>Subsumption (non-neutral)</td>
<td>Subsumption (total)</td>
<td>Non-subsumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>--------------</td>
<td>-----------------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>---------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td>S,M,F,ANLI</td>
<td>22.2</td>
<td>24.2</td>
<td>23.3</td>
<td></td>
<td>41.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELECTRA</td>
<td>S,M,F,ANLI</td>
<td>3.3</td>
<td>52.5</td>
<td>29.1</td>
<td></td>
<td>53.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBERT</td>
<td>S,M,F,ANLI</td>
<td>68.9</td>
<td>35.4</td>
<td>51.3</td>
<td></td>
<td>18.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>MNLI</td>
<td>27.8</td>
<td>18.2</td>
<td>22.8</td>
<td></td>
<td>37.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S,M,F,ANLI</td>
<td>21.1</td>
<td>33.3</td>
<td>27.5</td>
<td></td>
<td>59.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBERT</td>
<td>S,M,F,ANLI</td>
<td>33.3</td>
<td>38.4</td>
<td>36.0</td>
<td></td>
<td>49.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td>MNLI</td>
<td>36.7</td>
<td>46.5</td>
<td>41.8</td>
<td></td>
<td>40.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S,M,F,ANLI</td>
<td>44.4</td>
<td>23.2</td>
<td>33.3</td>
<td></td>
<td>58.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>MNLI</td>
<td>45.6</td>
<td>26.3</td>
<td>35.4</td>
<td></td>
<td>33.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M,F,ANLI</td>
<td>52.2</td>
<td>37.4</td>
<td>44.4</td>
<td></td>
<td>54.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M,F,Ling,DocNLI</td>
<td>86.7</td>
<td>17.2</td>
<td>50.3</td>
<td></td>
<td>36.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Models’ performance on instances requiring subsumption reasoning.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th>Model’s link</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELECTRA</td>
<td>S,M,F,ANLI</td>
<td><a href="https://huggingface.co/ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli">https://huggingface.co/ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli</a></td>
</tr>
<tr>
<td>SBERT</td>
<td>S,M,F,ANLI</td>
<td><a href="https://huggingface.co/usc-isi/sbert-roberta-large-anli-mnli">https://huggingface.co/usc-isi/sbert-roberta-large-anli-mnli</a></td>
</tr>
<tr>
<td>BART</td>
<td>MNLI</td>
<td><a href="https://huggingface.co/facebook/bart-large-mnli">https://huggingface.co/facebook/bart-large-mnli</a></td>
</tr>
<tr>
<td></td>
<td>S,M,F,ANLI</td>
<td><a href="https://huggingface.co/ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli">https://huggingface.co/ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli</a></td>
</tr>
<tr>
<td>ALBERT</td>
<td>S,M,F,ANLI</td>
<td><a href="https://huggingface.co/ynie/albert-xlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli">https://huggingface.co/ynie/albert-xlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli</a></td>
</tr>
<tr>
<td>DeBERTa-v3</td>
<td>MNLI</td>
<td><a href="https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli">https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli</a></td>
</tr>
<tr>
<td></td>
<td>M,F,ANLI</td>
<td><a href="https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli">https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli</a></td>
</tr>
<tr>
<td></td>
<td>M,F,Ling,DocNLI</td>
<td><a href="https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c">https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c</a></td>
</tr>
</tbody>
</table>

Table 15: Links to the models we use to test on GQNLI.