UNIVERSITY OF HELSINKI
FACULTY OF ARTS
DEPARTMENT OF LANGUAGES

Bachelor's Thesis

Word Class Dropping: Comparing Model and Human Performance on Systematically Corrupted NLI Data

Mitja Sainio

Bachelor's Programme in Languages - Linguistics

Supervisor: Aarne Talman 23.05.2023

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Humanistinen tiedekunta	Koulutusohjelma – Utbildningsprogram – Degree Programme Kielten kandiohjelma

Opintosuunta – Studieinriktning – Study Track
Kielitieteet

Tekijä – Författare – Author
Mitja Sainio

Työn nimi – Arbetets titel – Title
Word Class Dropping: Comparing Model and Human Performance on Systematically Corrupted NLI Data

Työn laji – Arbetets art – Level Kandidaatintutkielma	Aika – Datum – Month and year 5/2023	Sivumäärä – Sidoantal – Number of pages 15

Tiivistelmä – Referat – Abstract

Laajalti käytettyjen luonnollisella kielellä päättelyn (engl. *natural language inference*, NLI) suorituskykyä mittaavien data-aineistojen, kuten MNLI:n ja SNLI:n, on esitetty sisältävän annotaatioartifakteja, joita tunnistamalla Järjestelmä voi saavuttaa korkean suorituskyvyn testeissä. Tämän vuoksi artifakteja sisältävä data ei tarjoa tosiasiallista arviota järjestelmän luonnollisen kielen ymmärtämisen (engl. *natural language understanding*, NLU) kyvyistä. Jotta voidaan määrittää, tarjoaako data-aineisto järjestelmälle riittävän haasteen, on siis välttämätöntä arvioida, missä määrin siinä esiintyy kyseisiä artifakteja.

Yksi hiljattain ehdotettu metodi tämän tekemiseen on sanaluokkapoisto (engl. *word class dropping*), systemaattisen korruption muoto, jossa datasta poistetaan kokonaisia sanaluokkia. Sanaluokkapoiston muokkaama NLI-data voi vaikuttaa ihmisistä järjettömältä, jolloin järjestelmän suorituskyvyn pitäisi olla verrattain matala, sillä NLU-kykyä mitataan suhteessa ihmisten kielenymmärrykseen. Tässä tutkimuksessa kielimallin suorituskykyä MNLI-datalla verrataan ihmisannotoijien suorituskykyyn. Suhteessa korkeampi mallin suorituskyky on merkki siitä, että malli ei hyödynnä NLU:ta NLI-tehtävän suorittamisessa, vaan havaitsee ja hyödyntää annotaatioartifakteja.

Tutkimuksen tulokset osoittavat, että kielimalli saavuttaa 25% ihmisannotoijia korkeamman suorituskyvyn sanaluokkapoiston muokkaamalla datalla. Tämä tarkoittaa, että malli ei osoita NLU-kykyä, vaan käyttää annotaatioartifakteja saavuttaakseen korkeamman suorituskykynsä suhteessa ihmisiin. Sanaluokkapoisto on siis tehokas tapa arvioida NLI-datan laatua, sillä se paljastaa, sisältääkö data-aineisto annotaatioartifakteja.

Avainsanat – Nyckelord – Keywords
luonnollisen kielen käsittely, luonnollisen kielen ymmärtäminen, koneoppiminen

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopiston kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Abstract

Widely-used natural language inference (NLI) benchmark datasets such as SNLI and MNLI are reported to contain annotation artifacts, whose detection enables systems to achieve high performance on these benchmarks. As a consequence, data containing annotation artifacts do not effectively evaluate the natural language understanding (NLU) capabilities of systems. To determine whether a dataset poses an appropriate challenge to a system, it is therefore necessary to assess the extent to which these artifacts are present in it.

One recently proposed method is word class dropping, a form of systematic corruption which involves the removal of entire word classes. NLI data affected by word class dropping can become nonsensical to humans, which should be reflected in lower system performance, since it is measured relative to a human gold standard. In this study, the performance of a language model on MNLI data affected by word class dropping is compared with that of human annotators. Model performance that is superior to that of humans shows that the model is not using NLU in the NLI task, but rather detecting and making use of annotation artifacts.

The results of the study show that the language model achieves 25% higher performance than human annotators on data affected by word class dropping. This means that the model is not displaying NLU, but rather using annotation artifacts to achieve its superior performance in relation to the human annotators. Word class dropping is therefore an effective evaluator of NLI data quality, as it reveals the presence of annotation artifacts.

Contents

Chapter 1

Introduction

Natural language inference (NLI), also known as *recognizing textual entailment*, is a natural language processing (NLP) task that constitutes determining whether a natural-language hypothesis can be inferred from a given premise (MacCartney and Manning, 2008). In practical terms, NLI involves the classification of sentence pairs according to the inferential relation (entailment, neutral, contradiction) between the two sentences. An example of each class of sentence can be seen in Table 1.1. NLI is useful as a benchmark task for evaluating natural language understanding (NLU), which is an important part of such NLP problems as question answering, translation, and dialog (Williams et al., 2018).

Label	**Premise**	**Hypothesis**
entailment	"One of our number will carry out your instructions minutely."	"A member of my team will execute your orders with immense precision."
contradiction	"At the end of Rue des Francs-Bourgeois is what many consider to be the city's most handsome residential square, the Place des Vosges, with its stone and red brick facades."	"Place des Vosges is constructed entirely of gray marble."
neutral	"Conceptually cream skimming has two basic dimensions - product and geography."	"Product and geography are what make cream skimming work."

Table 1.1: NLI sentence pairs from the in-domain MNLI-matched training set.

Recent research has suggested that the usefulness of NLI as an assessor of NLU capability has been overestimated. Gururangan et al. (2018) found that just seeing the hypothesis allows a simple text classifier to identify many of the gold labels in widely-used benchmarks such as the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) corpora. Large-scale NLI datasets, such as the previously mentioned corpora, are created using hypothesis sentences produced by crowdworkers, who adopt certain strategies for generating different kinds of hypotheses. As a result of these strategies, the crowdsourced inference data contains artifacts that suggest the gold label. An NLI system can then succeed at an NLI task simply by detecting these annotation artifacts (Gururangan et al., 2018). This makes it difficult to evaluate the true language understanding capability of a given system.

In response to these findings, Talman et al. (2021) proposed a diagnostic test suite for assessing the quality of NLI datasets. The test suite is comprised of four tests, one of which is their own proposed diagnostics, word class dropping, which involves the removal of entire word classes from a dataset. Sentences affected by this method often appear nonsensical to humans, making it difficult to observe their inferential relations. It should therefore follow that NLI systems should experience similar difficulty, since the performance of such systems is measured relative to language understanding displayed by humans. Thus Talman et al. (2021) argue that if an NLI system performs well on data corrupted by word class dropping, the data does not provide a realistic assessment of the NLU capability of the system.

In order to assess the effectiveness of word class dropping as a test for NLI data quality, this study compares its effect on the performance of a fine-tuned DistilBERT (Sanh et al., 2019) language model and that of human annotators. Performance is measured by evaluating the class label predictions made by the DistilBERT model and the annotators. Word class dropping is carried out by removing the nouns from an evaluation set chosen from the MNLI corpus.

Chapter 2

Related Work

The Stanford NLI (SNLI) Corpus (Bowman et al., 2015) is a widely-used NLI challenge corpus. When released, SNLI presented a significant advancement in NLI data. At 570,152 sentence pairs, it was significantly more extensive than other NLI corpora annotated using non-automatic methods. (Bowman et al., 2015) The Multi-Genre NLI (MNLI) Corpus (Williams et al., 2018) was released a few years later to address the fact that the sentences in SNLI represent only one text genre, image captions. Important NLU phenomena such as temporal reasoning, belief and modality appear infrequently in the caption texts that comprise SNLI, which is reflected in the near-human performance that could be achieved by a model. Since such performance indicates that the corpus does not provide enough of a challenge to serve as an effective NLU benchmark, a new corpora was created to represent a wider variety of text genres, where more diverse NLU phenomena could be observed. (Williams et al., 2018)

Gururangan et al. (2018) suspected that the annotation artifacts present in large-scale NLI corpora such as SNLI and MNLI are due to crowdworkers adopting heuristics for quickly and efficiently generating hypothesis sentences. After discovering that using more than half of the MNLI gold labels and two thirds of the SNLI gold labels could be predicted correctly with the hypothesis alone, they conducted a statistical analysis of the data. The artifacts found by Gururangan et al. (2018) were categorized into lexical choice and sentence length.

According to Gururangan et al. (2018), entailment hypotheses were found to contain generic counterparts to more specific words used in the premise. Words like "dog", "guitar" and "beach" might be replaced with "animal", "instrument" and "outdoors". Additionally, exact numbers seem to be replaced with approximates, and references to specific gender are replaced with words like "person" and "human". Contradiction hypotheses are strongly indicated by the presence of negation words ("no", "never", "nobody"). Some more non-negative words were found to be used often to contradict some specific state. "Sleeping" is used to indicate no other activity can take place, "naked" contradicts descriptions of clothing. Additionally, "cat" was frequent in contradiction sentences, since many premises mention dogs. Neutral hypotheses seem to contain modifiers ("tall", "sad", "popular") and superlatives ("first", "favorite", "most"). Gururangan et al. (2018) speculate that this might result from a strategy for introducing information that is not entailed by premise, but still plausible considering the context. Cause and purpose clauses ("because I learned it", "to catch a stick") also appeared often in neutral hypotheses. Examples of different artifacts can be seen in Table 2.1.

Label	**Premise**	**Hypothesis**
entailment	"That goddamned **hamster** dance."	"That darn little **animal** dance."
contradiction	"Vrenna and I both fought him and he nearly took us."	"**Neither** Vrenna **nor** myself have ever fought him."
neutral	"CDC Centers for Disease Control and Prevention CERT/CC CERTa Coordination Center"	"The CDC in New York was **rounded** and **grey**."

Table 2.1: Sentence pairs from the MNLI-matched training set displaying typical annotation artifacts as described by Gururangan et al. (2018). The entailment hypothesis contains a generic counterpart ("animal") to a word used in the premise ("hamster"), the contradiction hypothesis contains negation words ("neither", "nor") and the neutral hypothesis contains modifiers ("rounded", "grey").

Gururangan et al. (2018) also observed that the number of word tokens in the hypotheses generated by crowdworkers displays correlation with the inference class of the sentence. In SNLI, neutral hypotheses are generally long and entailment hypotheses shorter. This phenomenon is also observed in MNLI, but to a lesser extent. Gururangan et al. (2018) speculate this is due to the presence of more diverse genres.

The diagnostic test suite proposed by Talman et al. (2021) consists of four tests: the hypothesis only baseline (Gururangan et al., 2018; Poliak et al., 2018) that, word-order shuffling (Pham et al., 2021), swapping premises and hypotheses (Wang et al., 2019b), and their own test, word class dropping. Word class dropping is performed by removing one or more word classes from a dataset. Talman et al. (2021) experimented with fine-tuned BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models using data from the ANLI (Nie et al., 2020) and MNLI benchmarks. Some of the models were fine-tuned with the original (unmodified) training data and others with systematically corrupted training data. The models were evaluated on both the original evaluation data and corrupted evaluation data. Talman et al. (2021) compared the effect of removing different word classes and found that content words (adverbs, nouns, verbs) had the greatest impact on model performance. The performance of a BERT model fine-tuned on the original MNLI training data decreased by 17% when nouns were removed from the evaluation set.

Estimates of human performance have been reported on some NLI datasets. Later revisions of the release paper accompanying the General Language Understanding Evaluation (GLUE) (Wang et al., 2018) benchmark include a human baseline established with six NLP researchers annotating 50 sentence pairs sampled from the different datasets that make up GLUE (Wang et al., 2019a). GLUE consists of nine NLU tasks that test the following capabilities: judging grammatical acceptability, sentiment analysis, detecting semantic similarity, and NLI. The tasks are all based on pre-existing datasets and the combined performance scores are meant to provide an overview of the NLU capabilities of a system. One of the NLI tasks in GLUE is performance on the matched (in-domain) and mismatched (out-of-domain) MNLI test sets. Wang et al. (2019a) reported an average accuracy of 80% for the human annotators across all tasks. A Fleiss' κ of 0.73 was reported for inter-annotator agreement. According to the interpretation provided for the measure by Landis and Koch (1977), this reflects a substantial level of agreement.

In addition to the baseline supplied by Wang et al. (2019a), another estimation was carried out separately for each GLUE task by Nangia and Bowman (2020). Their baseline is established using five crowdsourced annotations for 500 sentence pairs from each of the GLUE datasets. Nangia and Bowman (2020) reported an average human accuracy of 87% across all tasks in the benchmark. For the MNLI-matched test set, which is analogous to the MNLI-matched development set used in this study, the reported average accuracy is 92%. No such baseline seems to have been published for the MNLI-matched development set. The crowdsourcing used by Nangia and Bowman (2020) is comparable to the data collection method used for creating the MNLI corpus, so their reported figure for the MNLI-matched test set is arguably a good human baseline for this study.

Chapter 3

Datasets

3.1 The Multi-Genre NLI (MNLI) Corpus

The experiments in this study are performed using data from the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018). MNLI was designed as an improvement over the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). The main advantage that MNLI has over SNLI is the variety of genres represented in the corpus. MNLI consists of 433k sentence pairs from ten different genres of written and spoken modern standard American English.

The sentence pairs in the corpus consist of a premise, a hypothesis and one of three labels: entailment, neutral or contradiction. The premise sentences are selected from freely available text sources. For each premise there is a hypothesis written by a human annotator. The hypotheses are chosen in a way that all three classes are equally represented in the resulting corpus. The label assigned to each sentence pair is validated by four additional annotators, who relabel the sentence pair without seeing the original label. Finally, a gold label is assigned to each pair based on a simple majority vote. Pairs that do not receive a three-vote consensus are marked with '-' in the gold label field. These pairs are not to be used in normal performance evaluation (Williams et al., 2018).

MNLI is partitioned by a train/test/development split. The training set contains sentence pairs from only five of the ten genres in the corpus.[1] The test and development sets are further divided into in-domain and out-of-domain sets that are respectively called "matched" and "mismatched" in the MNLI release paper. The matched sets contain sentence pairs from the five genres that are present in the training set. The mismatched sets contain pairs from the other five, so the

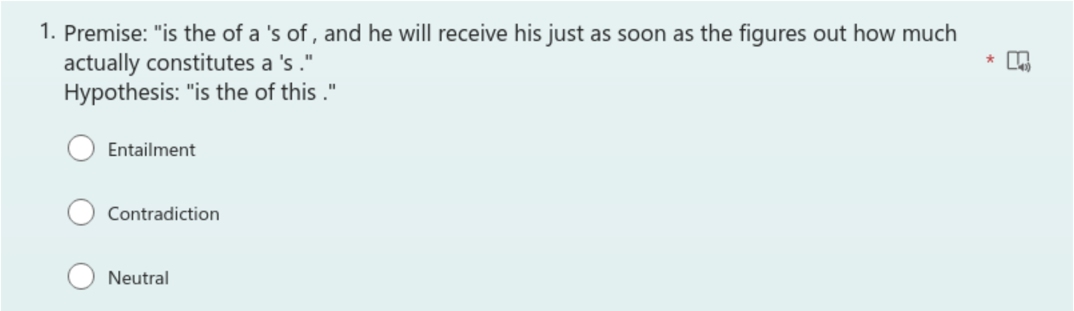
[1] The genres present in the training set are FICTION, GOVERNMENT, SLATE, TELEPHONE and TRAVEL. The genres not included in the set are 9/11, FACE-TO-FACE, LETTER, OUP and VERBATIM.

pairs in these sets do not closely resemble those in the training set. No premise sentence occurs in more than one set.

In this study, the performance of the DistilBERT model and the human annotators is compared using data from the MNLI-matched development set. A matched set is arguably best suited for evaluation, since human annotators are presumed to have familiarity with each of the ten genres represented in MNLI. The development set is used for evaluation because the MNLI test sets are not publicly available.

3.2 Corrupted Evaluation Data

Performance on corrupted data is compared using the MNLI-NOUN-SUBSET dataset created for this study. It is a randomly sampled 50-pair subset of the systematically corrupted MNLI-NOUN (Talman et al., 2021) evaluation set.[2] MNLI-NOUN was created by removing the nouns from the MNLI-matched development set as identified by the Natural Language Toolkit part-of-speech tagger.[3] The nouns are not replaced by a placeholder token, but rather removed altogether from the data. This leaves no syntactic clues concerning the removed nouns, apart from an additional whitespace, as seen in Figure 3.1. These spaces might be interpreted as missing information by human annotator and language model alike.



1. Premise: "is the of a 's of , and he will receive his just as soon as the figures out how much actually constitutes a 's ." *
Hypothesis: "is the of this ."

○ Entailment

○ Contradiction

○ Neutral

Figure 3.1: Screen capture displaying the interface used to collect human annotations in this study.

It bears noting that not all nouns were identified by the tagger: a few nouns such as "misdemeanor", "sundown" and "fisherman" remain in the dataset. While similar evaluation sets exist for other word classes, MNLI-NOUN is used, since Talman et al. (2021) specifically mentioned nouns as a word class whose removal would result in sentences that are nonsensical to humans. The MNLI-NOUN-SUBSET data are quite representative of larger MNLI-NOUN set, since label

[2] https://github.com/Helsinki-NLP/nli-data-sanity-check
[3] https://www.nltk.org/api/nltk.tag.html

distribution in the two sets is very similar (cf. Table 3.1).

Label	**MNLI-NOUN**	**MNLI-NOUN-SUBSET**
entailment	35%	36%
neutral	32%	32%
contradiction	33%	32%

Table 3.1: Label distribution for MNLI-NOUN and MNLI-NOUN-SUBSET.

Chapter 4

Model and Data Collection

4.1 DistilBERT

The model used in this study is the uncased DistilBERT model (DistilBERT-`base`) (Sanh et al., 2019). DistilBERT is a masked-language model pre-trained through knowledge distillation via the supervision of BERT (Devlin et al., 2019). The DistilBERT model was chosen for this study because it is lighter and faster than the BERT-`base` model used by Talman et al. (2021), and therefore requires fewer computational resources. Sanh et al. (2019) report that it retains 97% of BERT performance on GLUE, while being 60% faster and having 40% fewer parameters on the GLUE task STS-B. On the MNLI development set, DistilBERT retains 94.8% of BERT performance. In this study, the pre-trained DistilBERT-`base` model is fine-tuned for three epochs on the MNLI training dataset using a batch size of 16. The training and evaluation scripts are based on those provided in a tutorial by Aarne Talman.[1]

4.2 Human Annotations

Human performance on the corrupted MNLI data is evaluated with an annotation task. An invitation to the online task was distributed via the electronic mailing list and instant messaging group of the linguistics subject association of the University of Helsinki. 15 annotators completed the task. In the task, the annotators are asked to label the 50 sentence pairs in MNLI-NOUN-SUBSET according to their inferential relation (entailment, neutral, contradiction). The instructions for the task contain a brief description and an example sentence pair for each inferential relation. The sentence pairs chosen as examples are the first ones in the MNLI training set that consist of complete sentences. The annotation interface can be seen in Figure 3.1.

[1] `https://github.com/aarnetalman/Notebooks/blob/main/natural-language-inference-with-pytorch-and-transformers.ipynb`

Chapter 5

Evaluation

The prediction accuracy of the fine-tuned DistilBERT-`base` model is evaluated on the original MNLI-matched development set, MNLI-NOUN and MNLI-NOUN-SUBSET. The second row of Table 5.1 shows the DistilBERT model results on the different datasets. There is a noticeable decrease in performance in response to word class dropping. Performance is slightly higher on MNLI-NOUN-SUBSET than MNLI-NOUN. The second column of Table 5.2 shows label distribution for the DistilBERT-`base` model predictions. The model seems to assign entailment classifications in favor contradiction classifications, while the frequency of neutral classifications is equivalent to that found in the gold labels.

Table 5.3 shows a confusion matrix of the predictions made by the DistilBERT-`base` model. The model displays similar accuracy for entailment and contradiction sentence pairs, correctly predicting the labels for nearly two thirds of these. Prediction accuracy for neutral labels is considerably higher. The model classifies correctly over four fifths of all neutral sentence pairs in MNLI-NOUN.

	MNLI-NOUN	MNLI-NOUN-SUBSET	MNLI-matched dev/test
BERT	70%	-	84%
DistilBERT	68%	70%	79%
Human	-	56%	92%

Table 5.1: Prediction accuracy (%) for the BERT-`base` model fine-tuned on the original MNLI-matched training set by Talman et al. (2021), the DistilBERT-`base` model fine-tuned in this study, and the human annotators. Accuracy is reported on MNLI-NOUN, MNLI-NOUN-SUBSET and an original MNLI dataset, which is either the MNLI-matched development set (BERT, DistilBERT) or the test set (human annotators). The figures for the BERT model are reported by Talman et al. (2021). The MNLI-matched test set figure for the human annotators is the baseline established by Nangia and Bowman (2020).

Label	**Gold**	**Model**	**Human**
entailment	36%	40%	28%
neutral	32%	32%	56%
contradiction	32%	28%	16%

Table 5.2: Label distribution for the MNLI-NOUN-SUBSET gold labels and the DistilBERT-`base` model and human predictions on MNLI-NOUN-SUBSET.

		Predicted labels		
		entailment	neutral	contradiction
Actual labels	entailment	**65%**	5%	30%
	neutral	13%	**81%**	6%
	contradiction	21%	14%	**65%**

Table 5.3: Confusion matrix of the DistilBERT-`base` model predictions on MNLI-NOUN-SUBSET.

Human performance is evaluated on MNLI-NOUN-SUBSET using the annotation task. The second row of Table 5.1 shows the human results. The figure for "unmodified MNLI" is not based on an evaluation on the MNLI-matched development set, but rather the human baseline on the MNLI-matched test set established by Nangia and Bowman (2020). Performance for MNLI-NOUN is marked with '-' since human performance was not evaluated on the full set. The human annotators have an average accuracy of 56% on MNLI-NOUN-SUBSET with a Fleiss' κ of 0.40 (cf. Table 5.1). According to Landis and Koch (1977), this reflects a fair level of agreement. Word class dropping seems to affect human performance dramatically, since prediction accuracy on MNLI-NOUN-SUBSET is 39% lower than the human baseline. For comparison, DistilBERT-`base` model performance is only 11% percent lower on MNLI-NOUN-SUBSET than the MNLI-matched development set.

		Predicted labels		
		entailment	neutral	contradiction
Actual labels	entailment	**47%**	47%	6%
	neutral	15%	**81%**	4%
	contradiction	18%	42%	**40%**

Table 5.4: Confusion matrix of the human predictions on MNLI-NOUN-SUBSET.

Over half of all human-predicted labels are neutral (cf. Table 5.2). Consequently, there are fewer entailment and still fewer contradiction classifications in the human predictions than in the gold labels. Table 5.4 shows a confusion matrix of the human predictions. The human annotators label correctly 81% of the neutral sentence pairs, which is similar to the DistilBERT model accuracy on these sentence pairs. Prediction accuracy for human annotators on entailment and contradiction sentence pairs is 33% lower than for the DistilBERT model. Compared to the frequencies of incorrect entailment and neutral classifications, very few sentence pairs are erroneously classified as contradiction by humans.

Chapter 6

Discussion

Word class dropping seems to have a much greater effect on human performance than model performance, as the performance of the DistilBERT model is 25% greater than that of the human annotators. This indicates that NLI data corrupted by noun removal truly are largely nonsensical to humans, as suggested by Talman et al. (2021). Model performance that is superior to that of human annotators does not reflect real language understanding capability, since NLI performance is evaluated in relation to gold labels assigned by human annotators.

While the human annotators perform poorly in comparison to the baseline established by Nangia and Bowman (2020), human performance is still well above the 34% most-frequent-class baseline (cf. Table 3.1). This indicates that although the removal of nouns makes sentences hard to understand for humans, the annotators are still able to make some inferences between the premise and hypothesis. It is even possible that the human annotators leverage some of the annotation artifacts found by Gururangan et al. (2018). In the absence of important content words, human annotators might for example resort to the detection of negation words for identifying contradictions. The entailment artifacts presented by Gururangan et al. (2018) contained many nouns, so entailment sentence pairs should be the most difficult to classify correctly using annotation artifacts. It is difficult to say how the human annotators decided how to classify these sentence pairs.

The high performance of both the DistilBERT model and the human annotators in predicting neutral labels are likely to have different explanations. The DistilBERT model might achieve a high prediction accuracy on neutral sentence pairs by detecting the generally greater length of neutral hypotheses, as suggested by Gururangan et al. (2018), although this phenomenon is reported to be not as prevalent in MNLI as in SNLI. The average length of hypotheses labelled as neutral by the model is ten tokens, which is 25% greater than the subset average of eight tokens and 11% greater than the MNLI-NOUN average of nine tokens. It is unlikely that prediction accuracy on neutral sentence pairs is due to the detection of modifiers and superlatives in neutral hypotheses, as only 42% of the hypotheses labeled as neutral by the model contain these.

Human annotators probably achieve their high accuracy on neutral sentence pairs simply by labeling over half of all pairs as neutral. A possible explanation for the prevalence of neutral classifications in the human predictions is that because many sentence pairs might be difficult to understand, the neutral label is often chosen, since it denotes the absence of any entailment or contradiction between premise and hypothesis. Humans are less likely than the DistilBERT model to label pairs as neutral by detecting modifiers and superlatives, since only 32% of the hypotheses labelled mostly (50% of annotations or more) as neutral by humans contain these. The average length of hypotheses labelled mostly as neutral by humans is ten tokens, which is equal to that of the DistilBERT model. It is therefore possible that hypothesis length also affected human predictions concerning sentence pairs labelled as neutral.

Chapter 7

Conclusion

The discovery of annotation artifacts in widely used NLI benchmarks such as SNLI and MNLI indicate that the success of NLI systems has been overestimated. Systems can achieve high performance on these datasets simply by detecting the artifacts, which are a result of strategies adopted by crowdworkers for generating hypothesis sentences quickly and effectively. While crowdsourcing allows for the creation of large-scale datasets, this creation process has been shown to produce datasets that do not provide a realistic evaluation of the language understanding capability of an NLI system. In response, several diagnostic tests have been created to assess the quality of NLI data by identifying the presence of annotation artifacts.

This study examines the effectiveness of word class dropping (Talman et al., 2021) as a test for NLI data quality. Effectiveness is assessed by comparing the performance of a DistilBERT language model and human annotators on data affected by word class dropping. The results show that human performance on the corrupted data was significantly poorer than that of the DistilBERT model, which supports the presumption in Talman et al. (2021) that NLI data affected by noun removal is nonsensical to humans. This suggests the effectiveness of word class dropping as an assessor of data quality and as a valid addition to the diagnostic test suite proposed by (Talman et al., 2021). NLI systems that achieve high performance on data corrupted by word class dropping must clearly do so using means other than language understanding.

Bibliography

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL `https://aclanthology.org/N18-2017`.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2529310`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 1907.11692, 2019.

Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL `https://aclanthology.org/C08-1066`.

Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pages 4566–4575. Association for Computational Linguistics, 2020.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL `https://aclanthology.org/2020.acl-main.441`.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.98. URL `https://aclanthology.org/2021.findings-acl.98`.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL `https://aclanthology.org/S18-2023`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. NLI data sanity check: Assessing the effect of data corruption on model performance. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 276–287, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL `https://aclanthology.org/2021.nodalida-main.28`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv*, 1804.07461, 2019a.

Haohan Wang, Da Sun, and Eric P. Xing. What if we simply swap the two text fragments? a straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019b. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33017136. URL `https://doi.org/10.1609/aaai.v33i01.33017136`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.