Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6883–6888
Marseille, 11–16 May 2020

Towards the Necessity for Debiasing Natural Language Inference Datasets

Mithun Paul Panenghat, Sandeep Suntwal, Faiz Rafique, Rebecca Sharp, Mihai Surdeanu
Department of Computer Science, University of Arizona
Tucson, Arizona
{mithunpaul, sandeepsuntwal, faizr, bsharp, msurdeanu}@email.arizona.edu

Abstract

Modeling natural language inference is a challenging task. With large annotated data sets available it has now become feasible to train complex neural network based inference methods which achieve state of the art performance. However, it has been shown that these models also learn from the subtle biases inherent in these datasets (Gururangan et al., 2018). In this work we explore two techniques for delexicalization that modify the datasets in such a way that we can control the importance that neural-network based methods place on lexical entities. We demonstrate that the proposed methods not only maintain the performance in-domain but also improve performance in some out-of-domain settings. For example, when using the delexicalized version of the FEVER dataset, the in-domain performance of a state of the art neural network method dropped only by 1.12% while its out-of-domain performance on the FNC dataset improved by 4.63%. We release the delexicalized versions of three common datasets used in natural language inference. These datasets are delexicalized using two methods: one which replaces the lexical entities in an overlap-aware manner, and a second, which additionally incorporates semantic lifting of nouns and verbs to their WordNet hypernym synsets.

1. Introduction

The task of natural language inference (NLI) is considered to be an integral part of natural language understanding (NLU). In this task, which can be seen as a particular instance of recognizing textual entailment (RTE) (Fyodorov et al., 2000; Condoravdi et al., 2003; Bos and Markert, 2005; MacCartney and Manning, 2009), a model is asked to classify if a given sentence (premise) *entails*, *contradicts* or is *neutral* given a second sentence (hypothesis).

In order to advance any task in natural language processing (NLP), quality data sets are quintessential. Some such data sets which have enabled the advancement of NLI (and fact verification) are SNLI (Bowman et al., 2015) MNLI (Williams et al., 2017), FEVER (Thorne et al., 2018), and FNC (Pomerleau and Rao, 2017).

However these datasets are not devoid of biases (subtle statistical patterns in a dataset, which could have been introduced either due to the methodology of data collection or due to an inherent social bias). For example, (Gururangan et al., 2018) and Poliak et al. (2018) show that biases were introduced into the MNLI dataset by certain language creation choices made by the crowd workers. Similarly, Schuster et al. (2019) show that in the FEVER, the REFUTES label (same as the *contradicts* label mentioned above) highly correlates with the presence of negation phrases.

These biases can be readily exploited by neural networks (NNs), and thus have influence on performance. As an example, Gururangan et al. (2018) demonstrate that many state of the art methods in NLI could still achieve reasonable accuracies when trained with the hypothesis alone. Similarly, Suntwal et al. (2019) show that some NN methods in NLI with very high performance accuracies are heavily dependent on lexical information. Further (Yadav et al., 2019) show that this issue is relevant not just in NLI but in other NLP applications such as question answering (QA) also.

This tendency of NNs to inadvertantly exploit such dataset artifacts is likely worsened by the fact that currently the success of NLP approaches is almost exclusively measured by empirical performance on benchmark datasets. While this emphasis on performance has facilitated the development of practical solutions, they may lack guidance as they are often not motivated by more general linguistic principles or human intuition. This makes it difficult to accurately judge the degree to which these methods actually extract reasonable representations, correlate with human intuition or understand the underlying semantics (Dagan et al., 2013).

In this work we postulate that altering these datasets based on lexical importance is beneficial for organizing research and guiding future empirical work. While the technique of delexicalization (or masking) has been used before (Zeman and Resnik, 2008), we have expanded it by incorporating semantic information (the assumption that meaning arises from a set of independent and discrete semantic units) (Peyrard, 2019). Since these techniques are general and compatible with most existing semantic representations, we believe they can be further extended onto datasets used for other NLP tasks. Thus, by enabling integration of these techniques into the training pipeline, we hope to control lexicalization in the datasets which the NN methods possibly depend upon.

In particular, the contributions of our work are:

(1) To motivate further research in using delexicalized datasets, we present the delexicalized versions of several benchmark datasets used in NLI (e.g., FEVER, Fake News Challenge, SNLI, and MNLI), along with the corresponding software for the delexicalization. These datasets have been delexicalized using several strategies which are based on human intuition and underlying linguistic semantics. For example, in one of these techniques, lexical tokens are replaced or masked with indicators corresponding to their named entity class (Grishman and Sundheim, 1996). Our work differs from early works on delexicalization (Zeman and Resnik, 2008) in that we earmark overlapping entities (between the hypothesis and premise) with a unique id. Further we also explore semantic lifting to WordNet

Config.	Claim	Evidence
Lexicalized	With Singapore Airlines, the Airbus A380 entered commercial service.	The A380 made its first flight on 27 April 2005 and entered commercial service on 25 October 2007 with Singapore Airlines.
OA-NER	With `organization-c1`, the `misc-c1` entered commercial service.	The A380 made its `ordinal-e1` flight on `date-e1` and entered commercial service on `date-e2` with `organization-c1`.
OA-NER+SS Tags	With `organization-c1`, the `artifact-c1` `motion-c1` commercial `act-c1` .	The A380 `stative` its `ordinal-e1` `cognition-e1` on `date-e1` and `motion-c1` commercial `act-c1` on `date-e4` with `organization-c1`.

Table 1: Example illustrating our masking techniques, compared to the original fully lexicalized data.

synsets using Super Sense tags (Ciaramita and Johnson, 2003; Miller et al., 1990).

(2) We analyze and examine the effect of such delexicalization techniques on several state of the art methods in NLI and confirm that these methods can still achieve comparative performance in domain. We also show that in an out-of-domain set up, the model trained on delexicalized data outperforms that of the state of the art model trained on lexicalized data. This empirically supports our hypothesis that delexicalization is a necessary process for meaningful machine learning.

2. Masking Techniques

In Suntwal et al. (2019) the authors explore multiple methods for delexicalization. In this work we choose the two of their highest performing masking methods.

2.1. Overlap-Aware Named Entity Recognition

In our overlap-aware named entity recognition (OA-NER) technique, the tokens in a given dataset are first tagged as named or numeric entities (NEs) by the named entity recognizer (NER) of CoreNLP(Manning et al., 2014). Next, to capture the entity overlap between premise and hypothesis sentences, we uniquely enumerate the named entities. Specifically, in the claim (`c`) the first instance of an entity is tagged with `c1`. Subsequently wherever, in claim or evidence, this *same* entity is found next, it is replaced with this unique tag. In contrast, if an entity exists only in evidence, it is marked with an `e` tag. For example `person-c1` denotes the first time the proper name is found in claim, while `location-e3` indicates the third location found in evidence. An example of this is shown in Table 1.

2.2. OA-NER + Super Sense (SS) Tags

Our second type of masking relies on super sense tagging, a technique that uses sequence modeling to annotate phrases with their corresponding coarse WordNet senses (Ciaramita and Johnson, 2003; Miller et al., 1990). In this masking technique, in addition to the OA-NER replacement, other lexical items such as common nouns and verbs are replaced with their corresponding super sense tags. For example, as shown in Table 1, *Airbus A380* is replaced with *artifact* and *enter* is replaced with *motion* (more general abstractions captured by their WordNet hypernym synsets). Further unique overlap is also explicitly indicated, in the same manner as with OA-NER (see Table 1). This technique, specifically, will be used to explore the impact of semantic lifting (i.e., if a more coarse-grained type is possibly less domain dependent) in both the in-domain and out-of-domain settings.

3. Datasets and Methods

For analyzing the effects of various delexicalization operations we chose four popular datasets in NLI: Multi-genre NLI (MNLI) dataset (Williams et al., 2017), Fact Extraction and Verification (FEVER) dataset (Thorne et al., 2018), the Fake News Challenge (FNC) dataset (Pomerleau and Rao, 2017), and the Medical NLI (MedNLI) dataset (Romanov and Shivade, 2018). In MNLI the trained models were tested on both of the validation partitions, the matched partition (which serves as the in-domain partition) and mismatched (the out-of-domain) partition.

We ran these experiments using two high-performing NLI methods: the Decomposable Attention (DA) (Parikh et al., 2016) and the Enhanced Sequential Inference (ESIM) (Chen et al., 2016). All the methods were re-trained out of the box (without any parameter tuning) and tested on the corresponding evaluation partitions of the dataset.

4. Results and Discussion

Table 2 shows the performance of each of these methods (DA and ESIM) on both the lexicalized and delexicalized versions of the MNLI dataset. As shown in the table, the accuracies of both the methods increase when trained with the delexicalized version of the dataset. This aligns with our intuition that delexicalization helps towards de-biasing these datasets, and thus preventing the NN methods from being *distracted* by statistical patterns that are not meaningful for the task at hand.

While the ability of a NN method to derive reasonable representations within the training domain is important, it is also important to have the ability to transfer across domains. Hence, to test the effect of delexicalization on domain transferability we picked one of the methods, decomposable attention (DA), trained it in one domain and tested it in an out-of-domain setting (the DA was chosen since it was provided off-the-shelf with FEVER baseline code). These results can be seen in Table 3. Specifically, the table shows the accuracies in three settings, i.e., when the model

Method	MNLI matched lexicalized	MNLI matched delexicalized	MNLI mis-matched lexicalized	MNLI mis-matched delexicalized
DA	60.95%	64.52%	61.43%	64.86%
ESIM	68.84%	68.14%	69.40%	69.10%

Table 2: Performance of various high performing NN methods over lexicalized and delexicalized versions of the same dataset. 'Matched' is the in-domain partition of the MNLI validation dataset, and 'mis-matched' is the out-of-domain partition. The performance of both the methods remain close to each other in delexicalized and lexicalized versions of the same dataset, which validates that our delexicalization techniques preserve the original information of the text.

Train Domain Eval Domain	MNLI MedNLI	FEVER FNC	FNC FEVER
Lexicalized	51.47%	48.86%	41.16%
OA-NER	51.57%	53.59%	46.47%

Table 3: Performance accuracies of the Decomposable Attention against various masking techniques when tested out-of-domain. The "Train Domain" row indicates the training datasets, while the "Eval Domain" indicates the domain of the corresponding evaluation partitions. For example, one experiment trained the DA method on FEVER and evaluated the resulting model on the testing partition of FNC (column 3).

Masking strategy	FNC	FEVER
Lexicalized	68.99%	83.43%
OA-NER	65.85%	82.31%
OA-NER+SS	45.51%	75.26%

Table 4: Performance accuracies of the Decomposable Attention against various masking techniques when tested in-domain for FNC and FEVER datasets. The "Lexicalized" row shows the accuracies when DA was trained using the corresponding lexicalized data. This demonstrates that while delexicalization with OA-NER maintains the performance, the addition of Super Sense tags reduces the accuracy, emphasizing the fact that the amount of granularity to use is still an open problem.

was trained on MNLI and then tested on MedNLI datasets, when it was trained on FNC and tested on the FEVER datasets, and when the model was trained on FEVER and tested on FNC datasets. Note that in some cases of out-of-domain experiments, the label space of the source domain did not match with that of the target domain. Specifically, while the FEVER dataset consisted of data belonging to 3 classes, the FNC dataset had data points belonging to 4 classes. To enable us to evaluate using the official scoring measures of a target domain, we followed the label alignment approach used in Suntwal et al. (2019). For example, while the data points that belonged to the class *supports* were mapped to *agree*, and *refutes* to *disagree*, the ones in the class *not enough info* were further divided to align with the *unrelated* and *discuss* labels.

The experiments summarized in these tables highlight three observations: (a) The models trained on delexicalized data do not perform worse than the ones trained using lexicalized datasets; (b) in the two settings with texts where the named entities discussed are well covered by the NER used in this work (from FEVER to FNC, and from FNC to FEVER), the results demonstrate that the semantic lifting provided by our OA-NER method improves domain transfer considerably; and (c) we do not see a significant improvement in the transition from MNLI to MedNLI. We suspect this is because of the limited overlap of named entity types between the MedNLI and MNLI. For example while the named entity recognizer (CoreNLP) we used, focuses on PERSON, ORGANIZATION, etc. the MedNLI dataset contains more medical terms related diseases and symptoms. This possibly necessitates the importance of exploring using a domain-relevant NER.

Also this work touches upon questions about which granularity offers a good approximation of semantic meanings. For example, Table 4 shows the performance of Decomposable Attention against various delexicalization strategies. It can be seen that, while the addition of OA-NER tags does not change the performance significantly, the semantic lifting provided by the SS tags decreased the accuracies considerably in both cases (FEVER and FNC). This demonstrates that while generalizing away from lexical items is important, *how much* to generalize remains an open research problem.

5. Qualitative Analysis

To better understand the merits of the proposed delexicalization techniques for bias neutralization, we sample several data points and bin them into two categories: data points that were misclassified by the lexicalized model but are classified correctly by the delexicalized model, and vice versa.

Claim	Correct label	Lex model prediction	Delex model prediction	Bias in the lexicalized model	Most important entities
Did Argentina's President adopt a jewish baby to stop it from becoming a werewolf ?	Discuss	Disagree	Discuss	Argentina - Disagree : 54.78% Argentina - Discuss : 45.22%	Lex model: [n't, n't, Argentina] Delex model: [only, MISCc1, LOCATIONc1]
Trump fired Comey over mishandling of Clinton emails.	Disagree	Agree	Disagree	Clinton - Agree : 63.15% Clinton - Disagree : 36.85%	Lex model: [fired, Clinton] Delex model: [Only, PERSONc1]
U.S. confirms authenticity of second journalist beheading video.	Discuss	Agree	Discuss	U.S - Discuss : 82.35% U.S - Agree : 17.65%	Lex model: [U.S.] Delex model: [LOCATIONc1]

Table 5: Data points in which the model trained on lexicalized data made an incorrect prediction while the model trained on the data delexicalized with OA-NER made the right prediction.

Class	Percentage Misclassified
Agree	46.23%
Discuss	31.84%
Disagree	11.64%
Unrelated	10.27%

Table 6: The percentage of labels that were misclassified by the model that was trained on delexicalized data.

	Agree	Discuss	Disagree	Unrelated
Agree	-	84.44%	14.06%	1.48%
Discuss	39.7%	-	29.03%	31.18%
Disagree	11.76%	79.41%	-	8.82%
Unrelated	10%	80%	10%	-

Table 7: Confusion matrix that highlights the distribution of the incorrectly predicted labels. Rows indicate the incorrectly predicted labels and columns indicate the corresponding original correct labels.

5.1. Positive Changes

Table 5 shows several data points in which the model trained on lexicalized data made an incorrect prediction while the model trained on delexicalized data made the right prediction. As column three shows, the model trained on delexicalized data was able to overcome the bias, which possibly enabled the model trained on lexicalized data to make the incorrect prediction. For example, in the second data point the bias of *Clinton* towards the label *Agree* (i.e., the percentage of data points where the entity *Clinton* co-occurred with the label *Agree*) is 63.15%.

However, the model trained on delexicalized data was able to predict the label with a lower bias (*Disagree* with 36.85%). Further, column four shows the entities in which the corresponding model had placed the highest importance (as derived from their corresponding attention weights) for each of the trained models. In the first example it can be seen that while the model trained on lexicalized data considered the entity *Argentina* to be very important, the model trained on delexicalized data gives importance to the another overlap aware tag (MISCELLANEOUSc1) along with relegating the overlap aware tag of *Argentina* (LOCATIONc1) to lower importance. This indicates a possible decoupling from the bias that the model has achieved, along with promoting other entities more important to the given task.

5.2. Negative Changes

Similar to the above task, but to better understand the limits of the proposed delexicalization techniques, we sampled data points where the model trained on delexicalized data made incorrect predictions. The distribution of such incorrectly predicted labels, is given in table 6. As seen from this table, of all the incorrect predictions made, the maximum were of the label *agree* (46.23%) followed by the labels *discuss*, *disagree* and *unrelated.*

In table 7 we show the distribution of the incorrectly predicted labels against their original labels, in the form of

Claim	HP is better not together – company to split into enterprise and PC/printer businesses.
Evidence	HP's home-focused and business divisions have frequently seemed at odds with each other, and apparently the company agrees. The Wall Street Journal claims that the tech giant is about to split into two companies, one focused on PCs and the other dedicated solely to corporate hardware and services. If the report is accurate, the separation could be announced as early as Monday. The exact reasoning behind the move hasn't been mentioned, but the PC-centric group would be headed by one of its existing executives, Dion Weisler; current CEO Meg Whitman would run the business group and keep an eye on the other company by serving as its chairman of the board. However true the rumor may be, such a move wouldn't be all that surprising – much of the computing industry has been restructuring and rescaling to cope with a world where the PC's role is rapidly evolving . Source : Wall Street Journal
Label	Discuss

Table 8: An example of a data point whose original label was *discuss* and the model trained on overlap-aware delexicalized data predicted as *agree*.

	Agree	**Disagree**	**Discuss**	**Unrelated**
PERSON-c1	62.39%	21.54%	3.627%	12.42%
DATE-e1	57.70%	23.87%	3.17%	15.23%
PERSON-e1	52.52%	21.16%	4.08%	22.22%
Original label distribution	55.05%	20.43%	4.14%	20.37%

Table 9: Relative percentages of some OA-NER tags with respect to their labels along with the original label distribution in the training data.

a confusion matrix. The table focuses solely on the data points that are misclassified, so, unlike standard confusion matrices, the diagonal is empty here.

A case of particular interest in this analysis are the high percentages of wrongly predicted labels (*agree* (84.44%), *disagree* (79.41%), *unrelated* (80%)) when the original correct label was *discuss*. This suggests that the complexity and subtlety of language understanding gets particularly pronounced in cases of classifying neutral (*discuss*) texts. An example of such an incorrectly classified data point is presented in Table 8.

In table 9 we show the relative percentages of the OA-NER tags with respect to their classes along with the original label distribution in the training data. While the previous analyses indicate that we did overcome some biases, this table suggests that we have created new ones, which in-turn affect the quality of the model. This highlights that precisely determining the *right* amount of granularity needed for delexicalization to minimize bias remains an open research problem.

6. Conclusion

In this paper we investigated the need for delexicalization techniques to reduce the potential dependence of NN methods on lexicalized items in NLI tasks. We specifically explore two masking techniques to delexicalize datasets: the first one replaces the lexical entities in an overlap-aware manner, and the second one additionally incorporates semantic lifting of nouns and verbs.

Our experiments show that delexicalization achieves comparative results *in-domain* with the state of the art methods trained on lexicalized data. Importantly, we show that methods trained on delexicalized data transfer considerably better *out-of-domain*, which confirms the importance of delexicalization in NLI tasks for domain transferability. While there is still room for exploration in delexicalization techniques, we present this methodology as a means to conduct more meaningful machine learning experiments. To facilitate this, we release the delexicalized versions of MNLI, FEVER and FNC datasets.

One future direction of interest is exploring the right level of masking granularity needed for delexicalization, which is likely dependent on the task at hand. In addition, we would also like to explore combining delexicalization with knowledge distillation to transfer learning across domains.

7. Language Resource References

All the software and masked datasets for our proposed approach are open-source and publicly available. The datasets are available at
`https://osf.io/szdkn/?view_only=4845641a80624ac493ca14df34e68e8c`
and the code is available on GitHub at:
`https://github.com/clulab/releases/tree/master/lrec2020-masking`.

8. References

Bos, J. and Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 628–635. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., and Inkpen, D. (2016). Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Ciaramita, M. and Johnson, M. (2003). Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 168–175. Association for Computational Linguistics.

Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., and Bobrow, D. G. (2003). Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*, pages 38–45.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Fyodorov, Y., Winter, Y., and Francez, N. (2000). A natural logic inference system. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2)*. Citeseer.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

MacCartney, B. and Manning, C. D. (2009). An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*, pages 140–156. Association for Computational Linguistics.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Peyrard, M. (2019). A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Pomerleau, D. and Rao, D. (2017). Fake news challenge.

Romanov, A. and Shivade, C. (2018). Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Schuster, T., Shah, D. J., Yeo, Y. J. S., Filizzola, D., Santus, E., and Barzilay, R. (2019). Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Suntwal, S., Paul, M., Sharp, R., and Surdeanu, M. (2019). On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, (Short Papers)*, Hong Kong, November. Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yadav, V., Bethard, S., and Surdeanu, M. (2019). Alignment over heterogeneous embeddings for question answering. 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 2681–2691.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.