Numerical Correlation in Text

Daniel Spokoyny
Carnegie Mellon University

Chien-Sheng Wu
Salesforce AI Research

Caiming Xiong,
Salesforce AI Research

Abstract

Evaluation of quantitative reasoning of large language models is an important step towards understanding their current capabilities and limitations. We propose a new task, Numerical Correlation in Text, which requires models to identify the correlation between two numbers in a sentence. To this end, we introduce a new dataset, which contains over 2,000 Wikipedia sentences with two numbers and their correlation labels. Using this dataset we are able to show that recent numerically aware pretraining methods for language models do not help generalization on this task posing a challenge for future work in this area.[1]

1 Introduction

Numerical reasoning tasks are one area where the performance of Large Language Models (LLMs) has not improved as drastically (Rae et al., 2021) as on other tasks. Good performance is critical for many downstream applications in areas such as fact checking, question-answering, or search. Different tasks have been proposed to evaluate the numerical reasoning capabilities of LLMs (Mishra et al., 2022).

We can analyze these tasks along two dimensions: diversity of knowledge required and how solvable the task is. Higher diversity ensures better coverage across different domains while higher solvability yields more interpretable metrics. Mathematical word problems (MWPs) are written in a way that the text of the problem is always sufficient to determine the exact unique answer and are therefore highly solvable. However, they lack in diversity since many MWP datasets are constructed from templates or are even fully synthetic.

In contrast, numerical cloze-style problems requires highly diverse knowledge since they can be easily formed from any text that includes numbers.

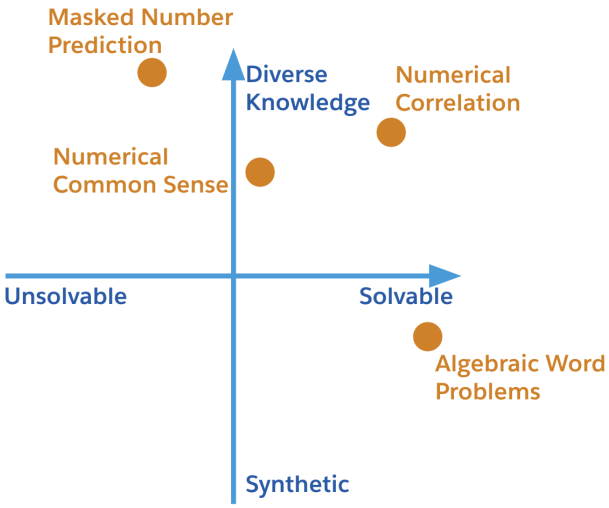

Figure 1: An illustrative plot of certain numerical evaluation tasks along the two dimensions of diversity and solvability. Our aim with numerical correlation is for the task to be both diverse and solvable.

A consequence of formulating cloze-style problems is that many texts do not provide sufficient information to determine the correct answer and have inherent uncertainty which results in a lower solvability. As an example from the NumerSense dataset (Lin et al., 2020), "Some plant varieties can grow up to <mask> feet tall." In Figure 1, we show an illustrative plot of tasks along these two dimensions. A good numeracy evaluation task should be both diverse and solvable.

In this work we propose Numerical Correlation in text, a new task that aims to retain both high diversity and high solvability. Given two numbers in text the task is to predict whether the numbers are positively, negatively or not correlated. For example: “Some plant varieties can grow up to 6 feet tall and require 20 liters of water a month”. We expect a positive correlation between the height of the plant and the amount of water it would need. This shows the key insight that predicting the correlation relationship between two numbers is possible with-

[1]Work completed during internship at Salesforce Research. Please direct correspondence to: dspokoyn@cs.cmu.edu

Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP), pages 33 - 39

# Ex	Text	Label
1.	The president travels on average **30** times a year on Air Force one a Boeing **747**.	Neutral
2.	A **2** bedroom, **1800** square feet house is hard to find in this neighborhood.	Positive
3.	To cook a 20 lb turkey place in the oven for **2** hours at **435** degrees.	Negative

Table 1: Explanations for the three examples: 1) the model of the plane should not change how often the president travels, 2) we expect more bedrooms to increase the size of the house, and 3) we expect an increase of temperature to decrease the cooking time.

out having to exactly predict the missing numbers. The task of numerical correlation requires a variety of commonsense reasoning skills but is trained with a cross-entropy objective and evaluated with a simple accuracy metric. We provide examples of sentences and their labels in Table 1.

Although correlation between two numbers can involve incredibly complex functions, we approximate the correlation to be linear and treat it as a three-way classification. We use a qualification task to select a group of Amazon Mechanical Turk (AMT) labelers and construct a dataset of Wikipedia sentences which contain two numbers and their correlation relationship.

We investigate the performance of four models: two general pretrained language transformers and two numerically aware models on our new dataset in a few-shot setting. When probed on the numerical correlation task we see that all models exhibit a plateau in their performance with only 6% of the training data. Further all models underperform the human baseline in both the finetuning and linear probing setting. Surprisingly, our results also indicate that existing numerically pretraining methods do not result in better performance on the numerical correlation task.

2 Dataset

2.1 Qualification

We used ten handwritten numerical correlation examples and had 100 AMT workers with 99%> approval rate label them. On average each question took around 1 minute to complete. Thresholding on 80% accuracy or above left us with 18 AMTlabelers. Examples and the instructions are shown in the Appendix Table 2 and Figure 5, respectively.

2.2 Annotation

We use the WikiConvert dataset (Thawani et al., 2021) which contains over 900k sentences with at least one measurement in each sentence. We use the three original correlation labels (Positive, Negative, Neutral)[2] and had each sentence labeled by three different AMTlabelers. We selected 1,000 random sentences that contain two measurements and another 1,000 sentences that contain two any two quantities.[3]

We used Krippendorff alpha to measure the inter-annotator agreement and found that the agreement was 0.55 (scale is [-1,1]). We computed an average "Jackknife" F1 score of 77 by choosing one label to be the ground truth and averaging the F1 score of the other two labels. We also observe that the time taken to label each sentence rose to 1.7 minutes on average, likely due to the increased difficulty to ascertain the correlation in random sentences.

2.2.1 Negative

Out of the 2,000 sentences only 42 were found to have a negative correlation which is too few data points to train or evaluate a model. For this reason we experimented with two strategies to generate more negative correlation examples: 1) editing a measurement in real sentence 2) providing a description of a real negative relationship and prompting labelers to provide a sentence as an example. In a small pilot we found that the first strategy was incredibly more time consuming to complete and so we only used the second strategy to generate negative correlation examples. We provided 60 descriptions of negative relationships and asked the three labelers to provide an example for each sentence.[4] In total our dataset consists of 124 sentences with negative correlation, 746 with positive correlation and 1,155 with neutral correlation.

[2]We introduce a fourth label (Unanswerable) which we advised the labelers to use sparingly when they were unsure of the answer

[3]We filtered out sentences that contained dates or where shorter than 64 characters in length.

[4]We hand filtered out sentences that did not properly follow the instructions.

	Test F1 (Neutral, Positive, Negative)				
w/ 10% Train	GenBERT	GeMM	RoBERTa-Base	Bart-Large	Human Jackknife
Linear Probing	33.0 (71.1 / 26.7 / 0.1)	37.9 (72.3 / 41.6 / 0)	23.7 (71.1 / 0 / 0)	**64.9 (76.6 / 57.7 / 60.4)**	~77
Finetuning	62.1 (77.5 / 59.3 / 49.6)	66.7 (77.9 / 65.5 / 56.8)	**69.6 (80.7 / 66.3 / 61.8)**	68.6 (* / * / *)	

Figure 2: Summary of the performance of the four models on the numerical correlation task with 10% of the training data.

3 Experiments

Given a sentence X and two numbers y_1 and y_2 in the text, we define the task of predicting the correlation between the two numbers as a classification task with the label set $C = \{Positive, Negative, Neutral\}$. We compare four models, two general pretrained language models (BART (Lewis et al., 2020) and RoBERTa (Liu et al., 2019)) and two numerically aware models (GeMM (Spokoyny et al., 2022) and GenBERT (Geva et al., 2020)). We conduct few-shot learning experiments where the model is trained on between 1% to 10% of the training data and the remaining data is split into a validation and test set evenly. We train all models with the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 and a batch size of 16. We use the majority vote labeling to choose the final label for each sentence in all subsequent experiments.[5] We report the test F1 scores averaged over 5 initialization seeds.

3.1 Supervised

We conduct few-shot linear probing as well as full finetuning experiments and plot the results in Figure 3 and Figure 4 respectively. For our linear probing experiments we freeze the parameters of the model and only train a linear classifier, $W_\theta \in \mathbb{R}^{d\times 3}$, where d is the hidden size of the model. We observed that BART performed better by a large margin (20 F1) as compared to the second best performing model, GeMM. However, all models experience a plateau in performance after only 6% of the training data.

Unlike the linear probing experiments, when we finetune the models we observe that all models (except GenBERT) converge to similar performance,

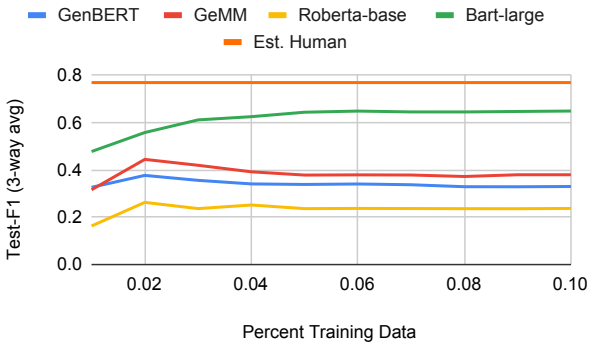

Figure 3: Linear probing experiments with 1% to 10% of the training data.

approximately 10 F1 points below human performance. The poor performance of GenBERT could be explained by the fact that it uses a BERT architecture whilst the other models are based on RoBERTa and BART. We present all of the supervised Test-F1 results with 10% of the training data in Figure 2.

3.2 Unsupervised

Since we observe the actual values of the both numbers we can probe a model in an unsupervised fashion to predict the correlation relationship. We do this by selecting one number (y_1) to be the target prediction and masking it's value in the sentence. We then probe the model to predict the value of the target (y_1) with different values of the other number (y_2). We use GeMM, a numerically pretrained model (Spokoyny et al., 2022) and denominate the model's prediction for the masked value as $\hat{Y}$.

We construct $\mathcal{N}$ examples, $\{X_1, X_\mathcal{N}\}$, by selecting values linearly spaced between $\{y_2 * 0.5, y_2 * 2\}$ and pass each example to the model to predict the $\mathcal{N}$ values of $\{\hat{Y}_1, \hat{Y}_\mathcal{N}\}$. We can then calculate the R-squared values of the linear regression for each

[5] In case of a tie we do not use the sentence in our data.

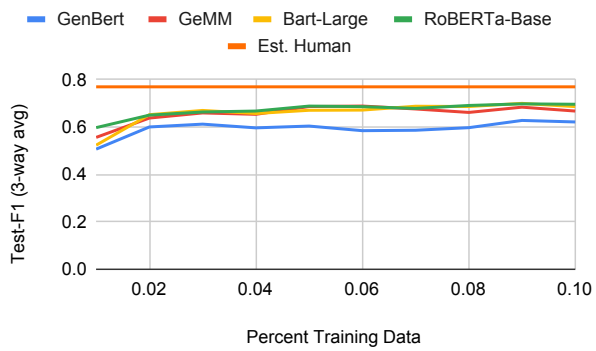

Figure 4: Full finetuning experiments with 1% to 10% of the training data.

pair of numbers in a sentence. We pick a threshold value τ and build a deterministic classifier which predicts "Neutral" if the R-squared value is less than τ, "Positive" if the R-squared value is greater than τ and the slope is positive, and "Negative" if the R-squared value is greater than τ and the slope is negative. When evaluated on a held out test set this classifier performs close to randomly guessing the label.

4 Related Work

4.1 Numerical Reasoning

An active area of research in NLP is focused on solving numerical reasoning tasks. There have been many datasets collected such as AQuA-RAT (Ling et al., 2017), Dolphin18K (Huang et al., 2016), Math23K (Wang et al., 2017), MathQA (Amini et al., 2019) which contain a mathematical question expressed in natural language and an answer. Benchmarks which aim to evaluate the general abilities of LLMs like BIG-bench, have also incorporated numerical reasoning tasks such as arithmetic questions or unit conversion (Srivastava and et al., 2022). To solve these problems a model needs to perform certain necessary calculations to arrive at the answer. Typically the value of the numbers provide no information to help disambiguate the derivation of the solution and can be treated symbolically. One key aspect of these tasks is that there exists no ambiguity in the answer.

4.2 Commonsense Reasoning

Another area of research has focused on cloze-style prediction of numbers in textual contexts. Certain works have limited the output space of numbers to small ranges (Lin et al., 2020), their exponent value (Chen et al., 2019) whilst others have aimed to produce distributions over the entire real number line (Spithourakis and Riedel, 2018; Spokoyny and Berg-Kirkpatrick, 2020). As opposed to the previous section, these tasks commonly do not have a correct answer but are ambiguous. A great advantage of numerical cloze-style reasoning is the ubiquity of available data in different forms and domains. However, it is difficult to measure progress and interpret the evaluation metrics such as likelihood for these types of commonsense tasks.

There are other NLP tasks which have concentrated on the difficulties that arise when numbers are present in a text. Ravichander et al. (2019) proposed EQUATE, a benchmark quantitative reasoning in natural language inference while other works have focused on quantity entailment (Roy et al., 2015). Dubey et al. (2019) built a dataset where the numerical values were useful to predict the sentiment of sarcastic tweets. Sundararaman et al. (2022) proposed a classification task of numbers into entities (Count, Size, Year, Percentage, Date, Age), while similar work has considered the problem of solving numeric Fused-Heads (Elazar and Goldberg, 2019). Our work on the correlation task focuses on a particular relationship between two quantities in text. However there are others potential relationships between numbers in text that could be explored such as causation.

5 Conclusion

We introduced a new task of predicting numerical correlation in text and build an annotated dataset to evaluate models on this task. Using this dataset we show that pretrained language models have poor performance on this task and that current methods to add numerically aware pretraining to models are not effective. We identified that there exists a large gap between human performance and the best supervised model. In the future we hope to expand our annotation to include the slope of the correlation. We believe that predicting both the slope and correlation type of two numbers can be improve interpretability in numerical question answering and commonsense reasoning applications. In future work we also plan to expand the dataset to capture numerical correlation relationships in longer chunks of text such as paragraphs and documents.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Ha-

jishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313, Florence, Italy. Association for Computational Linguistics.

Abhijeet Dubey, Lakshya Kumar, Arpan Somani, Aditya Joshi, and Pushpak Bhattacharyya. 2019. "when numbers matter!": Detecting sarcasm in numerical portions of text. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–80, Minneapolis, USA. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2019. Where's my head? Definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 887–896, Berlin, Germany. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. *ArXiv*, abs/2005.00683.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

Daniel Spokoyny and Taylor Berg-Kirkpatrick. 2020. An empirical investigation of contextualized number prediction. In *EMNLP*.

Daniel Spokoyny, Ivan Lee, Zhao Jin, and Taylor Berg-Kirkpatrick. 2022. Masked measurement prediction: Learning to jointly predict quantities and units from textual context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 17–29, Seattle, United States. Association for Computational Linguistics.

Aarohi Srivastava and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Liyan Xu, and Lawrence Carin. 2022. Improving downstream task performance by treating numbers as entities.

Avijit Thawani, Jay Pujara, and Filip Ilievski. 2021. Numeracy enhances the literacy of language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6960–6967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

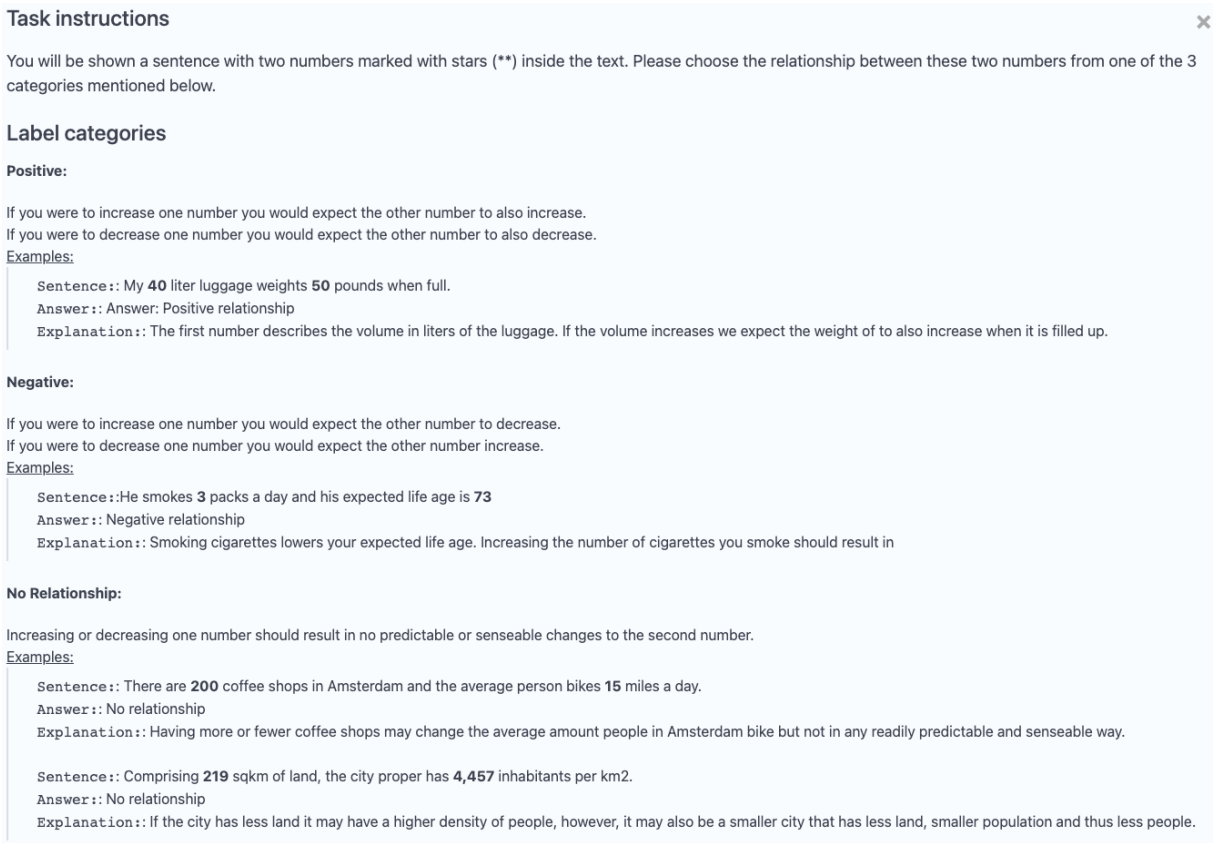

Figure 5: Instructions given to the labellers for the qualification task.

# Ex	Text	Label
1.	I wear my nike shoes out in only **3** months because the soles are only **1/2** an inch thick.	Positive
2.	To cook a 20 lb turkey place in the oven for **2** hours at **435** degrees.	Negative
3.	Jordan trained for his race by running **5** miles at a pace of **10** mph.	Negative
4.	The president travels on average **thirty** times a year on Air Force one a Boeing **747**.	No Relationship
5.	My house has **2** bedrooms and is **1800** square feet.	Positive
6.	Blackthorn was one of **39** original **180** feet seagoing buoy tenders built between 1942-1944.	No Relationship
7.	The family bought a **two** ton pickup truck with 180 hp and a fuel efficiency of **25** miles per gallon.	Negative
8.	My subaru has a **4** cylinder and **150** horse power enginer.	Positive
9.	Like all Type UB III submarines UB-102 carried **10** torpedoes and was armed with a **10** cms deck gun.	No Relationship
10.	The Triple Crown of Canoe Racing consists of three separate marathon races with a total distance of **308** miles over **5** days of racing.	Positive

Table 2: The ten examples used to qualify AMTworkers.