From "good work" to "better work" Generating Phrase Analogies in Large VAEs

Joshua Rivera
New York University
jcr567@nyu.edu

Karan Sagar
New York University
karan@karansag.org

Evan Silverman
New York University
es4753@nyu.edu

Abstract

In general, regularities of sentence embeddings have been studied less than that of word embeddings. One issue is the large sentence space: word embeddings correspond to fixed size vocabularies in their original space, whereas sentence embeddings' original space is significantly larger. Variational autoencoders (VAEs) (Kingma and Welling, 2014) offer a solution by mapping every point in the latent space to one in the original space as a distribution. We study the latent space of OPTIMUS, the first large-scale pre-trained Variational Auto-Encoder (Li et al., 2020), by using so-called "phrase analogies" ($s_a : s_b : s_c : s_d$) for sentences s_a, s_b, s_c, s_d. To see if OPTIMUS preserves relationships, we generate custom datasets similar to Zhu and de Melo (2020) and evaluate the analogy relationship using vector arithmetic on the latent space ($v_{pred} = v_c + v_b - v_a$). We find OPTIMUS does not generate sentences well enough in NLI to preserve relationships in a meaningful way, while syntactic relationships show mixed results depending on category.

1 Introduction

We explore the latent space of OPTIMUS, a new large-scale variational auto-encoder (VAE) (Li et al., 2020). We do so by assessing its performance on sentence analogies, analogies of the form $s_a : s_b : s_c : s_d$ for sentences s_a, s_b, s_c, and s_d. Based on the arithmetic vector transformation between the corresponding latent vectors of s_a and s_b, we apply the transform to the latent vector of s_c and compare the decoded result to the (label) s_d. Specifically,

$$v_{pred} = v_c + v_b - v_a$$

where $enc(s_i) = v_i$. See Table 1 for an example.

Since one of the cited goals of OPTIMUS is to produce a more semantically meaningful latent space than traditional sentence embeddings, we assess its performance on that metric. Our evaluation methods are varied: BLEU, exact-match, manual, and external model evaluation.

Sentence embeddings have been useful in the past for capturing properties of text and have been previously used in downstream tasks (Conneau et al., 2018b). While work has been done to produce universal sentence embeddings (Kiros et al., 2015; Conneau et al., 2018a; Cer et al., 2018), there remains room to explore regularities of these embeddings. Evaluating these regularities can be challenging due to the lack of a direct inverse function from the embedding (latent) space to the sentence space. Word embeddings, on the other hand, can rely on a nearest-neighbor computation from their finite vocabulary. Some approaches to mitigate this include probing (Conneau et al., 2018b) and natural language generation (Kerscher and Eger, 2020; Wang and Lepage, 2020). Zhu and de Melo (2020) also explored sentence analogies using a constructed syntactic and relationship analogy dataset. We take their approach in our dataset construction.

Variational auto-encoders (Kingma and Welling, 2014; Bowman et al., 2016), of which OPTIMUS is one, are auto-encoders that output a relatively low-dimensional latent space to provide a high-level feature representation of the sentence. This can help, for example, guide sentence generation (Li et al., 2020). Because VAEs decode, intuitively, from regions in space, they provide an easy way to decode the latent space back into the sentence space. Until recently, VAEs have not shown comparable performance on standard GLUE benchmarks (Li et al., 2020). However, OPTIMUS, "the first large-scale VAE", has shown good performance here. As a result, we believe exploration of sentence geometry via phrasal syntactic and relationship analogies may allow us to better understand how models like OPTIMUS preserve relationships in their latent space.

Our contributions are the following: we gen-

s_a	s_b	s_c	*(Generated)* s_d
A man stopping on the sidewalk with his bike to have a smoke.	two men stopping on the sidewalk with his bike to have a smoke.	A boy chases after a bird on a pier.	*two boys chase after a bird on a pier.*

Table 1: OPTIMUS evaluation chosen from randomly selected set of 50 (/10k) in "plurals" dataset

erate custom datasets for in-depth evaluation of OPTIMUS's latent space regularity with regard to analogies. We do this along multiple axes. We then evaluate these categories of analogy, with both broad breakdown of syntactic and NLI along with insights into specific subcategories.

The paper is structured as follows: Section 2 presents related work, while Section 3 details dataset construction. Section 4 presents OPTIMUS generation and scoring, while Section 5 discusses results. Section 6 suggests opportunities for further study, while section 7 presents ethical considerations.

2 Related Work

Sentence Embeddings Previous work has explored sentence embeddings in some depth. Skip-thought (Kiros et al., 2015) trains an encoder-decoder architecture, while InferSent (Conneau et al., 2018a) uses NLI data to train a BiLSTM network. Cer et al. (2018) trains a "Universal Sentence Encoder", while Reimers and Gurevych (2019) trains a Siamese network on NLI data that significantly outperformed previous methods.

Evaluation These and other models have been evaluated through probing and related methods. SentEval (Conneau et al., 2018a) provides a toolkit to assess sentence embeddings by using them as features in transfer tasks. Conneau et al. (2018b) introduces a comprehensive set of probing tasks to assess sentence embeddings. Vec2Sent (Kerscher and Eger, 2020) uses an RNN optimized for natural language generation (NLG) to probe sentence embeddings.

VAEs Variational auto-encoders were introduced by Kingma and Welling (2014) as well as Rezende et al. (2014) and expanded upon by Bowman et al. (2016). Bowman et al. (2016) has a Gaussian prior on the latent space, which provides a lower-bound for the log-likelihood of the data. This allows arbitrary points that reasonably fall under the space determined by the prior to decode to valid sentences. Other properties include decoding of any homotopy. Bowman et al. (2016) also dealt with the so-called "KL-vanishing problem"; many hyperparameter choices converged the latent space toward the Gaussian prior. However, Bowman et al. (2016) showed that performance was worse than corresponding LSTMs.

OPTIMUS On the other hand, OPTIMUS Li et al. (2020) was the first large-scale VAE. It uses a BERT-based encoder and GPT-2-based decoder to achieve better results than BERT on GLUE benchmarks. The authors also claim strong latent space manipulation and guided language generation capabilities.

Analogy Datasets We take the same approach here in our dataset construction as in Zhu and de Melo (2020). That is, we use lexical analogies to derive syntactic analogies and use existing relationship analogies from NLI datasets. However, unlike Zhu and de Melo (2020) the language-generation capabilities of OPTIMUS allow us to assess our analogy directly rather than through a nearest-neighbors calculation. We believe this provides a potentially stronger result. Note that Kerscher and Eger (2020) also uses NLG to assess sentence embeddings via analogy, but that trains a separate de-novo decoder for existing embedding methods and comes with caveats similar those of probing methods.

3 Data

3.1 Generation

We use SNLI as a template to generate data for both syntactic and NLI analogies.

Syntactic To construct the syntactic analogies, we combined a modified version of Google's lexical analogies dataset (Mikolov et al., 2013) with SNLI (Bowman et al., 2015). In particular, SNLI sentences served as phrase templates for specific lexical pair replacements from the Google dataset. For example, given the SNLI sentence "The woman is a competitive athlete" and the antonym pair "competitive,uncompetitive", we construct "The woman is a uncompetitive athlete". Each lexical pair, of course, only applies to SNLI sentences where the word appears.

After generating pairs of sentences in this manner with a single lexical relationship (but not necessarily the same words), we take a Cartesian product of all pairs, truncate the rows to about 6 million, and randomly sample 10,000. The output will be quadruplets (s_a, s_b, s_c, s_d) (s_a, s_b) where the analogy $s_a : s_b :: s_c : s_d$ holds. For example, (s_a, s_b) and $(s_c, s_d$ might be phrases where a word has been replaced with its antonym.

This method was applied for antonyms, plurals (both singular to plural and plural to singular), and comparatives. For the latter two, examples of lexical replacement include "some → one" and "more → less".

NLI We took the SNLI test and training sets and limited to only pairs in a "contradiction" or "entailment" relationship. We then joined the set of contradiction pairs to itself and the set of entailment pairs to itself. That is, a quadruplet produced would have (s_a, s_b) and (s_c, s_d) in the same NLI relationship. Similar to above, we limited the joined set to 10 million and randomly sampled 10,000. This was done separately for the SNLI test and training data.

3.1.1 Category Descriptions

Opposites Each resulting sentence contains one word replaced with its antonym with a negation prefix such as "un-", "im-", etc.

Comparative Each resulting sentence contains the comparative form of an adjective and modifies the original sentence such that the standard adjectival form can replace it.

Plural Each resulting sentence contains some singular noun phrase modified such that it has been changed to its plural form and any quantifiers have been modified.

Entailment These sentences are in an entailment relationship as per the NLI task.

Contradiction These sentences are in a contradiction relationship as per the NLI task.

4 Evaluation

4.1 OPTIMUS

Model We used OPTIMUS trained on SNLI with $\beta = 1.0$ and a latent dimension of 768. This was the largest latent space size. While language generation improves with this size, we sacrifice some accuracy on latent space manipulation [1].

Generation We performed the arithmetic analogy operation $v_{pred} = v_c + v_b - v_a$. When decoding the latent output v_{pred}, we used a temperature of 1.0 with no logit limitations (i.e,. `top_p=1.0, top_k=0`).

4.2 Scoring

In order to assess whether the decoded output from our analogy transformation matches the gold standard output, we use a variety of metrics. For syntactic data, we assess a BLEU score, an exact match, and a manually-scored sample of 200 data points. For relationship data, we use the NLI label predicted by a SoTA pre-trained model fine-tuned on NLI data.

BLEU Our BLEU scoring uses the NLTK library's [2] translation BLEU function with the standard 4-gram and stripped punctuation. If either the gold standard sentence or the generated sentence was fewer than 4 tokens, we used a correspondingly smaller n-gram to prevent a BLEU score of 0.

Exact Match The exact-match assesses a binary output between our gold label and prediction through case-insensitive string comparison.

Manual scoring Because assessment of "correctness" in can be challenging to do in an automated way, we take a sample of 50 predicted values from each dataset and assessed them manually.

NLI Model Prediction We merely want the relationship between sentence pairs to be preserved in the NLI analogy transformation. However, our prediction may be semantically and syntactically different from the gold label s_d while still being correct. As a result, we use the predicted label of a near SotA (Conneau et al., 2020) model fine-tuned on NLI data to assess the correctness of our prediction. In particular, we chose an XLM-R fine-tuned on XNLI data (Conneau et al., 2018c)[3]. The model fine-tunes XLM-R on both the MNLI training data and all the XNLI validation and test data. This was

[1] See author's note at https://github.com/ChunyuanLI/Optimus/blob/master/doc/optimus_finetune_language_models.md

[2] https://github.com/nltk/nltk

[3] https://huggingface.co/joeddav/xlm-roberta-large-xnli

Category	Avg(BL)	Med(BL)	Exact
Plurals	.297	.2242	6.32%
Opposites	.1658	.0002	0.1%
Comparatives	.3579	.3117	10.28%

Table 2: Results showing the average and median BLEU scores, along with exact match percentage, among different syntactic categories

chosen for its ease of access and focus on zero-shot classification. Its measured performance on the SNLI dev set was near 86%, so we considered it sufficient for determine the existence of an effect. Note that we do not require the cross-lingual transfer capabilities here.

Since "neutral" pairs were excluded from the NLI dataset, we compare results of both three-way classification (standard NLI evaluation) and two-way classification (the greater logit of $\{entailment, contradiction\}$).

5 Results

Syntactic The overall results can be found in Table 2. Comparatives performed well in general, but had a tendency to reverse the comparative (e.g. old → younger or short → longer). For plurals, OPTIMUS had good performance with natural numbers less than or equal to four. With larger numbers, OPTIMUS might either perform no transformation or substitute the wrong plural number. With multiple word numbers (e.g., two hundred) it shows improvement in accuracy when a dash (-) was used.

The opposites dataset had low scores, potentially as a consequence of the substitutions being semantically illogical. See Section 6 for more details.

NLI While the data in Table 3 appears to suggest greater accuracy than random (where random selection would reflect the SNLI training splits by NLI relationship), manual inspection tells a different story (see qualitative sample in Table 6, appendix).

For one, qualitative analysis of the higher "contradiction" accuracies shows that this accuracy is likely a function of OPTIMUS's predilection to output nonsense. OPTIMUS's "hypothesis" output, even if it makes sense grammatically, often does not make sense logically. XLM-R marks this as a contradiction, perhaps because it often shares, superficially, some noun phrases with the premise sentence.

On the entailment side, while some classifications are legitimate, many occur as a result of OPTIMUS generating truncated hypothesis sentences. Single noun phrases were sometimes scored as "entailment" by XLM-R.

Category	3-way	2-way
SNLI Test Set	–	–
Overall	49.63%	64.61%
Entailment (51.2%)	43.09%	55.62%
Contradiction (48.8%)	56.48%	74.04%
SNLI Train Set	–	–
Overall	48.69%	62.57%
Entailment (51.1%)	42.73%	54.71%
Contradiction (48.7%)	55.07%	70.98%

Table 3: NLI Results: Accuracy of NLI results as measured by XLM-R using both three-way and two-way classification

6 Further Study & Improvements

Datasets The construction of our "opposites" dataset often produced nonsense phrases. For example, lexical replacement of "decided" by an antonym in "one person has decided to take a hike in the mountain" and "one person has undecided..."; the latter sentence is unusual, if not meaningless. We should discount results from the opposites dataset.

On the "plurals" side, we noticed simple template replacement produced some logically incoherent results. For example, "one man wearing a sweater" became "ten men wearing a sweater". OPTIMUS appeared, on manual examination, to bias against producing these logically incoherent sentences. The plural scores may therefore be deflated. We would look for a more sophisticated mechanism to convert singular nouns to plural ones, among other issues.

NLI scoring XLM-R and similar NLI systems are not designed to evaluate machine-generated outputs. The automated NLI scoring often labeled logically incoherent sentences as "contradiction" or "entailment". We would look for further improvements to scoring, perhaps by putting the output of OPTIMUS through a grammaticality or common-sense checker. Ideally, we might have a rules-based step that counterweights the model-based score.

Syntactic Within every syntactic dataset, some decoded results had no modifications compared to s_c. While this should be marked as a total failure, the nonzero BLEU scores could inflate our aggregate metrics.

Transform	Initial	OPTIMUS Prediction
woman to man	the woman is a homemaker the woman is a nurse the woman is a housekeeper the woman is a teacher	the man is a builder the man is a nurse the man is a waiter the man is a teacher
man to woman	the man is a doctor the man is a lawyer the man is a politician the man is an engineer	the woman is a doctor the woman is a lawyer the woman is a politician the woman is an engineer

Table 4: Cursory examination of analogies against gender stereotypes

7 Ethical Considerations

Wikipedia Many potential ethical issues in this paper are inherited from OPTIMUS itself. Pre-training data was generated from English Wikipedia; see Zhao et al. (2018) or Hube and Fetahu (2019) for issues here. For work debiasing sentence representations of BERT, see Liang et al. (2020).

SNLI SNLI, which was used both to fine-tune and generate datasets, has bias issues as seen in Rudinger et al. (2017).

Occupational Stereotypes We ran a quick sanity check to see if OPTIMUS would do semantic transformations into nontraditional occupations by gender. We used a relatively gender-neutral s_a, s_b (s_a =: "the woman walks" and s_b =: "the man walks", and vice versa) to generate output from occupationally stereotyped sentences. As seen in table 4, OPTIMUS converted most sentences correctly. However, note that it did not convert into male "homemaker" and "housekeeper", suggesting some subtle bias does exist that could signal a need for further study.

8 Collaboration Statement

All team members participated in developing core ideas. Joshua and Evan designed and generated the initial datasets, while Karan wrote the code to integrate OPTIMUS. Syntactic analysis and evaluation was done by Joshua and Evan, while Karan focused on NLI evaluation and scoring. Joshua presented our findings to the class. Everyone participated in the writing of this paper and did some work on each part of the project.

We are also grateful for the support and guidance of Alex Warstadt (warstadt@nyu.edu) in the writing of this paper.

9 Code

The code can be viewed on our repository on GitHub. [4]. Data can be found on S3: please contact the authors for access.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. In *In submission to: EMNLP demonstration*, Brussels, Belgium.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2018a. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of*

[4] https://github.com/karansag/phrase-analogies-large-vae

the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018b. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018c. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 195–203, New York, NY, USA. Association for Computing Machinery.

Martin Kerscher and Steffen Eger. 2020. Vec2Sent: Probing sentence embeddings with natural language generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1729–1736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3294–3302, Cambridge, MA, USA. MIT Press.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. pages 5502–5515.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Liyan Wang and Yves Lepage. 2020. Vector-to-sequence models for sentence analogies. In *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, 2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020, pages 441–446. Institute of Electrical and Electronics Engineers Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *EMNLP*, pages 4847–4853.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Category	Avg(MG)	Med(MG)
Plurals	0.44	1
Opposites	0	0
Comparatives	0.8$\bar{3}$	0

Table 5: Results showing the statistics of the manual grading of the syntactic categories.
(MG=Manual Grade)

A Manual Grading

As mentioned in section 4.2, we manually scored the correctness of the predicted solution for 50 randomly chosen pairs for each type of transformation (250 in total). Each of the 3 authors of this paper independently scored them then had the results aggregated.

A.1 Syntactic Grading

For the syntactic analogy pairs datasets, we used a scale of 0-2 where 0=:Completely Wrong, 1=:Partially Correct and 2=:Correct. The grading results, as can be seen in Table 5 , show, in general, OPTIMUS was able to predict many phrases correctly or partially correctly, especially in the plurals dataset, but certainly lacks in some areas leading for the most common score overall being a grade of 0.[5]

A.2 NLI Grading

For NLI, we determined which of the gold labels best describe the results (Entailment, Contradiction, Neutral). Overall, the generated phrases were difficult to grade since they were often only superficially similar, but did not necessarily fit into any of the categories neatly. This is why the results appeared high, but are misleading, since OPTIMUS struggled as well, but defaulted these difficult ones to Contradiction by default.

B Additional Category Details

In Section 3, we mentioned how the the dataset of pairs were generated and what transformation was used between, but each type has additional factors to consider.

B.1 NLI

See Table 6 for a sample of the qualitative examination of NLI results

Label	Premise / *Generated Hypothesis*
entailment	People at a farmer's market. *the people at the farmer.*
entailment	Four African boys playing soccer. *four men are soccer.*
contradiction	A child is surrounded by a field of sunflowers *the child is at home losing*
contradiction	A young child plays on a beach next to an empty chair. *two children sit on a river beach.*
contradiction	A group of people are standing on steps in front of a building. *a man is walking on the steps of a man in the painting.*

Table 6: Sample of NLI results scored as correct by XLM-R

B.2 Plurals

OPTIMUS evaluated various types of plural phrases that had different success scores. The plurals could have a definite or an indefinite/no article as well as have a number modifier. Each variation of a plural phrase had different levels of success with some features, notably low specific quantifiers receiving high median BLEU scores.

See Table 7 for an in depth breakdown of median BLEU scores by plural modifiers.

B.3 Comparative

Comparatives can appear in the grammar in many locations. The most frequent after a noun with a copula ("to be"), but they can also appear as an addition between commas. This structure was accounted for and measured, but it appeared too infrequent to be significant as its own category. However, we did see that it does tend to do the transformation correctly, but generally makes mistakes within the rest of the phrase.

Example:
S_a=:"The house is bigger than the cat"
S_b=:"The house is big"
S_c=:"A ball, taller than a man, is chased"
S_d=:"A ball is tall"

pred=: "a ball boy is tall,"
(The comma was in the predicted output)

B.4 Opposites

Opposites was the hardest to generate and even still has many problems. The main issue, is even if a word has an direct antonym when comparing just

[5] Note that OPTIMUS' generation struggled with illogical semantics making most predicted phrases ungradable

Subcategory	Article	BLEU median
to-two	indefinite	0.3247
to-twenty	definite	0.3078
to-four	indefinite	0.3042
to-three	indefinite	0.2707
from-single	all	0.2691
to-various	indefinite	0.2631
to-some	indefinite	0.2032
to-six	indefinite	0.1821
to-many	indefinite	0.1821
to-five	indefinite	0.166
to-twenty	indefinite	0.1112
to-two hundred	definite	0.0626
to-ten	indefinite	0.0003
to-nine	indefinite	0.0003
to-two hundred	indefinite	0.0003
to-twenty two	indefinite	0.0002
to-one hundred	indefinite	0.0002
to-six	definite	0.0

Table 7: Subcategory breakdown of plural syntactic transformations. "to-<X>" denotes an analogy that should insert <X> in the transformation. Article denotes the article type of the original sentence.

lexical analogies in terms of part of speech and/or semantics, it probably will not work in a phrase due to how and when speakers actually use the words. This lead to many phrases being illogical, however, it did help show that OPTIMUS cannot make correct transformations between arbitrary sets of strings, they must be follow some logic and structure.

C Unexpected Generation

In some cases, OPTIMUS predicted completely unexpected solutions. We assume that, for the most part, it is a byproduct of OPTIMUS' initial training for the Wikapedia corpus, and that modified training would reduce the frequency of such results.

C.1 Irregular characters

Occasionally the prediction would insert tags like <unk> or <bos>, or in other cases, characters from different writing systems like Chinese.

Example:
S_a =: "A dog is fetching a stick out of very clear water."
S_b =: "A dog is fetching a stick out of very unclear water."
S_c =: "A man is about to put on an impressive juggling display."
S_d=: "A man is about to put on an unimpressive juggling display."

pred =: "a man is about to perform an elaborate <unk> on 双."

C.2 Irregular words

In a low number of cases, we discovered meaningless strings that were generated as part of the prediction. Most notably "aEStreamFrameay" which was unexpectedly generated a few times.

Example:
S_a =: "A dog is fetching a stick out of very clear water."
S_b =: "A dog is fetching a stick out of very unclear water."
S_c =: "Puppies at a park during a competitive game of catch."
S_d =: "Puppies at a park during a uncompetitive game of catch."

pred =: "puppies during a game at a park during aEStreamFrameay competition."