# WikiAtomicEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse Supplementary Material

**Manaal Faruqui** **Ellie Pavlick** **Ian Tenney** **Dipanjan Das**
Google AI Language

## 1 Human Annotation

### 1.1 Insertions Flagged as Errors

There are several categories of edit which humans were instructed to flag as errors, including grammatical errors, spam insertions, cases when the base sentence was not a complete sentence, and cases when the insertion phrase was a complete sentence. Several examples of flagged errors are given in Table 2.

### 1.2 Sources of Human Disagreement

Table 1 shows the overall rate at which annotators agree with the original editor, as well as the number of unique opinions observed across all annotators. For example, for an English edit with 5 annotators, if two agree with the original editor and choose index $i$, and the remaining three agree with each other on a different index $j$, $j \neq i$, we mark that example as having two "unique opinions".

| | Agree w/ Orig | Number Unique Opinions 1 | 2 | 3 | ≥4 |
|---|---|---|---|---|---|
| en | 66% | 50% | 38% | 11% | 2% |
| es | 72% | 33% | 45% | 20% | 3% |
| de | 85% | 67% | 25% | 8% | <1% |

Table 1: Measures of human agreement on edit insertion task. The leftmost column shows the percentage of time that individual annotators agree with the original editor. The following columns depict a histogram for the number of unique opinions represented across the human judgments (annotators plus original editor).

More interesting than how often humans disagree on this task is *why* they disagree. To better understand this, we take a sample of 100 English sentences in which at least one human annotator disagreed with the original editor and no annotator marked the edit as an error. We then manually inspect the sample and record whether or not the annotators' choices of different insertion points give rise to sentences with different semantic meaning or to sentences with different discourse structure. Understanding these sources of human disagreement is theoretically important from a computational perspective, as it helps us understand the degree to which these edits can be treated as part of language modeling–i.e. the edits can be understood/predicted by looking at the corpus alone–versus being part the fully grounded language understanding problem, in which "world knowledge" is required in order to fully appreciate the function of the edit.

In particular, we consider three categories for the observed disagreements: 1) the sentences are **meaning equivalent** from a model-theoretic perspective, 2) the sentences contain **significant differences in meaning** from a model-theoretic perspective, or 3) the sentences contain **minor differences or ambiguities** in meaning from a model theoretic perspective but would likely be considered equivalent from the point of view of a layperson. We also include an error category, for when the disagreement stems from a single annotator making an erroneous choice. Examples of each category are given in Table 3.

We take a sample of 100 English sentences in which at least one human annotator disagreed with the original editor and no annotator flagged the edit as "erroneous". We then manually inspect the sample and sort them according to the above-described categories. Table 4 shows our results. We found 49% to be meaning equivalent (i.e. the edit's location effected discourse structure only), and 22% to have significant differences in meaning (i.e. the edit's location fundamentally changed the meaning of the sentence. An additional 13% exhibited minor differences or ambiguities in meaning, and in the remaining 16% of

| |
|---|
| **New article name goes here** The StarBlaster is a ride at Canobie Lake Park.The ride is known as a S&S Power Double Shot Tower Ride at about 80ft of height . (*Insertion phrase not valid*) |
| The connections between neurons form neural circuits **MEOW RAWR RAWR WOOF WOOF SNORT MOO** that generate an organism 's perception of the world and determine its behavior . (*Spam*) |
| The two way trade current stand at $ 340 million **beetween 2010 - 2011** which was described by the Deputy High Commissioner of Bangladesh , Ruhul Alam Siddique as ' negligible when taking into account the combined population ' ( of both countries ) . (*Misspelling in insertion*) |
| It was n't until the 1970 's when there was massive public investment in agriculture **that India became free of famine** . (Base sentence not complete sentence) |
| Sam Woods ( 10 May 1846 – 23 November 1915 ) was a British trade unionist and politician who served as a Member of Parliament ( MP ) in the 1890s **he also was a famous wine connoisseur** . (*Should be separate sentences*) |

Table 2: Examples of edits marked as errors by annotators. Our explanation is given in parentheses.

| |
|---|
| ***Meaning Equivalent*** |
| Paul Wheelahan**, the son of a mounted policeman,** was born in Bombala, South Wales. . . |
| Paul Wheelahan was born in Bombala, South Wales**, the son of a mounted policeman,**. . . |
| ***Minor Difference / Ambiguity*** |
| She moved to Australia **in 1964** and attended the University of New South Wales. . . |
| She moved to Australia and attended the University of New South Wales **in 1964**. . . |
| ***Significant Difference in Meaning*** |
| . . . he and Bart have to share a raft with Ned Flanders and **his youngest son,** Todd Flanders. |
| . . . he and **his youngest son,** Bart have to share a raft with Ned Flanders and Todd Flanders. |

Table 3: Examples of sentences falling into three disagreement categories, defined in terms of the semantics of the edited sentence. See text for a detailed explanation.

cases, the disagreement appeared to be due to annotator error.

| | |
|---|---|
| Meaning Equivalent | 49 |
| Significant Differences in Meaning | 22 |
| Minor Differences/Ambiguities | 13 |
| Annotator Error | 16 |

Table 4: Analysis of 100 sentences for which at least one annotator disagreed with the gold label and no annotator marked as an error.

## 2 Corpus Analysis for Spanish and German

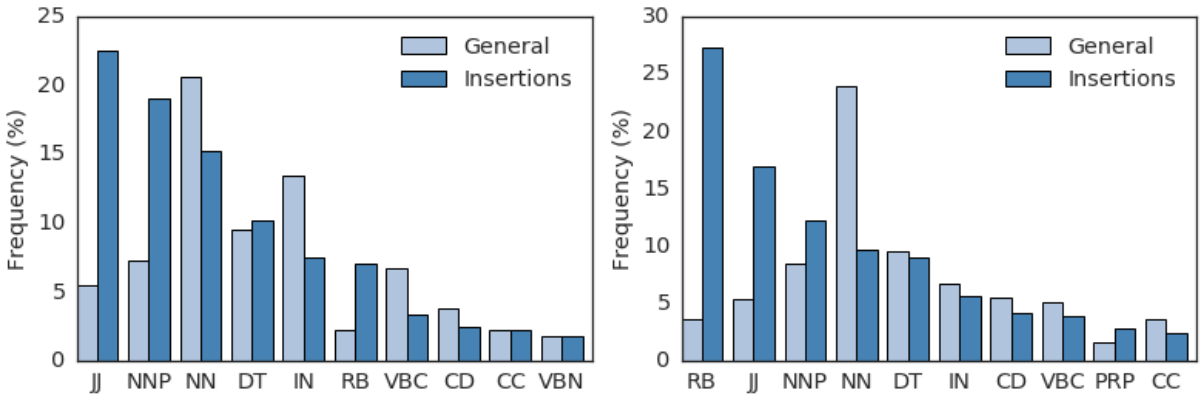
We apply the same methods as the English analysis in the main paper to produce Figure 1 and Table 5. We attribute the comparatively high rate of NN in the German General Wikipedia corpus to the fact that German is morphologically complex and unknown (out-of-vocabulary) words are often tagged as NN.

The higher-than-baseline and lower-than-baseline trends we observed for words in Spanish and German are also similar to what we see for English. English glosses are given for convenience. For Spanish, NNPs seem to be skewed slightly by presence of Wikipedia-specific words/pages, which affect Spanish more than the other languages since Spanish has less overall data. Note that we manually removed words from those displayed if they were due to obvious tagging errors. Specifically: for German, we removed the *"1"* which appeared as the second most common JJ; for Spanish, we removed *"I"* and *"II"* and the third and forth most common JJs. For German, we also omitted redundant conjugations of same word (e.g. *"ehemaligen"/"ehemalige"* were the top 2 JJs).

## 3 Prediction Insertion Location

### 3.1 Language Model Baseline

We train an LSTM language model with the same architecture as in Jozefowicz et al. (2016), who obtained SOTA results on language modeling on the one billion words benchmark for English (Chelba et al., 2013) – a 2 layer LSTM with a hidden size of




(a) Spanish (b) German

Figure 1: Most frequent POS tags for English single-word insertions. Dark blue bars show the relative frequency among inserted phrases and light blue bars show the relative frequency among phrases observed in Wikipedia in general.

| RB | JJ |
|---|---|
| auch/also 113:37 | ehemaligen/former 15:1 |
| jedoch/however 40:7 | deutschen/German 13:4 |
| dann/then 37:6 | neuen/new 8:1 |
| bereits/already 19:3 | heutigen/today 6:1 |
| heute/today 19:3 | politische/political 5:1 |
| zuletzt/last 2:81 | folgenden/following 1:42 |
| nicht/not 16:89 | zentralen/central 1:20 |
| aus/out 1:68 | bzw/or 1:15 |
| so/so 9:33 | eigenes/own 1:13 |
| neu/new 2:24 | freie/free 1:9 |

(a) Spanish

| JJ | NNP |
|---|---|
| gran/great 15:2 | Manuel 6:1 |
| actual/current 10:1 | Francisco 6:1 |
| estadounidense/American 9:1 | Antonio 5:1 |
| Real/Real 11:3 | Juan 6:2 |
| profesional/professional 8:1 | Carlos 5:1 |
| libre/free 2:92 | Wikipedia 1:120 |
| nueva/new 4:20 | Alemania 1:2 |
| personal/personal 1:11 | Francia 1:2 |
| siguiente/following 1:10 | J 1:2 |
| final/last 1:9 | D 1:2 |

(b) German

Table 5: Higher-than-baseline words and lower-than-baseline words for common POS tags in Spanish insertions.

length 8192, word embedding of length 1024. The code for training this model is publicly available.[1] The size of training set for each language model is shown in Table 6, and the tuning was done on 1000 randomly sampled sentences from Wikipedia.

[1] https://github.com/tensorflow/models/tree/master/research/lm_1b

### 3.2 Discriminative Insertion Model

The sentence encoder reads in the word embeddings of each word, denoted as $\mathbf{w}_i$ in the sentence as input.

$$\mathbf{h}^s_1, \cdots, \mathbf{h}^s_{|s|} = f(\mathbf{w}_1, \cdots, \mathbf{w}_{|s|}) \quad (1)$$

Here, $\mathbf{h}^s_i$, is the contextualized word representation for every word $s_i$ in the sentence. Similarly, we also compute representations for the insertion text as $\mathbf{h}^p_1, \cdots, \mathbf{h}^p_{|p|}$, and use the representation of the last word of the phrase as the phrase representation $\mathbf{p} = \mathbf{h}^p_{|p|}$. Now, we take cross-product between the phrase embedding and the contextualized sentence embedding at every word index and then pass it through a feed-forward neural network with sigmoid non-linearity to predict whether or not the insertion should be made at that index:

$$y_i = \sigma(\mathbf{W} * [\mathbf{p} \otimes \mathbf{h}^s_i] + b_i) \quad (2)$$

where, $\mathbf{W}$ and $b_i$ are parameters of the feed-forward neural network. $y_i \in [0, 1]$ is the prediction probability at index $i$. We train the model using cross-entropy loss and compute accuracy by selecting the index with the highest score as the insertion index.[2]

| Language | Dicsr | LM |
|---|---|---|
| de | 124 | 855 |
| en | 629 | 1015 |
| es | 64 | 411 |
| fr | 108 | 510 |
| it | 44 | 296 |
| ja | 96 | 376 |
| ru | 32 | 170 |
| zh | 58 | 406 |
| Avg. | 144.3 | 504.8 |

Table 6: The number of tokens (in millions) for training the discriminative model and the language model.

### 3.3 Experimental Setting

We use FastText (Mikolov et al., 2018; Grave et al., 2018)[3] word vectors of length 300, originally trained on more than 600 billion word tokens each from Common Crawl corpus for each language. For obtaining the sentence representation we tried two RNN models: biLSTM (Schuster and Paliwal, 1997), and GRU (Cho et al., 2014), and a self-attention model: transformers (Vaswani et al., 2017). The hidden unit was fixed to a length of 256 with 2 layers in all cases. We also experimented with whether the word embeddings should be trainable or not. We use dropout for regularization with keep probability tuned on $d = 0.8$ and $d = 0.9$. We use Adam optimizer with gradient clipping (Kingma and Ba, 2015) with a fixed learning rate of $0.001$. The batch size is set to 8 and the model is trained for 5 million steps. We found biLSTM to perform the best as an encoder and keeping the word-embeddings non-trainable better than training them. The no. of tokens per language that was used for training the discriminative model and the baseline language model is shown in Table 6.

## 4 Predicting Insertion Phrases

For the inputs $w_i$, we use the same FastText embeddings as the discriminative insertion model, and for the output text we use a learned wordpiece model (Wu et al., 2016; Schuster and Nakajima, 2012) with a vocabulary of 16000. The sentence encoder is a two-layer bidirectional RNN with highway connections (Srivastava et al., 2015), and the decoder is a single-layer RNN with Luong-style attention (Luong et al., 2015). We use GRU cells (Cho et al., 2014) with orthogonal initialization and a hidden size of 256 units. We use dropout with keep probability $p = 0.8$, a batch size of 128, and train for 700000 steps (approximately 4 epochs) using the Adadelta (Zeiler, 2012) optimizer with gradient clipping and a fixed learning rate of 1.0.

[2] A careful reader might note that we have used sigmoid instead of softmax in the output layer for prediction, this is because there can be multiple valid points of insertion for a text in a given sentence and at inference time, we might want to select all such points for evaluation.

[3] https://fasttext.cc/docs/en/crawl-vectors.html

## References


Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proc. of EMNLP*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proc. of LREC*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proc. of LREC*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *Proc. of ICASSP*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.