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

Manaal Faruqi[*] **Ellie Pavlick**[*] **Ian Tenney** **Dipanjan Das**
Google AI Language

Abstract

We release a corpus of 43 million *atomic edits* across 8 languages. These edits are mined from Wikipedia edit history and consist of instances in which a human editor has inserted a single contiguous phrase into, or deleted a single contiguous phrase from, an existing sentence. We use the collected data to show that the language generated during editing differs from the language that we observe in standard corpora, and that models trained on edits encode different aspects of semantics and discourse than models trained on raw, unstructured text. We release the full corpus as a resource to aid ongoing research in semantics, discourse, and representation learning.

1 Introduction

Written language often undergoes several rounds of revision as human authors determine exactly what information they want their words to convey. On Wikipedia, this process is carried out collectively by a large community at a rate of nearly two revisions per second (Yang et al., 2017). While Wikipedia's revision history contains arbitrarily complex edits, our corpus and analysis focuses on *atomic insertion edits*: instances in which an editor has inserted a single, contiguous span of text into an existing complete sentence (Table 1). This restriction allows us to make several assumptions which we believe make the data an especially powerful source of signal. Namely, we can assume that 1) some information was not communicated by the original sentence, 2) that information *should* have been communicated (according to a human editor), and 3) that information *is* communicated by the inserted phrase. Thus, we believe that a large data set of such edits is inherently valuable for researchers modeling inference and discourse and that the data can yield insights about representation at both the phrase and the sentence level.

Adding new relevant information
She died there **in 1949** after a long illness.
Refining claim/Resolving ambiguity
Finlay announced he'd be on the 1000th episode of "WWE Monday Night Raw"**, but he wasn't.**
Improving Discourse/Fluency
It is **also** being evaluated as a potential biological control for the invasive plant . . .

Table 1: Example atomic insertions (in bold) from the corpus and the types of semantic and discourse phenomena that such edits capture.

We mine Wikipedia edit history to create a corpus of 43 million atomic insertion and deletion edits covering 8 languages. We argue that the corpus contains distinct semantic signals not present in raw text. We thus focus our experiments on answering the following questions:

1. How is language that is inserted during editing different from general Wikipedia text?
2. What can we learn about language by observing the editing process that we cannot readily learn by observing only the final edited text?

Specifically, the contributions of this paper are:

- A new corpus (WikiAtomicEdits) of 26M atomic insertions and 17M atomic deletions covering 8 languages (§3 and §4): `http://goo.gl/language/wiki-atomic-edits`.
- Linguistic analysis showing that inserted language differs measurably from the language observed in general Wikipedia text (§5).

[*] Both authors contributed equally.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 305–315
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

- Language modeling experiments showing that models trained on WikiAtomicEdits encode different aspects of semantics and discourse than models trained on raw, unstructured text (§6).

2 Theoretical Motivation

We borrow the idea of an *atomic edit* from prior work in natural language inference, specifically natural logic (Lakoff, 1970; Van Benthem, 1986). MacCartney (2009) defines an atomic edit e applied to a natural language expression s as the insertion, deletion, or substitution of a sub-expression p such that both the original expression s and the resulting expression $e(s)$ are well-formed semantic constituents. E.g. $s =$ *"She died from an illness"*, $p =$ *"in 1949"*, and $e(s) =$ *"She died in 1949 from an illness"*. This formulation is desirable because it exposes a relationship between the surface form and the semantics of natural language while remaining agnostic about the underlying semantic representation. That is, the difference in "meaning" between s and $e(s)$ is exactly the "meaning" of p (in context), regardless of how that meaning is represented.

We adopt this philosophy in creating our corpus. We focus our analysis specifically on atomic insertion edits. We make the assumption that editors on Wikipedia are attempting to communicate true information[1] and to do so effectively. Insertion edits are thus particularly interesting because the underlying generation process admits the following assumptions:

1. The original sentence s does not effectively communicate some piece of information.

2. A reasonable reader of s would like/expect this information to be communicated.

3. This information is communicated by the inserted phrase p (in the context of $e(s)$).

We therefore believe that the supervision provided by insertion edits can improve our understanding of semantics, discourse, and composition, and that the data released will be valuable for research in these areas. The goal of our experiments is to establish that the signal provided in these edits is distinct from what one could easily obtain given currently available text corpora.

Language	Ins	Del	Total
German	3.3	1.9	5.2
English	13.7	9.3	23.0
Spanish	1.4	0.9	2.3
French	2.0	2.0	4.0
Italian	1.0	0.6	1.6
Japanese	2.2	1.3	3.5
Russian	1.4	0.9	2.3
Chinese	0.7	0.4	1.1
Total	25.7	17.2	42.9

Table 2: The number of instances (in millions) of atomic insertions/deletions for each language.

3 WikiAtomicEdits: Corpus Creation

3.1 Extracting Edits

Wikipedia edits can be accessed through Wikipedia dumps. The edits are stored as diffs on the entire Wikipedia page, meaning some processing is required to reconstruct the changes that were made at the sentence level. We use historical snapshots of each Wikipedia document and compare against subsequent snapshots to extract sentence-level edits. We strip the HTML tags and Wikipedia markup of the page and then run a sentence splitter (Gillick, 2009) to obtain a list of sentences for each snapshot. Rather than run a full, quadratic-time (Myers, 1986) sequence alignment to compare the two lists of sentences, which is infeasible for long articles, we propose an efficient precision-oriented approximation.

Given n sentences in one snapshot ("base") and m sentences in a subsequent one ("edited"), we assume that most edits are local and restrict our attention to a fixed-size window. For each sentence s_i in the base snapshot, we compute pairwise BLEU scores (Papineni et al., 2002) between s_i and the sentences $\{t_j\}_{j=i-k}^{i+k}$ ($k = 5$) in the edited snapshot. We consider the sentence with the highest BLEU score in this window as a candidate. If the sentences are not identical and the difference consists of an insertion or deletion of a single contiguous phrase[2], we add this example to the corpus. For each article, we run this algorithm over the most recent 100,000 snapshots as of February 2018. We extract edits for 8 languages. Statistics are shown in Table 2.

[1] This is true for the majority of edits, although about 13% of edits are "spam" (§4.3).

[2] We use the Python 2.7 `difflib` library to compute a minimal diff at the byte level.

3.2 Insertions vs. Deletions

We use the algorithm described above to extract both atomic insertions and atomic deletions. However, we chose to omit the deletions from our linguistic (§5) and language modeling (§6) analyses for two reasons. First, our intuition is that spans which are deleted by an editor are more likely to be "bad" phrases (e.g. spam, false information, or grammatical errors introduced by a previous editor). To confirm this, we manually inspected 100 of each type of edit. We found that indeed deletions contained a higher proportion of spam text and malformed English (16/100) than did insertions (7/100). Second, while insertions permit a clean set of assumptions about the relationship between the original sentence and the edited sentence (§2), it is more difficult to make generalizations about deletions. Specifically, it is difficult to say whether the original sentence *should not* communicate the information in the deleted phrase (i.e. the phrase contains false, irrelevant, or otherwise erroneous information) or rather the original sentence/surrounding context *already communicates* the information in the deleted phrase (i.e. the deleted phrase is redundant). As such, deletions are a noisier target for analysis. Nonetheless, we recognize that the deletions provide a related and likely useful signal. We thus include deletions in our corpus but leave their deeper linguistic analysis for future work.

4 Corpus Quality & Reproducibility

4.1 Annotation

Given the data collected as above, we now investigate whether the extracted edits are sufficiently clean to be useful for computational language analysis and modeling. To do this, we focus our attention specifically on the English, Spanish, and German subcorpora, as these are languages for which we could find a sufficient number of native speakers to perform the necessary annotation for our analysis. Thus, the discussion and results in this section may not be representative of the other languages in the corpus.

We are interested specifically in two questions. First, we want to measure the overall corpus quality: how many of the inserted phrases represent meaningful edits and how many are simply noise (e.g. from editor or preprocessing error)? Second, we want to understand, at least in part, the reproducibility of the corpus: could we expect a different group of human editors to produce the same edits as those observed?

	en	es	de
No Error	78%	55%	85%
Possible Error	13%	30%	9%
Clear Error	9%	15%	6%

Table 3: Corpus quality for three languages for which we were able to collect annotations. "No Error"/"Clear Error" means annotators agreed unanimously that the edit was/was not an error; "Possible Error" means annotations were mixed.

To address these questions, we collect annotations in a semi-generative manner. Each annotator is shown a sentence s and a phrase p to be inserted, and is asked to insert p into s in order to form a new sentence $e(s)$. If s is not a complete and well-formed sentence, or if there is no location at which p can be inserted such that $e(s)$ would be a complete and well-formed sentence, annotators are instructed to mark the edit as an error. We use the "error" labels in order to study corpus quality (§4.2) and use the annotators' insertion location to estimate reproducibility (§4.3).

We collect labels for 5,000 English edits, and 1,000 each for Spanish and German edits using a crowd-sourcing platform. We collect 5-way annotations for English and 3-way annotations for Spanish and German. Our choices of languages and the differing levels of redundancy were due to availability of annotators. We will release these 7,000 edits and their annotations with the corpus.

4.2 Corpus Quality

To measure corpus quality, we compute the proportion of edits marked as errors by our annotators. Table 3 shows our results. For English, in 78% of cases our annotators agreed unanimously that p could be inserted meaningfully into s (55% for Spanish; 85% for German). These numbers reassure us that, while there is some noise, the majority of the corpus represents legitimate edits with meaningful signal. For more discussion of the errors refer to Supplementary Material.

4.3 Agreement and Ambiguity

We next explore the extent to which the edits in the corpus are reproducible. In an ideal world, we would like to answer the question: given the same original sentences, would a different group of human editors produce the same edits? Answering this directly would require annotators with domain expertise and is infeasible in practice. However, we can use our crowdsourced annotation to answer a restricted variant of this question: given a sentence s and an insertable phrase p, do humans agree on where p belongs in s? We can measure agreement in this setting straightforwardly using exact match, and can interpret human performance as that of a "perfect" language model. I.e. we can interpret disagreement as evidence that reproducing the particular edit is dependent on exogenous information not available in the language of s alone (e.g. knowledge of the underlying facts being discussed, or of the author's individual style).

Based on our annotation experiment, we find that individual annotators agree with the original editor 66% of the time for English, 72% for Spanish, and 85% for German.[3] More interesting than how often humans disagree on this task, however, 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 simply to sentences with different discourse structure.

In particular, we consider three categories for the observed disagreements: 1) the sentences are **meaning equivalent** from a truth-conditional perspective, 2) the sentences contain **significant differences in meaning** from a truth-conditional perspective, or 3) the sentences contain **minor differences or ambiguities** in meaning (but would likely be considered equivalent from the point of view of most readers). 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 4. Note that the assessment of the truth conditions of the sentence and their equivalence is based on our judgment, and many of these judgments are subjective. We will release our annotations for this analysis with the corpus, to enable reproducibility and refinement in future research.

Table 5 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 cases, the disagreement appeared to be due to annotator error.

5 Corpus Linguistic Analysis

We now turn our attention to exploring the language in the corpus itself. In this section and in §6, our focus is on the questions put forth in the introduction: 1) how does the language that is inserted during editing differ from language that is observed in general? and 2) what can we learn about language by observing the editing process that we cannot readily learn by observing only raw text? Here, we explore these questions from a corpus linguistics perspective. The analysis in this section is based predominantly on the 14M insertion edits from the English subcorpus (Table 2).

5.1 Manual Categorization of Insertions

We first characterize the types of insertions in terms of the function they serve. Manually inspecting the edits, we identify four high-level categories. Note that we do not intend these categories to be formal or exhaustive, but rather to be illustrative of the types of semantic and discourse phenomena in the corpus: i.e. to give sense of the balance between semantic, pragmatic, and grammatical edits in the corpus. The categories we identify are as follows:

1. **Extension**: the explicit addition of new information that the author of the original sentence did *not* intend to communicate.[4]

2. **Refinement**: the addition of information that the author of the original sentence either intended to communicate or assumed the reader would already know. This category includes hedges, non-restrictive modifiers, and other clarifications or scoping-down of claims.

3. **Fluency / Discourse**: grammatical fixes, as well as the insertion of discourse connectives (*"thus"*), presuppositions (*"also"*), and editorializations (*"very"*).

[3] We consider cases which the annotator marks as "error" to be a disagreement with the original editor.

[4] Whether or not the author "intended" to communicate something is based on our judgment. Since this annotation is intended to be exploratory, we allow a degree of informality.

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 4: Examples of sentences falling into three disagreement categories, defined in terms of the truth conditions of the edited sentence. See text for a more detailed explanation.

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

Table 5: Analysis of 100 sentences for which at least one annotator disagreed with the gold label and no annotator marked as an error.

4. **Referring Expressions** (RE): changes in the name of an entity that do not change the underlying referent, such as adding a first name (*"Andrew"*) or a title (*"Dr."*). RE edits could fall within our definition of "refinement", but because they are especially prevalent we annotate them as a separate category.

We also define an **Error** category for spam, vandalism, and other "mistake" edits.

We manually categorize 100 randomly-sampled edits. The breakdown is shown in Table 6. In our sample, the majority (43%) were extensions, and the second most frequent where refinements (24%). No single category dominates and all are well-represented, suggesting that a variety of phenomena can be studied using this corpus.

5.2 Comparing Insertions to Raw Text

Understanding the high-level functions of edits, as above, provides some insight into the type of linguistic signals contained in the data. However, we are particularly interested in whether the language used for these functions is noticeably different from general Wikipedia text. That is: it is not obvious that the language humans use to e.g. extend or refine an existing claim should necessarily be different, in aggregate, from the language used to formulate these claims in general. We thus explore whether this is the case.

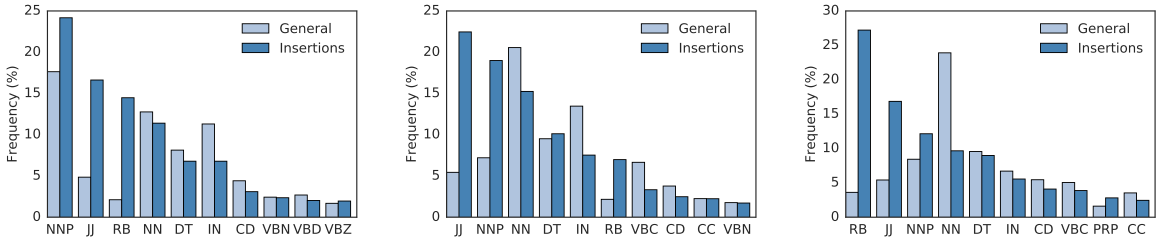
We first compare the distribution of parts of speech observed for the inserted phrases to the distribution of parts of speech that we observe in Wikipedia overall–i.e. in the sentences appearing in the final, published version of Wikipedia, not only the edit history. In order to compare the relative frequencies in a straightforward way, we look only at edits in which a single word was inserted.[5] Figure 1 shows our results for English, Spanish, and German. We see, for example, that in English, adjectives and adverbs combined make up nearly 30% of all inserted words, three and a half times higher than the frequency of adjectives/adverbs observed in the general Wikipedia corpus, and that proper nouns are inserted at a higher rate than would be suggested given their base frequency.

Looking more carefully, we see that the nature of the edits for each part of speech are qualitatively different as well. To explore this further, we look at which words appear at substantially higher rate as insertions than they do in the general Wikipedia corpus. We compute this as follows: for a word w with part of speech pos, we compute the number of times w occurs as an insertion per thousand insertions of any word of type pos, and compare this to the rate of occurrence of w per thousand occurrences of any word of type pos within the general Wikipedia corpus. Table 7 shows our results for English (Spanish and German are given in the Supplementary Material). In particular, we see that many words which are inserted at a significantly higher-than-baseline rate reflect "refinement"-type edits. Many of these are words which the original author may have commu-

[5] In our corpus 30% of inserted phrases are a single word, and 70% are less than five words. We compared frequencies for longer POS sequences as well, but it did not yield particular insight over looking at single POS tags.

Category	Freq.	Example
Extend	43%	The population was 39,000 in 2004**, measured at 29,413 at the 2011 Census**.
Refine	24%	. . . began an investigation into Savile 's **apparent** history of abuse. . .
RE	11%	**Andrew** Sugerman has been involved in the production of motion pictures. . .
Fluency	9%	Philippine coconut jam**, meanwhile,** is made from coconut cream. . .
Error	13%	The team **are well - known as a loser team in the past 5 years.The team** is. . .

Table 6: High-level categories into which we manually characterize edits, to understand the variety of phenomena captured by the corpus. Frequencies are based on our annotation of a sample of 100 edits.



(a) English

(b) Spanish

(c) German

Figure 1: Most frequent POS tags for English, Spanish, and German 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.

nicated implicitly but the editor chose to state explicitly, such as whether or not a person is a *"current"*/*"former"* public figure[6] or is *"famous"*. On the other hand, words which are inserted at a significantly lower-than-baseline rate are those which would be unlikely to be omitted by the original author. For example, if an event is famously the *"first"* or the *"only"* one of its kind, it is highly unlikely for the original author describing that event not to use these words explicitly.

6 Language Modeling Analysis

We next explore the corpus from a language modeling perspective. Again, we are interested in understanding how the signal captured by the editing process is distinct from that captured by the final edited text alone, and in characterizing the types of signals we can learn from modeling the insertions directly. We investigate this through two tasks: first, given a sentence s and insertable phrase p, predict the index i at which p should appear in s (§6.1), and second, given a sentence s and an index i, generate candidate phrases that would be appropriate to insert into s at i (§6.2).

	NNP	JJ	RB
Over	City 16:2	former 34:6	also 187:91
	Sir 7:1	current 11:2	currently 40:7
	US 7:1	famous 9:2	very 24:11
	John 6:3	professional 10:3	then 45:33
	Roman 4:1	fictional 5:1	allegedly 10:1
Under	New 1:5	first 9:29	not 35:96
	United 2:5	only 2:20	first 9:68
	I 2:4 9	other 12:26	all 1:35
	de 2:4	total 2:13	only 22:47
	School 1:3	such 3:13	about 4:29

Table 7: Words that appear as insertions at significantly higher rates (top row) and significantly lower rates (bottom row) than their rate of occurance in Wikipedia in general. We compute "rate" as simply the observed occurrence of the given word per thousand occurrences of any word with the given POS. Table shows each word followed by (rate as insertion):(rate in general)

6.1 Predicting Insertion Locations

Task. This task–given a phrase p and a sentence s, choose the best index i in s at which to insert p–is identical to the task we asked humans to perform in §4. We consider two simple models for performing this task: a basic language model and a discriminative model trained on the insertion data. We report performance as overall accuracy. We analyze whether a model which is trained to

[6] We note that the addition of *"former"* is likely tied to changes in the real world (Wijaya et al., 2015).

model insertions directly captures something different than a general language model in terms of the types of errors each model makes.

Models. We evaluate two models. First, we evaluate a standard language modeling baseline (**General LM**), in which we simply insert the phrase p at every possible point in s and chose the index which yields the lowest perplexity. We use the LSTM language model from Jozefowicz et al. (2016), which obtained SOTA results on language modeling on the one billion words benchmark for English (Chelba et al., 2013). We train this language model for each language on an average of $\sim$ 500 million tokens from Wikipedia. Second, we evaluate a discriminative model specifically trained on the insertion data (**Discriminative Model**). This model represents the base sentence using a sentence encoder that produces a context-dependent representation of every word index in the sentence, and then at test time, compares the learned representation of each index with the representation of the phrase p to be inserted. We use a 256-dimensional 2-layer biLSTM encoder, initialized with FastText 300-dimensional word vectors (Mikolov et al., 2018; Grave et al., 2018).[7] We hold out 50K and 10K insertion edits for each language as development and test sets, and use the remaining edits (insertions and deletions) as training data. This provides us with at least 1 million examples for training in each language (cf. Table 2). See Supplementary Material for additional details.

Results. Table 8 shows the accuracy of each model for each language. We see that the discriminitve model trained on insertions directly performs better than the general LM by at least 1% absolute accuracy on every language, and by 3.8% absolute on average. It is worth emphasizing that this performance improvement is despite the fact that the general LM was trained with, on average, four times the number of tokens[8] and is a much larger model–the general LM has $\sim$ 2 billion parameters (Jozefowicz et al., 2016) compared to $\sim$ 1 million for the discriminative model.

	General LM	Discr. Model
German	68.1	72.9
English	58.7	68.4
Spanish	67.0	70.1
French	69.9	73.4
Italian	69.0	72.9
Japanese	73.0	74.2
Russian	72.9	74.3
Chinese	65.5	68.9
Average	68.0	71.8

Table 8: Insertion accuracy on the test set.

More interesting than raw performance is the difference in the types of errors that the models make. For each model, we take a random sample of 50 examples on which the model made a correct prediction and 50 examples on which the model made an incorrect prediction. We annotate these 200 examples[9] according to the edit type classification discussed in §5.1. Table 9 shows the results. We find a significant difference[10] ($p < 0.01$) between the types of edits on which the General LM makes correct predictions and the types on which it makes incorrect predictions. Specifically, the General LM appears to be especially good at predicting location for fluency/discourse edits, and especially poor at predicting the location of refinement edits. In contrast, we do not see any significant bias in the errors made by the discriminative model compared to its correct predictions ($p = 0.23$). We interpret this as evidence that the insertion data captures some semantic signal that is not readily gleaned from raw text corpora.

6.2 Predicting Insertion Phrases

Task. In a final set of experiments, we explore a generative version of the language modeling task: given a sentence s and an specified index i, generate a phrase p which would be appropriate to insert into s at i. We are interested in what such a model can learn about the nature of how sentences are extended: what type of information would be relevant from a semantic perspective, and natural from a discourse perspective to insert at a given point? We train two models for this task, one trained on the training split of the WikiAtomicEdits corpus, and one baseline trained on a comparable set of phrasal insertions not derived from human edits. We evaluate on the same 10K held-out insertion

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

[8] The number of tokens in the WikiAtomicEdits is computed as the the total number of words in the edited sentence $e(s)$ after the insertion. Refer to Supplementary Material for more detailed statistics on the size of the dataset.

[9] To avoid bias, the 200 examples are shuffled and the annotator does not know which group (correct/incorrect, or which model) each example belongs to.

[10] We use the chi-squared test provided by scipy.stats.

	Base	General		Discr.	
	Freq.	LM		Model	
		✓	✗	✓	✗
Extend	25	21	19	25	21
Refine	14	7	18	13	14
RE	6	7	9	4	9
Fluency	5	15	4	8	6

Table 9: Relationship between model accuracy and insertion type, based on a sample of 50 correct (✓) and 50 incorrect (✗) predictions from each model. Base frequency is shown for reference and is based on our analysis from §5.1. The General LM shows a bias in accuracy by insertion type. This bias is not observed for the discriminative model.

edits as in §6.1, and measure performance using both a strict "exact match" as well as a softer similarity metric.

Model. We use an standard sequence-to-sequence model (Sutskever et al., 2014), modifying the input with a special token denoting the insertion point. For example, given the input [*" Angel " is a song recorded by <ins> pop music duo Eurythmics* .], the model would be trained to produce the target phrase [*the British*]. We use a two-layer bidirectional encoder using the same 300-dimensional FastText embeddings as in §6.1, and a sequence decoder with attention (Bahdanau et al., 2015) using a learned wordpiece model (Schuster and Nakajima, 2012) with a vocabulary of 16,000.

Experimental Design. We train one version of this model on the same set of 23M English examples as the discriminative insertion model from §6.1; we refer to the model trained on this data as **Edits**. For comparison, we train an identical model on a set of simulated insertions which we create by sampling sentences from Wikipedia and removing contiguous spans of tokens, which we then treat as the insertion phrases. To ensure that this data is reasonably comparable to the WikiAtomicEdits data, we parse the sampled sentences (Andor et al., 2016) and only remove a span if it represents a full subtree of the dependency parse and is not the subject of the sentence.[11] We generate 23M such "psuedo-edits" for training, the same size as the WikiAtomicEdits training set. We refer to the model trained on this data as **General**.

Results. We look at the top 10 phrases proposed by each model, as decoded by beam search. In addition to reporting standard LM perplexity, we compute two measures of performance, which are intended to provide an intuitive picture of how well each model captures the nature of the information that is introduced by the human editors. Specifically, we compute **Exact Match** as the proportion of sentences for which the model produced the gold phrase (i.e. the phrase inserted by the human editor) somewhere among the top 10 phrases. We also compute **Similarity@1** as the mean cosine similarity of each top-ranked phrase and respective gold phrase over the test set. We use the sum of the Glove embeddings (Pennington et al., 2014) of each word in the phrase as a simple approximation of the phrase vector.

Table 11 shows the results. We see that, compared to the model trained on General Wikipedia, the model trained on WikiAtomicEdits generates edits which are more similar to the human insertions, according to all of our metrics. Table 10 provides a few qualitative examples of how the phrases generated by the Edits model differ from those generated by the General model. Specifically, we see that the Edits model proposes phrases which better capture the discourse function of the human edit: e.g. providing context for/elaboration on a previously-stated fact. We note that this does not mean that training on Edits is inherently "better" than on General text, but rather that the supervision encoded by the WikiAtomicEdits corpus encodes aspects of language that are distinct from those easily learned from existing resources.

7 Related Work

Wikipedia Edits. Wikipedia edit history has been used as a source of supervision for a variety of NLP tasks, including sentence compression and simplification (Yamangil and Nelken, 2008; Yatskar et al., 2010), paraphrasing (Max and Wisniewski, 2010), entailment (Zanzotto and Pennacchiotti, 2010; Cabrio et al., 2012), and writing assistance (Zesch, 2012; Cahill et al., 2013; Grundkiewicz and Junczys-Dowmunt, 2014). User edits from Wikipedia and elsewhere have also been analyzed extensively for insight into the editing process and the types of edits made (Daxenberger and Gurevych, 2012, 2013; Yang et al., 2017)

[11] Not all of the inserted phrases in WikiAtomicEdits are well-formed constituents. However, generating psuedo-edits using this heuristic provided a cleaner, more realistic comparison than using fully-random spans.

She is cited as the first female superstar of Hindi Cinema **and India 's Meryl Streep**	
Edits	General
and is the best actress of the film	in japan
and is the best actress of the Indian cinema	in june
and is the best actress of the film industry	in 2011

He is married to Aida Leanca **and has two children**	
Edits	General
and has a daughter	in january
, and has a daughter	in june
, and has a daughter and a daughter	in january 2012

Table 10: Predicted phrase insertions from model trained on Edits vs. General corpus. The Edits model better captures the discourse function of the human edit, e.g. elaborating on the previously-stated fact, while the General model gives syntactically-appropriate but generic insertions.

	Edits	General
Log Perplexity	8.32	9.23
Exact Match	13.1%	8.0%
Similarity@1	0.54	0.48

Table 11: Comparison of how closely each model's generated phrases match the phrase inserted by the human editor. "Edits" was trained on WikiAtomicEdits and "General" was trained on comparable data not derived from human edits. We consider the top 10 phrases generated by each model.

and to better understand argumentation (Tan and Lee, 2014). Particular attention has been given to spam edits (Adler et al., 2011) and editor quality (Leskovec et al., 2010). Our work differs in that WikiAtomicEdits is much larger than currently available corpora, both by number of languages and by size of individual languages. In addition, our focus on atomic edits should facilitate more controlled studies of semantics and discourse.

Sentence Representation and Generation. We view the WikiAtomicEdits corpus as being especially valuable for ongoing work in sentence representation and generation, which requires models of what "good" sentences look like and how they are constructed. Recent work has attempted to model sentence generation by re-writing existing sentences, either using crowdsourced edit examples (Narayan et al., 2017) or unsupervised heuristics (Guu et al., 2018); in contrast, we provide a large corpus of natural, human-produced edits.

Also related is recent work in sentence representation learning from raw text (Kiros et al., 2015; Peters et al., 2018), bitext (McCann et al., 2017), and other supervised tasks including NLI (Conneau et al., 2017). Especially related is work on learning representations from weakly-labelled discourse relations (Nie et al., 2017; Jernite et al., 2017), as the WikiAtomicEdits corpus captures similar types of discourse signal.

Description of Data Release

Our full corpus is available for download at `http://goo.gl/language/wiki-atomic-edits`. The data contains 26M atomic insertions and 17M atomic deletions covering 8 languages. All sentences (both the original sentence s, and the edited sentence $e(s)$) have been POS-tagged and dependency parsed (Andor et al., 2016) as well as scored using a SOTA LM (Jozefowicz et al., 2016). We also release the 5K 5-way human insertion annotations for English, and 1K 3-way annotations each for Spanish and German, as described in §4.

8 Conclusion

We have introduced the WikiAtomicEdits corpus, derived from Wikipedia's edit history, which contains 43M examples of atomic insertions and deletions in 8 languages. We have shown that the language in this corpus is meaningfully different from the language we observe in general, and that models trained on this corpus encode different aspects of semantics and discourse than models trained on raw text. These results suggest that the corpus will be valuable to ongoing research in semantics, discourse, and representation learning.

References

B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proc. of CICLing*.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proc. of ACL*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Elena Cabrio, Bernardo Magnini, and Angelina Ivanova. 2012. Extracting context-rich entailment rules from wikipedia revision history. In *Proc. of the 3rd Workshop on the People's Web Meets NLP*.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust systems for preposition error correction using wikipedia revisions. In *Proc. of NAACL*.

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*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proce. of COLING*.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in wikipedia revisions. In *Proc. of EMNLP*.

Dan Gillick. 2009. Sentence boundary detection and the problem with the U.S. In *Proc. of NAACL*.

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

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing – Lecture Notes in Computer Science*, volume 8686, pages 478–490. Springer.

K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics (TACL)*.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

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

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proc. of NIPS*.

George Lakoff. 1970. Linguistics and natural logic. *Synthese*, 22(1/2):151–271.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Governance in social media: A case study of the Wikipedia promotion process. In *Proc. of ICWSM*.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Aurélien Max and Guillaume Wisniewski. 2010. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *Proc. of LREC*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proc. of NIPS*.

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

Eugene W Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proc. of EMNLP*.

Allen Nie, Erin D Bennett, and Noah D Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *arXiv preprint arXiv:1710.04334*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. of NIPS*.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proc. of ACL*.

Johan Van Benthem. 1986. *Natural Logic*. Springer Netherlands.

Derry Tanti Wijaya, Ndapandula Nakashole, and Tom Mitchell. 2015. "A spousal relation begins with a deletion of engage and ends with an addition of divorce": Learning state changing verbs from wikipedia revision history. In *Proc. of EMNLP*.

Elif Yamangil and Rani Nelken. 2008. Mining wikipedia revision histories for improving sentence compression. In *Proc. of ACL*.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proc. of EMNLP*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proc. of NAACL*.

Fabio Massimo Zanzotto and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from wikipedia using co-training. In *Proc. of the 2nd Workshop on The People's Web Meets NLP*.

Torsten Zesch. 2012. Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In *Proc. of EACL*.