NYTWIT: A Dataset of Novel Words in the New York Times

Yuval Pinter
Georgia Institute of Technology
Atlanta, GA, USA
uvp@gatech.edu

Cassandra L. Jacobs
University of Wisconsin
Madison, WA, USA
cjacobs2@wisc.edu

Max Bittker
School for Poetic Computation
New York, NY, USA
maxbittker@gmail.com

Abstract

We present the New York Times Word Innovation Types dataset, or **NYTWIT**, a collection of over 2,500 novel English words published in the New York Times between November 2017 and March 2019, manually annotated for their class of novelty (such as lexical derivation, dialectal variation, blending, or compounding). We present baseline results for both uncontextual and contextual prediction of novelty class, showing that there is room for improvement even for state-of-the-art NLP systems. We hope this resource will prove useful for linguists and NLP practitioners by providing a real-world environment of novel word appearance.

1 Introduction

Novel words, or Out-Of-Vocabulary words (OOVs), are a pervasive problem in modern natural language processing (Brill, 1995; Young et al., 2018). A common scenario in which this problem appears is that of a pre-trained model containing a word representation component such as an embedding table, encountering a previously-unseen word in a downstream task such as question answering or natural language inference. Multiple lines of work attempt to alleviate the downstream effect of OOV words (Müller and Schütze, 2011; Pinter et al., 2017), but most tend to focus on individual categories of OOVs: typographical errors (Sakaguchi et al., 2017), domain-specific terminology (Du et al., 2016), stylistic variability (Eisenstein, 2013; van der Goot, 2019), morphological productivity (Bhatia et al., 2016), or novel named entities (Hoffart et al., 2014). In reality, unseen texts contain all these classes of novelty, and more. OOVs are a typically presented as a significant challenge for generalization or understanding in noisy user-generated text (e.g. Twitter) and/or domain-specific content. Nevertheless, even large corpora that are narrow in domain (edited news stories) contain linguistic innovations, including but not limited to novel morphological processes, typographical errors, and loan words.

In this paper, we present a dataset of novel words in English relative to the corpus of articles published by the New York Times (NYT), as collected automatically in real time by a Twitter bot. We name it the New York Times Word Innovation Types corpus, or **NYTWIT** for short. We annotated each word for one of eighteen linguistically-informed **categories** of novelty within the context of the NYT corpus, as well as for its date of publication and a retrieval document identifier to enable context extraction.[1] To our knowledge, this is the first resource to include novel words along with their contextual information in addition to linguistically-informed annotation, a method that enables expansion beyond dictionary-based methods (Cook and Stevenson, 2010; Dhuliawala et al., 2016; Ahmad, 2000) and decontextualized neologisms (Kulkarni and Wang, 2018). In contrast with resources which provide examples and attestations to lexical forms, NYTWIT was constructed in a corpus-comprehensive manner where novelty guides curation and not vice versa.

In addition, we provide results for the task of classifying words into their categories based on word form and contextual information, a task which can both provide data for linguistic analysis of lexical

[1] Context article excerpts are not freely available without copyright licensing from the New York Times, who have ignored all contact attempts to date.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6509–6515
Barcelona, Spain (Online), December 8-13, 2020

enrichment and serve as a processing step for NLP systems which may work better if different modules are applied to different classes of novel words. We show that both character-level models and large pre-trained sentence encoders struggle on this task, illustrating the challenges of modeling language innovation.

We release the data at `https://github.com/yuvalpinter/nytwit` under the GNU General Public License v3.0. The project is ongoing, and this document pertains to version 1.1.

2 The New York Times Word Innovation Types Dataset

Our dataset is built upon two bots developed by the third author. The first stage of data collection relies on tweets from the NYT_First_Said bot[2], which operates by scraping new articles as they post on the NYT site and tweeting out novel words following a filtering process which we will describe at a high level.[3] After tokenizing on white space and punctuation, the precision-oriented script rejects capitalized words in order to avoid proper nouns (at the cost of missing sentence-initial true OOVs). `langid` (Lui and Baldwin, 2012) is used to reject non-English sentences, while still allowing loanwords in English sentences. Words are queried against the historical NYT search API to detect unpublished words.[4] For the time range of our collected corpus, November 7, 2017 to March 28, 2019, a bandwidth limit of five words per 30 minutes was imposed, but we confirmed that this did not have a substantial effect on OOV coverage, leaving our artifacts distributionally representative for the news domain.

An associated context bot replies to the tweets with links to the original articles.[5] We used the URLs from this bot's posts as the main reference for the words' contexts. For 17 words, the article URL was retrieved manually by searching for the target article directly.[6] As the articles are subject to edits long after publication, there is an increasing but small portion of articles which no longer contain the context, although at time of publication these mostly include the removal of typographical errors from the stories and which are ultimately filtered by our annotation process (see below).

2.1 Annotation

The extracted data was independently annotated and filtered by the first two authors. Initially, all 2,587 words were assigned one of 20+ tags inspired by the word formation literature (Kiparsky, 1982; Klymenko, 2019). Certain categories were filter categories intended to capture and exclude false positives from the final dataset: DUPLICATE for inflections of words already appearing in the dataset in a morphologically simpler form, e.g. *batchcode* and *batchcodes*; FOREIGN and PRP for foreign words and proper names (mostly all-lowercase Twitter usernames) which were not caught by the automatic filtering; SPACES and TYPO for unintended cases of space deletion and typographical errors which were not caught by NYT editors.[7] The filtered items are provided in the dataset under the label FILTERED.

Agreement between the annotators at the preliminary phase was 68% over all labels, and 0.65 Cohen's Kappa. Following category filtering, amounting to 40% of the original dataset, agreement over the remaining 1,550 words was calculated to be 65% at 0.61 Kappa. At the coarse-grained level, agreement on the four themes (lexical / morphological / syntactic / sociopragmatic) was 89% at 0.75 Kappa.[8]

The annotators then examined each other's annotations and agreed on some consolidation of rarely-occurring original labels, as well as introduction of new labels deemed useful post-hoc.

[2] `https://twitter.com/NYT_First_Said`

[3] The code for the bot is available at `https://github.com/MaxBittker/NYT-first-said`.

[4] We note that the search index relies on imperfect, although extensive, digitization artifacts. At the time of writing, in a sample of 450 terms from our dataset, [Cassie: "4"] were entries in the Oxford English Dictionary, nearly all of which belong to the domain or foreign categories.

[5] `https://twitter.com/NYT_Said_Where`

[6] One term lacks a context because neither the NYT search engine nor the API support the letter é.

[7] The overwhelming share of these words have indeed since been deleted from the NYT website.

[8] A reviewer noted that these are low agreement rates, and compared the task to part-of-speech annotation. We dispute the comparison, both on grounds of the novelty of the forms involved and of the mechanical syntactic nature of the majority of POS tagging decisions.

2.2 Novel Word Taxonomy

We describe the eighteen categories in the finalized dataset, organized by a thematic grouping not explicitly annotated. Counts for each category are provided [in brackets].

Lexical OOVs. We deem certain categories to arise from the fact that the NYT, while being interested in many aspects of life, has not had the chance to delve into each and every one at depth over its 168 years of existence. These are the DOMAIN label for technical terms from uncommon domains (e.g. *glossopoeia*) [285]; the INNOVATION label for terms coined with no discernable prevailing linguistic process (e.g. *swanicles*, a term from a work of fiction) [11]; and the ONOMATOPEIA label for sound-based sequences (e.g. *ktktk*) [23], which includes cases of verbatim vocalization such as *trololo*.

Morphological OOVs. In this group we include categories of words composed of meaning-carrying units present in existing English words which have appeared in the NYT before, manifested in a new form. In increasing order of syntactic and semantic novelty, they are: INFL, unseen inflections of existing wordforms: same part-of-speech, different syntactic attributes (e.g. *pennyloafers*) [53];[9] DERIV, unseen derivations of existing words into new parts-of-speech which carry no semantic distancing beyond that implicit in the new part-of-speech itself (e.g. *foamability*) [215]; AFFIX, affixation of very distinct base words which are typically derivational in nature but include a semantic charge (e.g. *extraphotographic*, *pizzaless*) [483]; AFFIX_LIBFIX, affixation of distinct base words with particles extracted from another word in a process known as *libfixation* (Zwicky, 2010) or *splintering* (Berman, 1961), where the liberated affix still elicits the originating word but can be freely attached to a growing selection of words (e.g. *dripware*) [18]; COMPOUND_COMP, a concatenation of two complete words each contributing essential semantics to the final form in a way we deem (subjectively, with help of context) to be compositional (e.g. *smellwalks*, strolls focusing on olfactory input) [121];[10] COMPOUND_NEW, a concatenation of base words resulting in a new semantic concept deemed remote from the bases (e.g. *nothingbuffet*, a play on *nothingburger*) [49]; and BLEND, a fusion of two or more base forms together where original characters are lost or shared, or new ones are added (e.g. *chipster*, a chicano hipster) [142].[11]

Syntactic OOVs. This group consists solely of the SYNTH category of tokens which synthesize multiple syntactic words into one form, a rare formation process in English limited typically to auxiliary contractions (e.g. *this'll*) [6].

Sociopragmatic OOVs. Words in this group exhibit an orthographic diversion from standard English usually intended as a statement of register or status, or as a faithful representation of a certain linguistic style or sentiment. ARCHAIC, a register of older variants of English or an ironic semblance of such (e.g. *shooketh*, a mock-archaic form of *shake* using Middle English morphology) [14]; DIALECT, a geographically- or demographically-specific form of a word typically spelled differently in the NYT (e.g. *skwarsh*, an r-full *squash*) [46]; INFIX, a morphological tool reserved in English for expletive emphasis (McCawley, 1978) (e.g. *unfreakingbelievable*) [2]; PHONAESTHEME, a phonological duplication phenomenon used in contemporary English nearly only as derisive echo reduplication borrowed from Yiddish (Wales and Ramsaran, 1990) (e.g. *schmarket*) [6]; LENGTHENING, a written manifestation of the expressive elongation of phonetic segments (e.g. *greaaaaat*) [53]; VARIANT, spelling alternations or intentional typos which are not intended to be read differently from the standard form of the word, used for branding and jest (e.g. *kyllyng*) [18]; and SPACES_SIC, the removal of whitespace to simulate breathlessness (e.g. *lineafterlineafterline*) [5].

2.2.1 Difficult Distinctions

Naturally, some annotation cases are not clear-cut, as evidenced by the imperfect inter-annotator agreement. We found the most challenging cases to be among the morphological categories, where an affix is either semantically null (DERIV / INFL) or not (AFFIX) (14% and 15% of disagreements, respectively);

[9] We include the negating prefixes `in-` and `un-`, which despite change a word's meaning, but retain its part-of-speech.

[10] One compound in our dataset, *dramatotherapy*, adds characters for cadence; another, *laysoccerperson*, is nonlinear.

[11] A single blend, *pregret*, has just one base fused with a prefix.

where a sense of the nearest in-vocabulary word can signal the difference between INFL and DERIV (3.4%); where an AFFIX_LIBFIX has been "liberated" enough from the underlying word such that it is now simply an AFFIX (does *cyber-* still envoke the full word *cybernetics*? Does *crypto-* envoke *cryptography*?); if it has not been liberated yet, it should be a BLEND or a COMPOUND. In addition, the pre-processing phase required a demarcation between DOMAIN and FOREIGN which was not easy to make given the heavy foreign-word influence in certain knowledge domains such as cuisine (e.g. *dinkelbrot*). Words adapted into English morphology would usually lead to a DOMAIN label (DOMAIN vs. COMPOUND: 4%). In many cases, we found the contexts in which the words were introduced to give sufficient disambiguation (so, e.g., *cybercoach* is an affix, but *cyberinvasion* is a compound).

We invite readers to email errata to either of the first two authors, or submit a pull request on Github.

3 OOV Classification Task

The task of classifying OOVs, i.e. assigning a novel word with a label from the taxonomy we defined above, can be beneficial from both an analytical linguistic standpoint, and from an NLP standpoint concerned with model performance on downstream language understanding tasks. To get a sense of the predictability of the various OOV classes in the dataset, we present several baselines for this straightforward task. The uniqueness of our dataset allows us to apply both type-level and context-dependent systems, the latter operating in the real-world scenario of encountering a word for the first time in the actual context of its introduction to the corpus.

First, our **Majority class** baseline assumes all OOVs are the result of *affixation*.

For all following models we trained a ridge classifier with default regularization parameters in `scikit-learn`. Scores for all supervised models are reported via 10-fold cross-validation using the same folds for all systems. Due to the class imbalance, we chose to implement training in such a way that upsampled rare classes with replacement at each iteration to equal frequency as the most common class. We report accuracy (ACC) and macro F1 scores.

Contextless features. We compare and contrast several input features to our classifier that only have access to the form of the OOV, without consideration of the context:

- **Character n-grams.** We extract bag-of-character features ranging from one to three characters for each OOV. The feature vocabulary is estimated on the training set and applied to the test set.
- **FastText.** We infer fasttext vectors (Bojanowski et al., 2017), applying its 3–6 character-ngram representations, from the subword model trained on English Wikipedia.[12]
- **ELMo embeddings.** We use the word-level embeddings from ELMo (Peters et al., 2018), obtained via a pre-trained character-level convolutional net for each OOV presented in isolation, with no surrounding sentence context.
- **BERT no-context.** We apply BERT-Base (Devlin et al., 2019) to the OOVs in a null "`[CLS] ___ [SEP] .`" context. The averaged top-layer vectors from all the OOV's word pieces are passed to the classifier.[13]

Context-aware features.

- **Character RNN.** We train a 2-layer forward- (backward-) character-level GRU language model on 100,000 Wikipedia documents and run it through the beginning (end) of the sentence until the OOV, then use the concatenated final hidden states from each direction as features.
- **ELMo.** We obtain contextualized embeddings for all words in our sentences and select the top layer representation associated with each OOV.
- **BERT.** We apply BERT-Base to the entire sentence in which the OOV appears, and use the averaged top-layer embeddings at the indices of each OOV.

[12] `wiki.en.bin` file obtained May 25, 2020.

[13] Using just the embedding of the final word piece produced similar results.

Contextless	Acc	F1	Contextual	Acc	F1
Majority class	.312	.026			
Character n-grams	**.484**	**.323**			
FastText	.433	.241	Character RNN	.128	.054
ELMo embeddings	.365	.203	ELMo	.324	.135
BERT no-context	.442	.288	BERT	.469	.269

Table 1: Baseline results for OOV classification ($N = 1550$, $|C| = 18$).

3.1 Results

The results, presented in Table 1, show that pretrained contextual models not only trail behind a contextless, un-pretrained character n-gram baseline, they even fail to improve over their own uncontextualized variants. An analysis of class-specific F1 scores across the different models exposed two general patterns in classifier performance: in all models, performance on the AFFIX class was in the top four, and the same for LENGTHENING except for Character RNN. We also observed that models that encode contextual, sentence-level properties are typically better at encoding genre phenomena (e.g. DOMAIN was a top-four category for BERT, Character RNN, fastText, and ELMo). However, for some classes of models, there was a clear benefit to memorizing word forms. All count-based feature representations (e.g. bag-of-character ngrams, bag-of-wordpieces) led to better performance on orthographic properties, namely PHONAESTHEME, SYNTH, and ONOMATOPOIEA. These results demonstrate the power that simple surface-form signals from character sequences still possess in meaningful NLP tasks. In future work, we will attempt to supplement the contextual models with auxiliary mechanisms and perform fine-tuning.

4 Conclusion

We presented a novel dataset of OOVs along with their contexts and linguistic novelty class annotations. We showed that contextual information in the form of other parts of the sentence provides some signal, but simple models relying on character n-gram information alone achieve high performance.

The availability of broader document contexts in which these neologisms occur enables many linguistic and technical applications. From the perspective of the study of language growth and formation, the dataset may be used to assess the morphological productivity of different affixes and roots, or the prevalence of the different word formation processes in a realistic setting; or perform in-depth analysis on any of the specific types of innovations we identified. In addition, the in-vivo nature of the dataset provides a reference for neologisms which may or may not be later adopted into everyday use, allowing diachronic studies anchored in the time of word introduction. Analysis of the phonological, morphological, and discourse-level properties of these words may provide insight into lexical adoption dynamics.

For NLP researchers, an important component of text applications is proper normalization and segmentation of word forms. Our experiment shows that popular word form encoders, such as ELMo or BERT's WordPiece, still have a long way to go in terms of recognizing the origins of a novel form. Errors at this stage might lead to inability to handle morphologically complex OOVs in downstream semantic applications (Pinter et al., 2020), although further study of such effects and of the utility of OOV classification in alleviating them is still necessary. Properly leveraging context for morphological decomposition of complex forms also remains an open problem.

The resource is an ongoing project; the repository includes plans for the next versions, including increasing the dataset size by including newer words from the bot, and annotating additional information such as part-of-speech tags.

Acknowledgments

We thank Jacob Eisenstein, Kyle Gorman, Arya McCarthy, Sandeep Soni, and the anonymous reviewers for their valuable notes. Yuval Pinter is a Bloomberg Data Science PhD Fellow. Cassandra Jacobs is supported on NSF BCS Grant 1849236 awarded to Maryellen MacDonald.

References

Khurshid Ahmad. 2000. Neologisms, nonces and word formation. In *Proceedings of the Ninth EURALEX International Congress*, pages 711–730.

JM Berman. 1961. Contribution on blending. *Zeitschrift für Anglistik und Amerikanistik*, 9:278–281.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 490–500, Austin, Texas, November. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying the source words of lexical blends in English. *Computational Linguistics*, 36(1):129–149.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Jinhua Du, Andy Way, and Andrzej Zydron. 2016. Using BabelNet to improve OOV coverage in SMT. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 9–15, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia, June. Association for Computational Linguistics.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd international conference on World wide web*, pages 385–396.

Paul Kiparsky. 1982. Word-formation and the lexicon. In *Proceedings of the Mid-America Linguistics Conference*, pages 3–29. University of Kansas.

Olga Klymenko. 2019. Twitterverse: The birth of new words. *Proceedings of the Linguistic Society of America*, 4(1):11–1.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. 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 1424–1434, New Orleans, Louisiana, June. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

James D McCawley. 1978. Where you can shove infixes. *Syllables and segments*, pages 213–221.

Thomas Müller and Hinrich Schütze. 2011. Improved modeling of out-of-vocabulary words using morphological classes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 524–528, Portland, Oregon, USA, June. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of EMNLP*.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Rob van der Goot. 2019. An in-depth analysis of the effect of lexical normalization on the dependency parsing of social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 115–120, Hong Kong, China, November. Association for Computational Linguistics.

Katie Wales and S Ramsaran. 1990. Phonotactics and phonaesthesia: the power of folk lexicology. *Studies in pronunciation of English. A commemorative volume in honour of AC Gimson*, pages 339–351.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Arnold Zwicky. 2010. Libfixes. *Arnold Zwicky's Blog*.