Will it Unblend?

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

Cassandra L. Jacobs
Department of Psychology
University of Wisconsin
Madison, WI, USA
cjacobs2@wisc.edu

Jacob Eisenstein
Google Research
Seattle, WA, USA
jeisenstein@google.com

For the token-based architectures that dominate contemporary natural language processing, a particularly difficult form of linguistic generalization arises from unseen phenomena at the word level. Such novel sequences of characters, morphemes, or phonemes are known as out-of-vocabulary (**OOV**) terms (Brants, 2000; Plank, 2016). Pretrained transformers like BERT (Devlin et al., 2019) handle OOV terms by **subtokenization**: segmenting all whitespace-delimited tokens into smaller units, from which any OOV term can be constructed (Sennrich et al., 2016). While this approach is well suited for phenomena like concatenative English morphology, many linguistic processes generate OOVs that cannot be cleanly decomposed into meaningful segments.

Our work addresses a challenging source of OOV terms: novel **blends** (Algeo, 1977), also known as portmanteaux. Blends are constructed from the combination of multiple **bases** into a new form, in which some characters are shared across both bases: for example, *shop* + *optics* = *shoptics*. In this way, blends differ from other lexical formations such as compounds (e.g., *water* + *melon* = *watermelon*), which are formed by simple concatenation.

Dataset and annotation. We collected a dataset of novel blends from the New York Times (NYT), starting from the output of a Twitter bot extracting all novel words with their originating contexts, a process described in Pinter et al. (2020). For each blend, we annotated the bases and the semantic relation between them, following the taxonomy defined by Tratz and Hovy (2010). We also define a character-level schema we call PAXOBS after its tagset, where each character in the blend is identified as being part of a single base (given sequential letters of the alphabet, so typically A or B), part of both (X), an extra-base prefix (P) or suffix (S), or a new character added to the blend for prosodic effect (O). Table 1 presents four examples of blends from our dataset, two of which are **linear**: all A characters precede all B characters; no O characters exist; and no interleaving X characters.

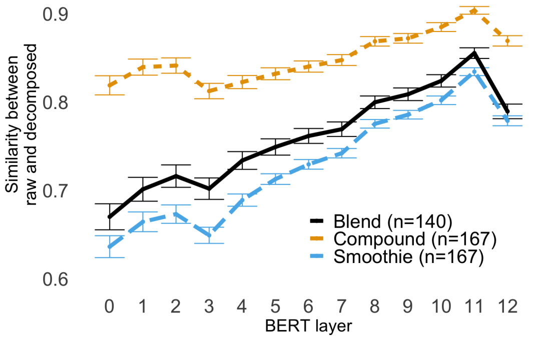

Figure 1: Similarity of BERT representations between base components of complex words, and naturally and artificially blended forms ("smoothies"). All representations computed in the original context in which the words appear. Error bars are standard error for the class.

Contextualized embeddings of blends. We show the inherent challenges presented by blends for pretrained contextualized word embeddings (e.g., BERT), which are a foundational component of natural language processing. To gauge how well BERT represents blends, we conduct a comparison with its treatment of a minimally-different control class of novel words, namely **lexical compounds**, which are forms where at least two bases are concatenated in full (e.g. *quizmaker*), without the character loss incurred in blends. We run BERT on the surrounding contexts of blends and compounds as-is, and compare the resulting representations with those when the novel words are separated into

Blend	PAXOBS	Bases	Semantic relation	Definition
hatriotism	AXXBBBBSSS	hate patriotism	ATTRIBUTE	Hate disguised as patriotism.
shoptics	AAXXBBBS	shop optics	LOC-PART-WHOLE	The social image projected when shopping.
innoventor	XXAAXBBXXX	innovator inventor	CAUSAL	A person who innovates by inventing.
thrupple	AAABOBBB	three couple	CONTAINMENT	A group of three people acting as a couple.

Table 1: A sample of the blends from the fully annotated dataset. Linear blends are underlined.

their bases (e.g. `[left-context hatriotism right-context]` vs. `[left-context hate patriotism right-context]`).[1] Figure 1 shows that compounds are represented much more similarly to their bases than blends are. This result can be due to either a functional preference for creating blends in certain semantic conditions, or due to the form-level pathology of blends, i.e. the missing and joined characters. We annotated all blends and compounds for their semantic relations and found that despite matching semantic roles, blends were still highly dissimilar from their decomposed forms in context. We thus subject the compounds to a process of artificial blending using an existing algorithm for string merging (Kulkarni and Wang, 2018) and repeat the experiment on these "smoothies". As shown in Figure 1, this process eliminates the representation gap, leading us to conclude that the difficulty in representing blends is due primarily to the complex relationship between their surface forms and meanings.

Unblending. As part of blend understanding, blends may be understood through segmentation and component identification, which we term *recovery*. We cast the problem as a two-step pipeline, beginning with detection of morphological boundaries within blends, followed by a selection of the correct bases from a list of candidates constructed given the segmentation and a vocabulary (similar to Cook and Stevenson, 2010). Even under favorable conditions, we find that systems proposed previously for similar tasks struggle on blends, showing limitations of form-based and distributional similarity approaches: BERT's WordPiece segmentation (which is based on byte-pair encoding, or BPE (Sennrich et al., 2016)) reaches an F1 score of .562 for segmentation, compared with .427 by a baseline which treats each character as its own segment. Neither a BPE model retrained on news data (to approximate NYT's domain) nor a character-level sequence tagger improve this result.

	Mean Reciprocal Rank			
Model	A	B	ω	P@1
Lower bound	.115	.257	.036	.014
Character RNN	.162	.368	.060	.021
Edit distance	.176*	.432*	.066	.014
fastText	.357*	.610*	.167	.127
GloVe	.449*	.734*	.188	.127
BERT RANKER	.392	.711	**.288**	**.264**
−CONTEXT	.379	.675	.147	.127

Table 2: Results for base recovery. MRR columns refer to the first base (A), the second base (B), or the pair composed of both bases (ω). *Results dependent on knowledge of the correct base on the other side.

We next investigate whether it is possible to recover the original bases given correct segmentation: when shown a substring of a blend corresponding to the portion originating in one of the bases, the system scores all possible candidates (bases beginning or ending with the substring) and we record the rank of the correct base in the scored list. We compare a novel unsupervised base recovery method we propose, BERT RANKER, against various baselines, in Table 2. On the precision at 1 metric, which measures the proportion of true base pairs ranked above all candidates, BERT RANKER performs twice as well as systems based on static embedding similarity (FASTTEXT and GLOVE, but there is still substantial room for improvement.

Conclusion. Our experiments show that even sophisticated methods struggle to parse and understand blends. We find that the use of context can improve models for segmentation of blends and recovery of the base components. Our results highlight the need for future work on our novel dataset and associated tasks.[2]

[1] We note that blends, or compounds, are not generally substitutable with the phrase composed of the originating bases, syntactically speaking. Our evaluation stands in for evaluation on a downstream task, which we see as a promising avenue for future work given our results.

[2] We release our code and data at http://github.com/yuvalpinter/unblend.

References

John Algeo. 1977. Blends, a structural and systemic view. *American speech*, 52(1/2):47–64.

Thorsten Brants. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of Applied Natural Language Processing*, pages 224–231.

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 NAACL*.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. In *Proceedings of NAACL*.

Yuval Pinter, Cassandra L Jacobs, and Max Bittker. 2020. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, Online. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of ACL*.