2021

Will it Unblend?

Yuval Pinter  
*Georgia Institute of Technology*, me@yuvalpinter.com

Cassandra L. Jacobs  
*University of Wisconsin*, cjacobs2@wisc.edu

Jacob Eisenstein  
*Google Research*, jeisenstein@google.com

Follow this and additional works at: https://scholarworks.umass.edu/scil

Part of the Computational Linguistics Commons

Recommended Citation

Pinter, Yuval; Jacobs, Cassandra L.; and Eisenstein, Jacob (2021) "Will it Unblend?," *Proceedings of the Society for Computation in Linguistics*: Vol. 4 , Article 62.  
DOI: https://doi.org/10.7275/dtcb-0n63  
Available at: https://scholarworks.umass.edu/scil/vol4/iss1/62

This Paper is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
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
their bases (e.g. left-context hatriotism right-context) vs. (left-context hate patriotism right-context)).

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 unfavorable 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.

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.

---

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](http://github.com/yuvalpinter/unblend).

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

<table>
<thead>
<tr>
<th>Blend</th>
<th>PAXOBS</th>
<th>Bases</th>
<th>Semantic relation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>hatriotism</td>
<td>AXXBBBSSS</td>
<td>hate patriotism</td>
<td>ATTRIBUTE</td>
<td>Hate disguised as patriotism.</td>
</tr>
<tr>
<td>shopitics</td>
<td>AXXBBSB</td>
<td>shop optics</td>
<td>LOC-PART-WHOLE</td>
<td>The social image projected when shopping.</td>
</tr>
<tr>
<td>innovator</td>
<td>XXAXBBXXXXX</td>
<td>innovator inventor</td>
<td>CAUSAL</td>
<td>A person who innovates by inventing.</td>
</tr>
<tr>
<td>thruple</td>
<td>AAABBBB</td>
<td>three couple</td>
<td>CONTAINMENT</td>
<td>A group of three people acting as a couple.</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Reciprocal Rank</th>
<th>P@1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Lower bound</td>
<td>.115</td>
<td>.257</td>
</tr>
<tr>
<td>Character RNN</td>
<td>.162</td>
<td>.368</td>
</tr>
<tr>
<td>Edit distance</td>
<td>.176*</td>
<td>.432*</td>
</tr>
<tr>
<td>fastText</td>
<td>.357*</td>
<td>.610*</td>
</tr>
<tr>
<td>GloVe</td>
<td>.449*</td>
<td>.734*</td>
</tr>
<tr>
<td>BERT RANKER</td>
<td>.392</td>
<td>.711</td>
</tr>
<tr>
<td>–CONTEXT</td>
<td>.379</td>
<td>.675</td>
</tr>
</tbody>
</table>
References


