Lexical Semantic Recognition

Nelson F. Liu
Stanford University
nfliu@cs.stanford.edu

Daniel Hershcovich
University of Copenhagen
dh@di.ku.dk

Michael Kranzlein
Georgetown University
{mmk119, nathan.schneider}@georgetown.edu

Abstract

In lexical semantics, full-sentence segmentation and segment labeling of various phenomena are generally treated separately, despite their interdependence. We hypothesize that a unified lexical semantic recognition task is an effective way to encapsulate previously disparate styles of annotation, including multiword expression identification/classification and supersense tagging. Using the STREUSLE corpus, we train a neural CRF sequence tagger and evaluate its performance along various axes of annotation. As the label set generalizes that of previous tasks (PARSEME, DiMSUM), we additionally evaluate how well the model generalizes to those test sets, finding that it approaches or surpasses existing models despite training only on STREUSLE. Our work also establishes baseline models and evaluation metrics for integrated and accurate modeling of lexical semantics, facilitating future work in this area.

1 Introduction

Many NLP tasks traditionally approached as tagging focus on lexical semantic behavior—they aim to identify and categorize lexical semantic units in running text using a general set of labels. Two examples are supersense tagging of nouns and verbs as formulated by Ciaramita and Altun (2006), and verbal multiword expression (MWE) identification and classification in the multilingual PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2018, 2020). By analogy with named entity recognition, we can use the term lexical semantic recognition (LSR) for such chunking-and-labeling tasks that apply to lexical meaning generally, not just entities. This disambiguation can serve as a foundational layer of analysis for downstream applications in natural language processing, and provides an initial level of organization for compiling lexical resources, such as semantic nets and thesauri.

In this paper, we tackle a more inclusive LSR task of lexical semantic segmentation and disambiguation. The STREUSLE corpus (see §2) contains comprehensive annotations of MWEs (along with their holistic syntactic status) and noun, verb, and preposition/possessive supersenses. We train a neural CRF tagger (Lafferty et al., 2001) using BERT embeddings (Devlin et al., 2019) and find that it obtains strong results as a first baseline for this task in its full form.

In addition, we ask: Does a tagger trained on STREUSLE generalize to evaluations like the PARSEME shared task on verbal MWEs (Ramisch et al., 2018) and the DiMSUM shared task on MWEs and noun/verb supersenses (Schneider et al., 2016)? Results show our LSR model based on STREUSLE is general enough to capture different types of analysis consistently, and suggest an integrated full-sentence tagging framework is valuable for explicit modeling of lexical semantics in NLP.1

2 LSR Tagging Frameworks

Our tagger is based on STREUSLE (Supersense-Tagged Repository of English with a Unified Semantics for Lexical Expressions; Schneider and Smith, 2015; Schneider et al., 2018),2 a corpus of web reviews annotated comprehensively for lexical semantic units and supersense labels. Specifically, there are three annotation layers: multiword expressions, lexical categories, and supersenses. The supersenses apply to noun, verb, and prepositional/possessive units. Figure 1 shows an example.

Many of the component annotations have been applied to other languages: verbal multiword expressions (Savary et al., 2017; Ramisch et al., 2018), noun and verb supersenses (e.g., Picca et al.,

1Code, pretrained models, and model and scorer output (all train/dev/test splits) can be found at https://nelsonliu.me/papers/lexical-semantic-recognition
2https://github.com/nert-nlp/streusle
We took our vehicle in for a repair to the air conditioning.

STREUSLE Annotation Layers

STREUSLE comprises the entire 55K-word Reviews section of the English Web Treebank (Bies et al., 2012), for which there are gold Universal Dependencies (UD; Nivre et al., 2020) graphs, and adopts the same train/dev/test split.

The lexical-level annotations do not make use of the UD parse directly, but there are constraints on compatibility between lexical categories and UPOS tags (see §3).

**Multiword expressions** (MWEs; Baldwin and Kim, 2010) are expressed as groupings of two or more tokens into idiomatic or collocational units. As detailed by Schneider et al. (2014a,b), these units may be contiguous or gappy (discontinuous). Each unit is marked with a binary *strength* value: idiomatic/noncompositional expressions are strong; collocations that are nevertheless semantically compositional, like “highly recommended”, are weak.

We use the term **lexical unit** for any expression that is either a strong MWE grouping of multiple tokens, or a token that does not belong to a strong MWE. Every token in the sentence thus belongs to exactly one lexical unit. The other layers of semantic annotation augment lexical units, and weak MWEs are groupings of (entire) lexical units.

**Lexical categories** (lexcats) describe the syntax of lexical units. They are similar to UPOS tags available in the UD annotations of the corpus, but are necessary in order to (a) express refinements relevant to the criteria for the application of supersenses, and (b) account for the overall syntactic behavior of strong MWEs, which may not be obvious from their internal syntactic structure.

**Supersenses** semantically classify lexical units and provide a measure of disambiguation in context. There are 3 sets of supersense labels: nominal, verbal, and prepositional/possessional. The lexcat determines which of these sets (if any) should apply. The MWE, lexcat, and supersense information over lexical units is serialized as per-token tags in a BIO-based encoding (details in §2.1.1).

2.1 Tag Serialization

STREUSLE specifies **token-level tags** to allow modeling lexical semantic recognition as sequence tagging. The Bb110o_~ tagging scheme (Schneider et al., 2014a) consists of 8 positional flags indicating MWE status: 0 applies to single-word expressions, B to the start of a new MWE, I_ to the continuation of a strong MWE, and I~ to the continuation of a weak MWE (if not continuing a strong MWE within the weak MWE). The lowercase counterparts o, b, i, i~ are the same except they are used within the gap of a discontinuous MWE. For MWE identification, local constraints on tag bigrams—e.g., that the bigrams (b, B) and (B, 0) are invalid, and that the sentence must end with I_, I~, or 0—ensure a valid overall segmentation into units (Schneider and Smith, 2015).

The lexcat and (where applicable) supersense information is incorporated in the first tag of each lexical unit. Thus B- . N-n . ARTIFACT indicates the

---

3Some preposition units are labeled with two supersenses drawn from the same label set; the scene role label represents the semantic role of the prepositional phrase marked by the preposition, and function label represents the lexical contribution of the preposition in itself (Schneider et al., 2018). The scene role and the function are identical by default.

4Though in named entity recognition it is typical to include the class label on every token in the multiword unit, STREUSLE does not do this because it would create a non-local constraint across gaps (that the tags at either end have matching lexcat and supersense information). A tagger would either need to use a more expensive decoding algorithm or would need to greatly enhance the state space so within-gap tags capture information about the gappy expression.

In STREUSLE there is actually a slight limitation due to the verbal lexcats, which distinguish between single-word and strong multiword expressions (see Appendix A): if a B- . s or I~ . s tag is followed by a gap, there is no local indication of whether the expression will be strong or weak (strength is indicated only after the gap). If the expression being started is strong, then one of the verbal MWE subtypes (V . V ID, etc.)
The Universal Semantic Tagset takes a similar approach (Bjerva et al., 2016; Abzianidze and Bos, 2017; Abdou et al., 2018), and defines a cross-linguistic inventory of semantic classes for content and function words, which is designed as a substrate for compositional semantics, and does not have a trivial mapping to STREUSLE categories. However, two shared task datasets consist of subsets of the categories used for STREUSLE annotations, on text from different sources.

**PARSEME Verbal MWEs.** The first such dataset is the English test set for the PARSEME 1.1 Shared Task (Ramisch et al., 2018), which covers several genres (including literature and several web genres) and is annotated only for verbal multiword expressions. The STREUSLE lexcats for verbal MWEs are identical to those of PARSEME; thus, a tagger that predicts full STREUSLE-style annotations can be evaluated for verbal MWE identification and subtyping by simply discarding the supersenses and the non-verbal MWEs and lexcats from the output.

**DiMSUM.** The second shared task dataset is DiMSUM (Schneider et al., 2016), which was annotated in three genres—TrustPilot web reviews, TED talk transcripts, and tweets—echoing the annotation style of STREUSLE when it contained only MWEs and noun and verb supersenses. DiMSUM does not contain prepositional/possessive supersenses or lexcats. It also lacks weak MWEs.

### 3 Modeling

We develop and evaluate a strong neural sequence tagger on the full task of lexical semantic recognition with MWEs and noun/verb/preposition/possession supersenses to assess the performance of modern techniques on the full joint tagging task. Our tagger feeds pre-trained BERT representations (Devlin et al., 2019) through a biLSTM. An affine transformation followed by a linear chain conditional random field produces the final output. For further implementation details, see Appendix B.

The predicted tag for each token is the conjunction of its MWE, lexcat, and supersense.7 There are 572 such tags in the STREUSLE training set, and only 12 unique conjoined tags in the development set are unseen during training (≈5% of the development set tagging space, corresponding to ≈0.2% of the tokens in the development set).

**Constrained Decoding.** A few hard constraints are imposed in tagging. To enforce valid MWE chunks, we use first-order Viterbi decoding with the appropriate corpus-specific constraints (e.g., for STREUSLE MWEs, the BbIiOo_~ tagset; see §2.1.1). The MWE constraint is applied during training and evaluation. In addition, a given token’s possible lexcats are constrained by the token’s POS tag and lemma. For instance, a token with the AUX UPOS tag can only take the AUX lexcats. However, if the token’s UPOS is AUX and its lemma is “be”, it can take either the AUX or V lexcats.

The POS and lemma constraints are only applied during evaluation; to avoid relying on gold POS/lemma annotations at test time we use an off-the-shelf system (Qi et al., 2018).

### 3.1 Experiments

We train the tagger on version 4.3 of the English STREUSLE corpus and evaluate on the STREUSLE, English PARSEME, and DiMSUM test sets (§2). The latter two are (zero-shot) out-of-domain test sets; the tagger is not retrained on the associated shared task training data.

We also compare to a model with static word representations by replacing BERT with the concatenation of 300-dimensional pretrained GloVe embeddings (Pennington et al., 2014) and the output of a character-level convolutional neural net-
work. Finally, we also establish an upper bound on performance by providing the model with gold POS tags and lemmas; note that the difference between gold and predicted POS tags and lemmas only applies to the constrained decoding.

### 3.2 Results and Discussion

Table 1 shows all standard STREUSLE evaluation metrics on the test set. For preposition supersones (SNACS), we compare to the results in Schneider et al. (2018), who performed MWE identification and supersonse labeling for prepositions only. Note that Schneider et al. (2018) used version 4.0 of the STREUSLE corpus, which is slightly different from the version we use (some of the SNACS annotations have been revised). However, our baseline tagger, even with GloVe embeddings, outperforms Schneider et al. (2018) on that subset. Using BERT embeddings with constraints POS tags and lemmas improves performance substantially; on preposition supersonse tagging, it even outperforms using gold POS tags and lemmas. Liu et al. (2019) also found that BERT embeddings improved SNACS labeling on STREUSLE 4.0, although they study a simplified setting (gold preposition identification, and only considering single words).

Table 2 shows standard PARSEME and DiSUM test set evaluation metrics, for models trained on the STREUSLE training set, in a zero-shot out-of-domain evaluation setting. On the PARSEME test set, our BERT-based model approaches the state-of-the-art MWE-based F-score and exceeds the best reported fully-supervised token-based F-score. However, on the DiSUM test set, the BERT model did not outperform the best shared task system, likely owing to the comparative difficulty of the full lexical semantic recognition task versus the restricted DiSUM setting.

These results demonstrate that pre-training contextualized embeddings on large corpora can help models generalize to out-of-domain settings.

Constrained decoding does not substantially impact the performance of our BERT model. In general, constraints with gold POS/lemmas perform the best, while not using POS/lemma constraints is

---

### Table 1: STREUSLE test set results (%)

<table>
<thead>
<tr>
<th># Gold</th>
<th>Full</th>
<th>−LC</th>
<th>−SS</th>
<th>NOUN</th>
<th>VERB</th>
<th>SNACS</th>
<th>MWE</th>
<th>VERB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Gold</td>
<td></td>
<td></td>
<td>Labeled</td>
<td>F</td>
<td>Labeled</td>
<td>Role</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>5381</td>
<td>986</td>
<td>697</td>
<td>485</td>
<td>433.5</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td>GloVe (Gold)</td>
<td>82.5</td>
<td>79.3</td>
<td>82.7</td>
<td>89.9</td>
<td>69.0</td>
<td>66.1</td>
<td>77.1</td>
</tr>
<tr>
<td>BERT</td>
<td>GloVe (Pred.)</td>
<td>81.0</td>
<td>77.5</td>
<td>81.7</td>
<td>87.9</td>
<td>68.0</td>
<td>65.7</td>
<td>75.1</td>
</tr>
<tr>
<td>BERT</td>
<td>GloVe (None)</td>
<td>82.0</td>
<td>77.1</td>
<td>82.7</td>
<td>89.1</td>
<td>69.6</td>
<td>64.9</td>
<td>76.8</td>
</tr>
<tr>
<td>Schneider et al.</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>55.7</td>
<td>58.2</td>
<td>66.7</td>
</tr>
</tbody>
</table>

### Table 2: PARSEME and DiMSUM zero-shot test set results (%)

<table>
<thead>
<tr>
<th>MWE-based</th>
<th>Token-based</th>
<th>MWEs</th>
<th>Supersonses</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36.1</td>
<td>45.5</td>
<td>40.3</td>
<td>40.2</td>
<td>52.0</td>
</tr>
<tr>
<td>34.1</td>
<td>45.9</td>
<td>39.2</td>
<td>37.1</td>
<td>52.2</td>
</tr>
<tr>
<td>36.2</td>
<td>45.3</td>
<td>40.3</td>
<td>40.4</td>
<td>51.8</td>
</tr>
<tr>
<td>33.8</td>
<td>32.7</td>
<td>33.5</td>
<td>37.3</td>
<td>31.8</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>36.0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

DiMSUM 1.0 (test, 16,500 words)

<table>
<thead>
<tr>
<th></th>
<th>DiMSUM 1.0 (test, 5,381 words)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Gold</td>
<td>STREUSLE 4.3</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>5381</td>
<td>986</td>
<td>697</td>
<td>485</td>
<td>433.5</td>
</tr>
<tr>
<td></td>
<td>Labeled F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VERB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SNACS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Labeled</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VERB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MWE ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LinkAvg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I have a new born daughter and she helped me with a lot.

Go down 1 block to Super 8.

beware they will rip u off.

Figure 3: Selected examples where the model without MWE constraints (first row under each sentence) produces a structurally invalid tagging. Incorrect tags are red; the ones that render the tagging structurally invalid are bold. The last row under each sentence is the gold annotation, and the middle row (if different from gold) is the model prediction with MWE constraints. (The first sentence ends with a period, omitted for brevity.)

often better than using predicted POS/lemmas. Removing the MWE constraints yields models with slightly higher overall tag accuracy, but results in invalid segmentations for a large proportion of sentences: 14% of STREUSLE sentences in the fully unconstrained model and 17% of sentences if only predicted POS and lemmas are used for constraints.

Three sentences out of those 17% appear in figure 3. The first shows both an omission of a “B-” tag needed to start an MWE (“new”) and a false positive gap without members of an MWE on either side (“me”). When the full set of constraints is used, the gold tagging is recovered. In the second sentence, there is a false positive yet structurally valid MWE (“Go down”) as well as an invalid start to an MWE that is never continued (“Super”), perhaps because it is rare for a number to continue an MWE (this happens <20 times in the entire corpus). Finally, in the third sentence, the model constrained only by POS and lemma is inclined toward the literal meaning of “rip”, whereas the MWE-constrained model recovers the gappy verb-particle construction “rip off”. Naturally, in other sentences, the MWE-constrained model sometimes suffers from false positive or false negative MWEs, but always produces a coherent segmentation.

4 Related Work

The computational study of MWEs has a long history (Sag et al., 2002; Diab and Bhutada, 2009; Baldwin and Kim, 2010; Ramisch, 2015; Qu et al., 2015; Constant et al., 2017; Bingel and Søgaard, 2017; Shwartz and Dagan, 2019), as does supersense tagging (Segond et al., 1997; Ciaramita and Altun, 2006). Vincze et al. (2011) developed a sequence tagger for both MWEs and named entities in English. Schneider and Smith (2015); Schneider et al. (2016) featured joint tagging of MWEs and noun and verb supersenses with feature-based sequence models. Richardson (2017) trained such a model on STREUSLE 3.0 as a noun, verb, and preposition supersense tagger (without modeling MWEs). For preposition supersenses, Goen and Goldberg (2016) incorporated multilingual cues; Schneider et al. (2018) experimented with feature-based and neural classifiers; and Liu et al. (2019), modeling supersense disambiguation of single-word prepositions only, found pretrretrained contextual embeddings to be much more effective even with simple linear probing models.

5 Conclusion

We study the lexical semantic recognition task defined by the STREUSLE corpus, which involves joint MWE identification and coarse-grained (supersense) disambiguation of noun, verb, and preposition expressions; this task subsumes and unifies the previous PARSEME and DiMSUM evaluations. We develop a strong baseline neural sequence model, and see encouraging results on the task. Furthermore, zero-shot out-of-domain evaluation of our baselines on partial versions of the task yields scores comparable to the fully-supervised in-domain state of the art.

Acknowledgments

We are grateful to anonymous reviewers as well as members of the NERT lab for their feedback on this work. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel. NL is supported by an NSF Graduate Research Fellowship under grant number DGE-1656518.
References


