Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut

Nathan Schneider **Emily Danchik** **Chris Dyer** **Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nschneid,emilydan,cdyer,nasmith}@cs.cmu.edu

Abstract

We present a novel representation, evaluation measure, and supervised models for the task of identifying the multiword expressions (MWEs) in a sentence, resulting in a *lexical semantic segmentation*. Our approach generalizes a standard chunking representation to encode MWEs containing gaps, thereby enabling efficient sequence tagging algorithms for feature-rich discriminative models. Experiments on a new dataset of English web text offer the first linguistically-driven evaluation of MWE identification with truly heterogeneous expression types. Our statistical sequence model greatly outperforms a lookup-based segmentation procedure, achieving nearly 60% F_1 for MWE identification.

1. **MW named entities:** *Prime Minister Tony Blair*
2. **MW compounds:** *hot air balloon*, *skinny dip*
3. **conventionally SW compounds:** *somewhere*
4. **verb-particle:** *pick up*, *dry out*, *take over*, *cut short*
5. **verb-preposition:** *refer to*, *depend on*, *look for*
6. **verb-noun(-preposition):** *pay attention (to)*
7. **support verb:** *make decisions*, *take pictures*
8. **other phrasal verb:** *put up with*, *get rid of*
9. **PP modifier:** *above board*, *at all*, *from time to time*
10. **coordinated phrase:** *cut and dry*, *more or less*
11. **connective:** *as well as*, *let alone*, *in spite of*
12. **semi-fixed VP:** *pick up where* <one> *left off*
13. **fixed phrase:** *scared to death*, *leave of absence*
14. **phatic:** *You're welcome. Me neither!*
15. **proverb:** *Beggars can't be choosers.*

Figure 1: Some of the classes of idioms in English. The examples included here contain multiple lexicalized words—with the exception of those in (3), if the conventional single-word (SW) spelling is used.

1 Introduction

Language has a knack for defying expectations when put under the microscope. For example, there is the notion—sometimes referred to as *compositionality*—that words will behave in predictable ways, with individual meanings that combine to form complex meanings according to general grammatical principles. Yet language is awash with examples to the contrary: in particular, idiomatic expressions such as *awash with* NP, *have a knack for* $\text{VP}_{\text{-}ing}$, *to the contrary*, and *defy expectations*. Thanks to processes like metaphor and grammaticalization, these are (to various degrees) semantically opaque, structurally fossilized, and/or statistically idiosyncratic. In other words, idiomatic expressions may be exceptional in form, function, or distribution. They are so diverse, so unruly, so difficult to circumscribe, that entire theories of syntax are predicated on the notion that constructions with idiosyncratic form-meaning mappings (Fillmore et al., 1988; Goldberg, 1995) or statistical properties (Goldberg, 2006) offer crucial evidence about the grammatical organization of language.

Here we focus on **multiword expressions** (MWEs): *lexicalized* combinations of two or more words that are exceptional enough to be considered as single units in the lexicon. As figure 1 illustrates, MWEs occupy diverse syntactic and semantic functions. Within MWEs, we distinguish (a) proper names and (b) lexical idioms. The latter have proved themselves a "pain in the neck for NLP" (Sag et al., 2002). Automatic and efficient detection of MWEs, though far from solved, would have diverse appli-

Transactions of the Association for Computational Linguistics, 2 (2014) 193–206. Action Editor: Joakim Nivre.
Submitted 12/2013; Revised 1/2014; Published 4/2014.

cations including machine translation (Carpuat and Diab, 2010), information retrieval (Newman et al., 2012), opinion mining (Berend, 2011), and second language learning (Ellis et al., 2008).

It is difficult to establish any comprehensive taxonomy of multiword idioms, let alone develop linguistic criteria and corpus resources that cut across these types. Consequently, the voluminous literature on MWEs in computational linguistics—see §7, Baldwin and Kim (2010), and Ramisch (2012) for surveys—has been fragmented, looking (for example) at subclasses of phrasal verbs or nominal compounds in isolation. To the extent that MWEs have been annotated in existing corpora, it has usually been as a secondary aspect of some other scheme. Traditionally, such resources have prioritized certain kinds of MWEs to the exclusion of others, so they are not appropriate for evaluating general-purpose identification systems.

In this article, we briefly review a shallow form of analysis for MWEs that is neutral to expression type, and that facilitates free text annotation without requiring a prespecified MWE lexicon (§2). The scheme applies to gappy (discontinuous) as well as contiguous expressions, and allows for a qualitative distinction of association strengths. In Schneider et al. (2014) we have applied this scheme to fully annotate a 55,000-word corpus of English web reviews (Bies et al., 2012a), a conversational genre in which colloquial idioms are highly salient. This article's main contribution is to show that the representation—constrained according to linguistically motivated assumptions (§3)—can be transformed into a sequence tagging scheme that resembles standard approaches in named entity recognition and other text chunking tasks (§4). Along these lines, we develop a discriminative, structured model of MWEs in context (§5) and train, evaluate, and examine it on the annotated corpus (§6). Finally, in §7 and §8 we comment on related work and future directions.

2 Annotated Corpus

To build and evaluate a multiword expression analyzer, we use the MWE-annotated corpus of Schneider et al. (2014). It consists of informal English web text that has been specifically and completely annotated for MWEs, without reference to any particular lexicon. To the best of our knowledge, this corpus is the first to be freely annotated for many kinds of MWEs (without reference to a lexicon), and is also the first dataset of social media text with MWE annotations beyond named entities. This section gives a synopsis of the annotation conventions used to develop that resource, as they are important to understanding our models and evaluation.

Rationale. The multiword expressions community has lacked a canonical corpus resource comparable to benchmark datasets used for problems such as NER and parsing. Consequently, the MWE literature has been driven by lexicography: typically, the goal is to acquire an MWE lexicon with little or no supervision, or to apply such a lexicon to corpus data. Studies of MWEs in context have focused on various subclasses of constructions in isolation, necessitating special-purpose datasets and evaluation schemes. By contrast, Schneider et al.'s (2014) corpus creates an opportunity to tackle *general-purpose* MWE identification, such as would be desirable for use by high-coverage downstream NLP systems. It is used to train and evaluate our models below. The corpus is publicly available as a benchmark for further research.[1]

Data. The documents in the corpus are online user reviews of restaurants, medical providers, retailers, automotive services, pet care services, etc. Marked by conversational and opinionated language, this genre is fertile ground for colloquial idioms (Nunberg et al., 1994; Moon, 1998). The 723 reviews (55,000 words, 3,800 sentences) in the English Web Treebank (WTB; Bies et al., 2012b) were collected by Google, tokenized, and annotated with phrase structure trees in the style of the Penn Treebank (Marcus et al., 1993). MWE annotators used the sentence and word tokenizations supplied by the treebank.[2]

Annotation scheme. The annotation scheme itself was designed to be as simple as possible. It consists of grouping together the tokens in each sentence that belong to the same MWE instance. While annotation guidelines provide examples of MWE groupings in a wide range of constructions, the annotator is not

[1] http://www.ark.cs.cmu.edu/LexSem/

[2] Because we use treebank data, syntactic parses are available to assist in post hoc analysis. Syntactic information was not shown to annotators.

	# of constituent tokens				*# of gaps*		
	2	**3**	**≥4**	**total**	**0**	**1**	**2**
strong	2257	595	172	3024	2626	394	4
weak	269	121	69	459	322	135	2
	2526	716	241	3483	2948	529	6

Table 1: Counts in the MWE corpus.

tied to any particular taxonomy or syntactic structure. This simplifies the number of decisions that have to be made for each sentence, even if some are difficult.

Further instructions to annotators included:

- Groups should include only the lexically fixed parts of an expression (modulo inflectional morphology); this generally excludes determiners and pronouns: ***made*** *the* ***mistake***, ***pride*** *themselves* ***on***.
- Multiword proper names count as MWEs.
- Misspelled or unconventionally spelled tokens are interpreted according to the intended word if clear.
- Overtokenized words (spelled as two tokens, but conventionally one word) are joined as multiwords. Clitics separated by the tokenization in the corpus—negative *n't*, possessive *'s*, etc.—are joined if functioning as a fixed part of a multiword (e.g., ***T 's Cafe***), but not if used productively.

Gaps. There are, broadly speaking, three reasons to group together tokens that are not fully contiguous. Most commonly, gaps contain internal modifiers, such as *good* in ***make*** *good* ***decisions***. Syntactic constructions such as the passive can result in gaps that might not otherwise be present: in *good* ***decisions*** *were* ***made***, there is instead a gap filled by the passive auxiliary. Finally, some MWEs may take internal arguments: *they* ***gave*** *me* ***a break***. Figure 1 has additional examples. Multiple gaps can occur even within the same expression, though it is rare: *they agreed to* ***give*** *Bob* ***a*** *well-deserved* ***break***.

Strength. The annotation scheme has two "strength" levels for MWEs. Clearly idiomatic expressions are marked as strong MWEs, while mostly compositional but especially frequent collocations/phrases (e.g., *abundantly clear* and *patently obvious*) are marked as weak MWEs. Weak multiword groups are allowed to include strong MWEs as constituents (but not vice versa). Strong groups are required to cohere when used inside weak groups: that is, a weak group cannot include only part of a strong group. For purposes of annotation, there were no constraints hinging on the ordering of tokens in the sentence.

Process. MWE annotation proceeded one sentence at a time. The 6 annotators referred to and improved the guidelines document on an ongoing basis. Every sentence was seen independently by at least 2 annotators, and differences of opinion were discussed and resolved (often by marking a weak MWE as a compromise). See Schneider et al. (2014) for details.

Statistics. The annotated corpus consists of 723 documents (3,812 sentences). MWEs are frequent in this domain: 57% of sentences (72% of sentences over 10 words long) and 88% of documents contain at least one MWE. $8,060/55,579$=15% of tokens belong to an MWE; in total, there are 3,483 MWE instances. 544 (16%) are strong MWEs containing a gold-tagged proper noun—most are proper names. A breakdown appears in table 1.

3 Representation and Task Definition

We define a **lexical segmentation** of a sentence as a partitioning of its tokens into segments such that each segment represents a single unit of lexical meaning. A *multiword* lexical expression may contain **gaps**, i.e. interruptions by other segments. We impose two restrictions on gaps that appear to be well-motivated linguistically:

- **Projectivity:** Every expression filling a gap must be completely contained within that gap; gappy expressions may not interleave.
- **No nested gaps:** A gap in an expression may be filled by other single- or multiword expressions, so long as those do not themselves contain gaps.

Formal grammar. Our scheme corresponds to the following extended CFG (Thatcher, 1967), where S is the full sentence and terminals w are word tokens:

$$S \rightarrow X^+$$
$$X \rightarrow \underline{w}^+ \ (Y^+ \ \underline{w}^+)^*$$
$$Y \rightarrow \underline{w}^+$$

Each expression X or Y is lexicalized by the words in one or more underlined variables on the right-hand side. An X constituent may optionally contain one or more gaps filled by Y constituents, which must not contain gaps themselves.[3]

[3] MWEs with multiple gaps are rare but attested in data: e.g., ***putting*** *me* ***at*** *my* ***ease***. We encountered one violation of the gap nesting constraint in the reviews data: *I* $have^2_1$ $nothing^2_1$ but^2_1 *fantastic things*^{2} to^2_1 say^2_1 . Additionally, the interrupted phrase

Denoting multiword groupings with subscripts, *My wife had* $taken_1$ *her '07*$_2$ *Ford*$_2$ *Fusion*$_2$ in_1 *for a routine* oil_3 $change_3$ contains 3 multiword groups—{*taken*, *in*}, {*'07*, *Ford*, *Fusion*}, {*oil*, *change*}—and 7 single-word groups. The first MWE is gappy (accentuated by the box); a single word and a contiguous multiword group fall within the gap. The projectivity constraint forbids an analysis like $taken_1$ *her '07*$_2$ *Ford*$_1$ *Fusion*$_2$, while the gap nesting constraint forbids $taken_1$ *her*$_2$ *'07* *Ford*$_2$ *Fusion*$_2$ in_1.

3.1 Two-level Scheme: Strong vs. Weak MWEs

Our annotated data distinguish two strengths of MWEs as discussed in §2. Augmenting the grammar of the previous section, we therefore designate nonterminals as strong ($\overline{X}$, $\overline{Y}$) or weak ($\tilde{X}$, $\tilde{Y}$):

$$
\begin{array}{lcl}
S & \rightarrow & \tilde{X}^{+} \\
\tilde{X} & \rightarrow & \underline{\overline{X}}^{+} \;\; (\tilde{Y}^{+} \;\; \underline{\overline{X}}^{+})^{*} \\
\overline{X} & \rightarrow & \underline{w}^{+} \;\; (\tilde{Y}^{+} \;\; \underline{w}^{+})^{*} \\
\tilde{Y} & \rightarrow & \underline{\overline{Y}}^{+} \\
\overline{Y} & \rightarrow & \underline{w}^{+}
\end{array}
$$

A weak MWE may be lexicalized by single words and/or strong multiwords. Strong multiwords cannot contain weak multiwords except in gaps. Further, the contents of a gap cannot be part of any multiword that extends outside the gap.[4]

For example, consider the segmentation: *he was willing to* $budge_1$ a_2 $little_2$ on_1 *the price which* $means^4$ a_3^4 lot_3^4 to^4 me^4. Subscripts denote strong MW groups and superscripts weak MW groups; unmarked tokens serve as single-word expressions. The MW groups are thus {*budge*, *on*}, {*a*, *little*}, {*a*, *lot*}, and {*means*, {*a*, *lot*}, *to*, *me*}. As should be evident from the grammar, the projectivity and gap-nesting constraints apply here just as in the 1-level scheme.

3.2 Evaluation

Matching criteria. Given that most tokens do not belong to an MWE, to evaluate MWE identification we adopt a precision/recall-based measure from the coreference resolution literature. The MUC criterion (Vilain et al., 1995) measures precision and recall of links in terms of groups (units) implied by the transitive closure over those links.[5] It can be defined as follows:

Let $a - b$ denote a link between two elements in the gold standard, and $a \hat{-} b$ denote a link in the system prediction. Let the $*$ operator denote the transitive closure over all links, such that $[\![a -^{*} b]\!]$ is 1 if a and b belong to the same (gold) set, and 0 otherwise. Assuming there are no redundant[6] links within any annotation (which in our case is guaranteed by linking consecutive words in each MWE), we can write the MUC precision and recall measures as:

$$P = \frac{\sum_{a,b:a \hat{-} b} [\![a -^{*} b]\!]}{\sum_{a,b:a \hat{-} b} 1} \qquad R = \frac{\sum_{a,b:a - b} [\![a \hat{-}^{*} b]\!]}{\sum_{a,b:a - b} 1}$$

This awards partial credit when predicted and gold expressions overlap in part. Requiring full MWEs to match exactly would arguably be too stringent, overpenalizing larger MWEs for minor disagreements. We combine precision and recall using the standard F_1 measure of their harmonic mean. This is the **link-based** evaluation used for most of our experiments. For comparison, we also report some results with a more stringent **exact match** evaluation where the span of the predicted MWE must be identical to the span of the gold MWE for it to count as correct.

Strength averaging. Recall that the 2-level scheme (§3.1) distinguishes *strong* vs. *weak* links/groups, where the latter category applies to reasonably compositional collocations as well as ambiguous or difficult cases. If where one annotation uses a weak link the other has a strong link or no link at all, we want to penalize the disagreement less than if one had a strong link and the other had no link. To accommodate the 2-level scheme, we therefore average $F_1^{\uparrow}$, in which all weak links have been converted to strong links, and $F_1^{\downarrow}$, in which they have been removed: $F_1 = \frac{1}{2}(F_1^{\uparrow} + F_1^{\downarrow})$.[7] If neither annotation contains any weak links, this equals the MUC

great gateways never^1 *before*^{1} , so_3^2 far_3^2 as_3^2 *Hudson knew*^{2} , *seen*^{1} *by Europeans* was annotated in another corpus.

[4] This was violated 6 times in our annotated data: modifiers within gaps are sometimes collocated with the gappy expression, as in on_2^1 a_2^1 *tight*^{1} $budget_2^1$ and $have_2^1$ *little*^{1} $doubt_2^1$.

[5] As a criterion for coreference resolution, the MUC measure has perceived shortcomings which have prompted several other measures (see Recasens and Hovy, 2011 for a review). It is not clear, however, whether any of these criticisms are relevant to MWE identification.

[6] A link between a and b is redundant if the other links already imply that a and b belong to the same set. A set of N elements is expressed non-redundantly with exactly $N - 1$ links.

[7] Overall precision and recall are likewise computed by averaging "strengthened" and "weakened" measurements.

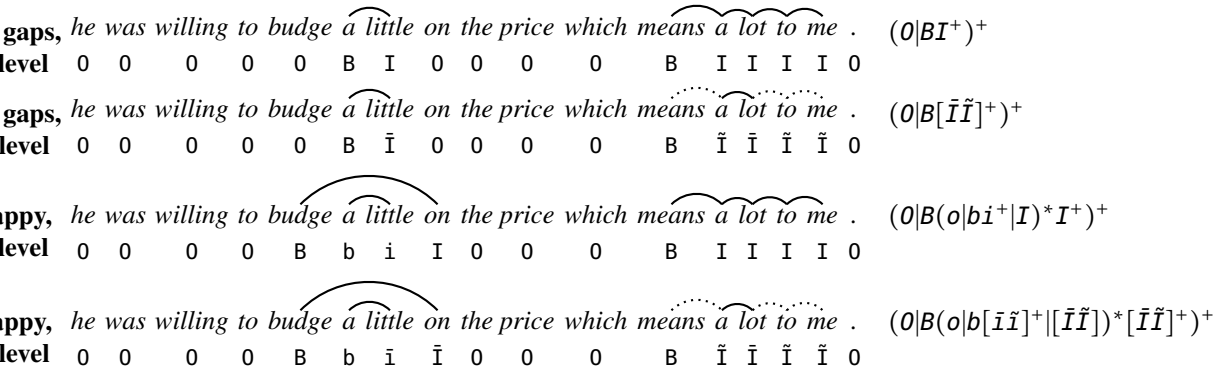

Figure 2: Examples and regular expressions for the 4 tagging schemes. Strong links are depicted with solid arcs, and weak links with dotted arcs. The bottom analysis was provided by an annotator; the ones above are simplifications.

score because $F_1 = F_1^{\uparrow} = F_1^{\downarrow}$. This method applies to both the link-based and exact match evaluation criteria.

4 Tagging Schemes

Following (Ramshaw and Marcus, 1995), shallow analysis is often modeled as a sequence-chunking task, with tags containing chunk-positional information. The BIO scheme and variants (e.g., BILOU; Ratinov and Roth, 2009) are standard for tasks like named entity recognition, supersense tagging, and shallow parsing.

The language of derivations licensed by the grammars in §3 allows for a tag-based encoding of MWE analyses with only bigram constraints. We describe 4 tagging schemes for MWE identification, starting with BIO and working up to more expressive variants. They are depicted in figure 2.

No gaps, 1-level (3 tags). This is the standard contiguous chunking representation from Ramshaw and Marcus (1995) using the tags {O B I}. O is for tokens **o**utside any chunk; B marks tokens **b**eginning a chunk; and I marks other tokens **i**nside a chunk. Multiword chunks will thus start with B and then I. B must always be followed by I; I is not allowed at the beginning of the sentence or following O.

No gaps, 2-level (4 tags). We can distinguish strength levels by splitting I into two tags: Ī for strong expressions and Ĩ for weak expressions. To express strong and weak contiguous chunks requires 4 tags: {O B Ī Ĩ}. (Marking B with a strength as well would be redundant because MWEs are never length-one chunks.) The constraints on Ī and Ĩ are the same as the constraints on I in previous schemes. If Ī and Ĩ occur next to each other, the strong attachment will receive higher precedence, resulting in analysis of strong MWEs as nested within weak MWEs.

Gappy, 1-level (6 tags). Because gaps cannot themselves contain gappy expressions (we do not support full recursivity), a finite number of additional tags are sufficient to encode gappy chunks. We therefore add lowercase tag variants representing tokens *within a gap*: {O o B b I i}. In addition to the constraints stated above, no within-gap tag may occur at the beginning or end of the sentence or immediately following or preceding O. Within a gap, b, i, and o behave like their out-of-gap counterparts.

Gappy, 2-level (8 tags). 8 tags are required to encode the 2-level scheme with gaps: {O o B b Ī ī Ĩ ĩ}. Variants of the inside tag are marked for strength of the incoming link—this applies gap-externally (capitalized tags) and gap-internally (lowercase tags). If Ī or Ĩ immediately follows a gap, its diacritic reflects the strength of the gappy expression, not the gap's contents.

5 Model

With the above representations we model MWE identification as sequence tagging, one of the paradigms that has been used previously for identifying *contiguous* MWEs (Constant and Sigogne, 2011, see §7).[8] Constraints on legal tag bigrams are sufficient to ensure the full tagging is well-formed subject to the regular expressions in figure 2; we enforce these

[8] Hierarchical modeling based on our representations is left to future work.

constraints in our experiments.[9]

In NLP, conditional random fields (Lafferty et al., 2001) and the structured perceptron (Collins, 2002) are popular techniques for discriminative sequence modeling with a convex loss function. We choose the second approach for its speed: learning and inference depend mainly on the runtime of the Viterbi algorithm, whose asymptotic complexity is linear in the length of the input and (with a first-order Markov assumption) quadratic in the number of tags. Below, we review the structured perceptron and discuss our cost function, features, and experimental setup.

5.1 Cost-Augmented Structured Perceptron

The structured perceptron's (Collins, 2002) learning procedure, algorithm 1, generalizes the classic perceptron algorithm (Freund and Schapire, 1999) to incorporate a structured decoding step (for sequences, the Viterbi algorithm) in the inner loop. Thus, training requires only max inference, which is fast with a first-order Markov assumption. In training, features are adjusted where a tagging error is made; the procedure can be viewed as optimizing the structured hinge loss. The output of learning is a weight vector that parametrizes a feature-rich scoring function over candidate labelings of a sequence.

To better align the learning algorithm with our F-score–based MWE evaluation (§3.2), we use a cost-augmented version of the structured perceptron that is sensitive to different kinds of errors during training. When recall is the bigger obstacle, we can adopt the following cost function: given a sentence $\mathbf{x}$, its gold labeling $\mathbf{y}^*$, and a candidate labeling $\mathbf{y}'$,

$$cost(\mathbf{y}^*, \mathbf{y}', \mathbf{x}) = \sum_{j=1}^{|\mathbf{y}^*|} c(y^*_j, y'_j) \quad \text{where}$$

$$c(y^*, y') = [\![y^* \neq y']\!] + \rho [\![y^* \in \{\mathtt{B}, \mathtt{b}\} \wedge y' \in \{\mathtt{O}, \mathtt{o}\}]\!]$$

A single nonnegative hyperparameter, ρ, controls the tradeoff between recall and accuracy; higher ρ biases the model in favor of recall (possibly hurting accuracy and precision). This is a slight variant of the recall-oriented cost function of Mohit et al. (2012). The difference is that we only penalize *beginning-of-expression* recall errors. Preliminary experiments showed that a cost function penalizing all recall errors—i.e., with $\rho [\![y^* \neq \mathtt{O} \wedge y' = \mathtt{O}]\!]$ as the second term, as in Mohit et al.—tended to append additional tokens to high-confidence MWEs (such as proper names) rather than encourage new MWEs, which would require positing at least two new non-outside tags.

Input: data $\left\langle \langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle \right\rangle_{n=1}^{N}$; number of iterations M
$\mathbf{w} \leftarrow \mathbf{0}$
$\overline{\mathbf{w}} \leftarrow \mathbf{0}$
$t \leftarrow 1$
for $m = 1$ **to** M **do**
  for $n = 1$ **to** N **do**
    $\langle \mathbf{x}, \mathbf{y} \rangle \leftarrow \langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle$
    $\hat{\mathbf{y}} \leftarrow \arg\max_{\mathbf{y}'} (\mathbf{w}^\top \mathbf{g}(\mathbf{x}, \mathbf{y}') + cost(\mathbf{y}, \mathbf{y}', \mathbf{x}))$
    if $\hat{\mathbf{y}} \neq \mathbf{y}$ **then**
      $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{g}(\mathbf{x}, \mathbf{y}) - \mathbf{g}(\mathbf{x}, \hat{\mathbf{y}})$
      $\overline{\mathbf{w}} \leftarrow \overline{\mathbf{w}} + t\mathbf{g}(\mathbf{x}, \mathbf{y}) - t\mathbf{g}(\mathbf{x}, \hat{\mathbf{y}})$
    end
    $t \leftarrow t + 1$
  end
end
Output: $\mathbf{w} - (\overline{\mathbf{w}}/t)$

Algorithm 1: Training with the averaged perceptron. (Adapted from Daumé, 2006, p. 19.)

5.2 Features

Basic features. These are largely based on those of Constant et al. (2012): they look at word unigrams and bigrams, character prefixes and suffixes, and POS tags, as well as lexicon entries that match lemmas[10] of multiple words in the sentence. Appendix A lists the basic features in detail.

Some of the basic features make use of *lexicons*. We use or construct 10 lists of English MWEs: all multiword entries in **WordNet** (Fellbaum, 1998); all multiword chunks in **SemCor** (Miller et al., 1993); all multiword entries in English **Wiktionary**;[11] the **WikiMwe** dataset mined from English Wikipedia (Hartmann et al., 2012); the **SAID** database of phrasal lexical idioms (Kuiper et al., 2003); the named entities and other MWEs in the WSJ corpus on the English side of the **CEDT** (Hajič et al., 2012);

[9] The 8-tag scheme licenses 42 tag bigrams: sequences such as B O and o I are prohibited. There are also constraints on the allowed tags at the beginning and end of the sequence.

[10] The WordNet API in NLTK (Bird et al., 2009) was used for lemmatization.

[11] `http://en.wiktionary.org`; data obtained from `https://toolserver.org/~enwikt/definitions/enwikt-defs-20130814-en.tsv.gz`

			LOOKUP				SUPERVISED MODEL			
preexising lexicons	entries	max gap	$\bar{P}$	$\bar{R}$	$\overline{F_1}$	σ	$\bar{P}$	$\bar{R}$	$\overline{F_1}$	σ
none	0	length					74.39	44.43	55.57	2.19
WordNet + SemCor	71k	0	46.15	28.41	35.10	2.44	74.51	45.79	56.64	1.90
6 lexicons	420k	0	35.05	46.76	40.00	2.88	76.08	**52.39**	**61.95**	1.67
10 lexicons	437k	0	33.98	47.29	39.48	2.88	75.95	51.39	61.17	2.30
best configuration with in-domain lexicon		1	**46.66**	**47.90**	**47.18**	2.31	**76.64**	51.91	61.84	1.65
			2 lexicons + *MWtypes*(train)$_{\geq 1}$				6 lexicons + *MWtypes*(train)$_{\geq 2}$			

Table 2: Use of lexicons for lookup-based vs. statistical segmentation. Supervised learning used only basic features and the structured perceptron, with the 8-tag scheme. Results are with the link-based matching criterion for evaluation.

Top: Comparison of preexisting lexicons. "6 lexicons" refers to WordNet and SemCor plus SAID, WikiMwe, Phrases.net, and English Wiktionary; "10 lexicons" adds MWEs from CEDT, VNC, LVC, and Oyz. (In these lookup-based configurations, allowing gappy MWEs never helps performance.)

Bottom: Combining preexisting lexicons with a lexicon derived from MWEs annotated in the training portion of each cross-validation fold at least once (lookup) or twice (model).

All precision, recall, and F_1 percentages are averaged across 8 folds of cross-validation on **train**; standard deviations are shown for the F_1 score. In each column, the highest value using only preexisting lexicons is underlined, and the highest overall value is bolded. The boxed row indicates the configuration used as the basis for subsequent experiments.

the **verb-particle constructions** (VPCs) dataset of (Baldwin, 2008); a list of **light verb constructions** (LVCs) provided by Claire Bonial; and two idioms websites.[12] After preprocessing, each lexical entry consists of an ordered sequence of word lemmas, some of which may be variables like <*something*>.

Given a sentence and one or more of the lexicons, lookup proceeds as follows: we enumerate entries whose lemma sequences match a sequence of lemmatized tokens, and build a lattice of possible analyses over the sentence. We find the shortest path (i.e., using as few expressions as possible) with dynamic programming, allowing gaps of up to length 2.[13]

Unsupervised word clusters. Distributional clustering on large (unlabeled) corpora can produce lexical generalizations that are useful for syntactic and semantic analysis tasks (e.g.: Miller et al., 2004; Koo et al., 2008; Turian et al., 2010; Owoputi et al., 2013; Grave et al., 2013). We were interested to see whether a similar pattern would hold for MWE identification, given that MWEs are concerned with what is lexically *idiosyncratic*—i.e., backing off from specific lexemes to word classes may lose the MWE-relevant information. Brown clustering[14] (Brown et al., 1992) on the 21-million-word Yelp Academic Dataset[15] (which is similar in genre to the annotated web reviews data) gives us a hard clustering of word types. To our tagger, we add features mapping the previous, current, and next token to Brown cluster IDs. The feature for the current token conjoins the word lemma with the cluster ID.

Part-of-speech tags. We compared three PTB-style POS taggers on the full REVIEWS subcorpus (**train+test**). The Stanford CoreNLP tagger[16] (Toutanova et al., 2003) yields an accuracy of 90.4%. The ARK TweetNLP tagger v. 0.3.2 (Owoputi et al., 2013) achieves 90.1% with the model[17] trained on the Twitter corpus of Ritter et al. (2011), and 94.9% when trained on the ANSWERS, EMAIL, NEWSGROUP, and WEBLOG subcorpora of WTB. We use this third configuration to produce automatic POS tags for training and testing our MWE tagger. (A comparison condition in §6.3 uses oracle POS tags.)

5.3 Experimental Setup

The corpus of web reviews described in §2 is used for training and evaluation. 101 arbitrarily chosen documents (500 sentences, 7,171 words) were held

[12] `http://www.phrases.net/` and `http://home.postech.ac.kr/~oyz/doc/idiom.html`

[13] Each top-level lexical expression (single- or multiword) incurs a cost of 1; each expression within a gap has cost 1.25.

[14] With Liang's (2005) implementation: `https://github.com/percyliang/brown-cluster`. We obtain 1,000 clusters from words appearing at least 25 times.

[15] `https://www.yelp.com/academic_dataset`

[16] v. 3.2.0, with `english-bidirectional-distsim`

[17] `http://www.ark.cs.cmu.edu/TweetNLP/model.ritter_ptb_alldata_fixed.20130723`

				LINK-BASED			EXACT MATCH		
configuration	M	ρ	$\lvert\mathbf{w}\rvert$	P	R	F_1	P	R	F_1
base model	5	—	1,765k	69.27	50.49	58.35	60.99	48.27	53.85
+ recall cost	4	150	1,765k	61.09	57.94	59.41	53.09	55.38	54.17
+ clusters	3	100	2,146k	63.98	55.51	59.39	56.34	53.24	54.70
+ oracle POS	4	100	2,145k	66.19	59.35	62.53	58.51	57.00	57.71

Table 3: Comparison of supervised models on **test** (using the 8-tag scheme). The base model corresponds to the boxed result in table table 2, but here evaluated on **test**. For each configuration, the number of training iterations M and (except for the base model) the recall-oriented hyperparameter ρ were tuned by cross-validation on **train**.

out as a final **test** set. This left 3,312 sentences/ 48,408 words for training/development (**train**). Feature engineering and hyperparameter tuning were conducted with 8-fold cross-validation on **train**. The 8-tag scheme is used except where otherwise noted.

In learning with the structured perceptron (algorithm 1), we employ two well-known techniques that can both be viewed as regularization. First, we use the average of parameters over all timesteps of learning. Second, within each cross-validation fold, we determine the number of training iterations (epochs) M by early stopping—that is, after each iteration, we use the model to decode the held-out data, and when that accuracy ceases to improve, use the previous model. The two hyperparameters are the number of iterations and the value of the recall cost hyperparameter (ρ). Both are tuned via cross-validation on **train**; we use the multiple of 50 that maximizes average link-based F_1. The chosen values are shown in table 3. Experiments were managed with the `ducttape` tool.[18]

6 Results

We experimentally address the following questions to probe and justify our modeling approach.

6.1 Is supervised learning necessary?

Previous MWE identification studies have found benefit to statistical learning over heuristic lexicon lookup (Constant and Sigogne, 2011; Green et al., 2012). Our first experiment tests whether this holds for comprehensive MWE identification: it compares our supervised tagging approach with baselines of heuristic lookup on preexisting lexicons. The baselines construct a lattice for each sentence using the same method as lexicon-based model features (§5.2). If multiple lexicons are used, the union of their entries is used to construct the lattice. The resulting segmentation—which does not encode a strength distinction—is evaluated against the gold standard.

Table 2 shows the results. Even with just the labeled training set as input, the supervised approach beats the strongest heuristic baseline (that incorporates in-domain lexicon entries extracted from the training data) by 30 precision points, while achieving comparable recall. For example, the baseline (but not the statistical model) incorrectly predicts an MWE in *places to **eat in** Baltimore* (because *eat in*, meaning 'eat at home,' is listed in WordNet). The supervised approach has learned not to trust WordNet too much due to this sort of ambiguity. Downstream applications that currently use lexicon matching for MWE identification (e.g., Ghoneim and Diab, 2013) likely stand to benefit from our statistical approach.

6.2 How best to exploit MWE lexicons (type-level information)?

For statistical tagging (right portion of table 2), using more *preexisting* (out-of-domain) lexicons generally improves recall; precision also improves a bit.

A lexicon of MWEs occurring in the non-held-out training data *at least twice*[19] (table 2, bottom right) is marginally worse (better precision/worse recall) than the best result using only preexisting lexicons.

6.3 Variations on the base model

We experiment with some of the modeling alternatives discussed in §5. Results appear in table 3 under both the link-based and exact match evaluation criteria. We note that the exact match scores are (as expected) several points lower.

[18] `https://github.com/jhclark/ducttape/`

[19] If we train with access to the full lexicon of *training set* MWEs, the learner credulously overfits to relying on that lexicon—after all, it has perfect coverage of the training data!—which proves fatal for the model at test time.

Recall-oriented cost. The recall-oriented cost adds about 1 link-based F_1 point, sacrificing precision in favor of recall.

Unsupervised word clusters. When combined with the recall-oriented cost, these produce a slight improvement to precision/degradation to recall, improving exact match F_1 but not affecting link-based F_1. Only a few clusters receive high positive weight; one of these consists of *matter*, *joke*, *biggie*, *pun*, *avail*, *clue*, *corkage*, *frills*, *worries*, etc. These words are diverse semantically, but all occur in collocations with *no*, which is what makes the cluster coherent and useful to the MWE model.

Oracle part-of-speech tags. Using human-annotated rather than automatic POS tags improves MWE identification by about 3 F_1 points on **test** (similar differences were observed in development).

6.4 What are the highest-weighted features?

An advantage of the linear modeling framework is that we can examine learned feature weights to gain some insight into the model's behavior.

In general, the highest-weighted features are the lexicon matching features and features indicative of proper names (POS tag of proper noun, capitalized word not at the beginning of the sentence, etc.).

Despite the occasional cluster capturing collocational or idiomatic groupings, as described in the previous section, the clusters appear to be mostly useful for identifying words that tend to belong (or not) to proper names. For example, the cluster with *street*, *road*, *freeway*, *highway*, *airport*, etc., as well as words outside of the cluster vocabulary, weigh in favor of an MWE. A cluster with everyday destinations (*neighborhood*, *doctor*, *hotel*, *bank*, *dentist*) prefers non-MWEs, presumably because these words are not typically part of proper names in this corpus. This was from the best model using non-oracle POS tags, so the clusters are perhaps useful in correcting for proper nouns that were mistakenly tagged as common nouns. One caveat, though, is that it is hard to discern the impact of these specific features where others may be capturing essentially the same information.

6.5 How heterogeneous are learned MWEs?

On **test**, the final model (with automatic POS tags) predicts 365 MWE instances (31 are gappy; 23 are weak). There are 298 unique MWE types.

POS pattern	#	examples (lowercased lemmas)
NOUN NOUN	53	*customer service*, *oil change*
VERB PREP	36	*work with*, *deal with*, *yell at*
PROPN PROPN	29	*eagle transmission*, *comfort zone*
ADJ NOUN	21	*major award*, *top notch*, *mental health*
VERB PART	20	*move out*, *end up*, *pick up*, *pass up*
VERB ADV	17	*come back*, *come in*, *come by*, *stay away*
PREP NOUN	12	*on time*, *in fact*, *in cash*, *for instance*
VERB NOUN	10	*take care*, *make money*, *give crap*
VERB PRON	10	*thank you*, *get it*
PREP PREP	8	*out of*, *due to*, *out ta*, *in between*
ADV ADV	6	*no matter*, *up front*, *at all*, *early on*
DET NOUN	6	*a lot*, *a little*, *a bit*, *a deal*
VERB DET NOUN	6	*answer the phone*, *take a chance*
NOUN PREP	5	*kind of*, *care for*, *tip on*, *answer to*

Table 4: Top predicted POS patterns and frequencies.

Organizing the predicted MWEs by their coarse POS sequence reveals that the model is not too prejudiced in the kinds of expressions it recognizes: the 298 types fall under 89 unique POS+strength patterns. Table 4 shows the 14 POS sequences predicted 5 or more times as strong MWEs. Some of the examples (*major award*, *a deal*, *tip on*) are false positives, but most are correct. Singleton patterns include PROPN VERB (*god forbid*), PREP DET (*at that*), ADJ PRON (*worth it*), and PREP VERB PREP (*to die for*).

True positive MWEs mostly consist of (a) named entities, and (b) lexical idioms seen in training and/or listed in one of the lexicons. Occasionally the system correctly guesses an unseen and OOV idiom based on features such as hyphenation (*walk - in*) and capitalization/OOV words (*Chili Relleno*, *BIG MISTAKE*). On **test**, 244 gold MWE types were unseen in training; the system found 93 true positives (where the type was predicted at least once), 109 false positives, and 151 false negatives—an unseen type recall rate of 38%. Removing types that occurred in lexicons leaves 35 true positives, 61 false positives, and 111 false negatives—a unseen and OOV type recall rate of 24%.

6.6 What kinds of mismatches occur?

Inspection of the output turns up false positives due to ambiguity (e.g., *Spongy and **sweet bread***); false negatives (*top to bottom*); and overlap (***get high** quality service*, gold *get **high quality** service*; ***live up** to*, gold ***live up to***). A number of the mismatches turn

scheme	$\lvert\mathcal{Y}\rvert$	ρ	$\overline{M}$	$\overline{\lvert\mathbf{w}\rvert}$	$\bar{\text{P}}$	$\bar{\text{R}}$	$\overline{F_1}$
no gaps, 1-level	3	100	2.1	733k	73.33	55.72	63.20
no gaps, 2-level	4	150	3.3	977k	72.60	59.11	65.09
gappy, 1-level	6	200	1.6	1,466k	66.48	61.26	63.65
gappy, 2-level	8	100	3.5	1,954k	73.27	60.44	66.15

Table 5: Training with different tagging schemes. Results are cross-validation averages on **train**. All schemes are evaluated against the full gold standard (8 tags).

out to be problems with the gold standard, like *having our water **shut off*** (gold *having our **water shut off***). This suggests that even noisy automatic taggers might help identify annotation inconsistencies and errors for manual correction.

6.7 Are gappiness and the strength distinction learned in practice?

Three quarters of MWEs are strong and contain no gaps. To see whether our model is actually sensitive to the phenomena of gappiness and strength, we train on data simplified to remove one or both distinctions—as in the first 3 labelings in figure 2—and evaluate against the full 8-tag scheme. For the model with the recall cost, clusters, and oracle POS tags, we evaluate each of these simplifications of the training data in table 5. The gold standard for evaluation remains the same across all conditions.

If the model was unable to recover gappy expressions or the strong/weak distinction, we would expect it to do no better when trained with the full tagset than with the simplified tagset. However, there is some loss in performance as the tagset for learning is simplified, which suggests that gappiness and strength are being learned to an extent.

7 Related Work

Our annotated corpus (Schneider et al., 2014) joins several resources that indicate certain varieties of MWEs: lexicons such as WordNet (Fellbaum, 1998), SAID (Kuiper et al., 2003), and WikiMwe (Hartmann et al., 2012); targeted lists (Baldwin, 2005, 2008; Cook et al., 2008; Tu and Roth, 2011, 2012); websites like Wiktionary and Phrases.net; and large-scale corpora such as SemCor (Miller et al., 1993), the French Treebank (Abeillé et al., 2003), the Szeged-ParalellFX corpus (Vincze, 2012), and the Prague Czech-English Dependency Treebank (Čmejrek et al., 2005). The difference is that Schneider et al. (2014) pursued *a comprehensive annotation approach* rather than targeting specific varieties of MWEs or relying on a preexisting lexical resource. The annotations are *shallow*, not relying explicitly on syntax (though in principle they could be mapped onto the parses in the Web Treebank).

In terms of modeling, the use of machine learning classification (Hashimoto and Kawahara, 2008; Shigeto et al., 2013) and specifically BIO sequence tagging (Diab and Bhutada, 2009; Constant and Sigogne, 2011; Constant et al., 2012; Vincze et al., 2013) for contextual recognition of MWEs is not new. Lexical semantic classification tasks like named entity recognition (e.g., Ratinov and Roth, 2009), supersense tagging (Ciaramita and Altun, 2006; Paaß and Reichartz, 2009), and index term identification (Newman et al., 2012) also involve chunking of certain MWEs. But our discriminative models, facilitated by the new corpus, *broaden the scope of the MWE identification task* to include many varieties of MWEs at once, including explicit marking of gaps and a strength distinction. By contrast, the aforementioned identification systems, as well as some MWE-enhanced syntactic parsers (e.g., Green et al., 2012), have been restricted to contiguous MWEs. However, Green et al. (2011) allow gaps to be described as constituents in a syntax tree. Gimpel and Smith's (2011) shallow, gappy language model allows arbitrary token groupings within a sentence, whereas our model imposes projectivity and nesting constraints (§3). Blunsom and Baldwin (2006) present a sequence model for HPSG supertagging, and evaluate performance on discontinuous MWEs, though the sequence model treats the non-adjacent component supertags like other labels—it cannot enforce that they mutually require one another, as we do via the gappy tagging scheme (§3.1). The lexicon lookup procedures of Bejček et al. (2013) can match gappy MWEs, but are nonstatistical and extremely error-prone when tuned for high oracle recall.

Another major thread of research has pursued *unsupervised* discovery of multiword types from raw corpora, such as with statistical association measures (Church et al., 1991; Pecina, 2010; Ramisch et al., 2012, *inter alia)*, parallel corpora (Melamed, 1997; Moirón and Tiedemann, 2006; Tsvetkov and Wintner, 2010), or a combination thereof (Tsvetkov and

Wintner, 2011); this may be followed by a lookup-and-classify approach to contextual identification (Ramisch et al., 2010). Though preliminary experiments with our models did not show benefit to incorporating such automatically constructed lexicons, we hope these two perspectives can be brought together in future work.

8 Conclusion

This article has presented the first supervised model for identifying heterogeneous multiword expressions in English text. Our feature-rich discriminative sequence tagger performs shallow chunking with a novel scheme that allows for MWEs containing gaps, and includes a strength distinction to separate highly idiomatic expressions from collocations. It is trained and evaluated on a corpus of English web reviews that are comprehensively annotated for multiword expressions. Beyond the training data, its features incorporate evidence from external resources—several lexicons as well as unsupervised word clusters; we show experimentally that this statistical approach is far superior to identifying MWEs by heuristic lexicon lookup alone. Future extensions might integrate additional features (e.g., exploiting statistical association measures computed over large corpora), enhance the lexical representation (e.g., by adding semantic tags), improve the expressiveness of the model (e.g., with higher-order features and inference), or integrate the model with other tasks (such as parsing and translation).

Our data and open source software are released at `http://www.ark.cs.cmu.edu/LexSem/`.

Acknowledgments

This research was supported in part by NSF CAREER grant IIS-1054319, Google through the Reading is Believing project at CMU, and DARPA grant FA8750-12-2-0342 funded under the DEFT program. We are grateful to Kevin Knight, Martha Palmer, Claire Bonial, Lori Levin, Ed Hovy, Tim Baldwin, Omri Abend, members of JHU CLSP, the NLP group at Berkeley, and the Noah's ARK group at CMU, and anonymous reviewers for valuable feedback.

A Basic Features

All are conjoined with the current label, y_i.

Label Features

1. previous label (the only first-order feature)

Token Features

Original token

2. $i = \{1, 2\}$
3. $i = |\mathbf{w}| - \{0, 1\}$
4. capitalized $\wedge$ $[\![i = 0]\!]$
5. word shape

Lowercased token

6. prefix: $[w_i]_1^k \big|_{k=1}^4$
7. suffix: $[w_i]_j^{|w|} \big|_{j=|w|-3}^{|w|}$
8. has digit
9. has non-alphanumeric c
10. context word: $w_j \big|_{j=i-2}^{i+2}$
11. context word bigram: $\mathbf{w}_j^{j+1} \big|_{j=i-2}^{i+1}$

Lemma Features

12. lemma + context lemma if one of them is a verb and the other is a noun, verb, adjective, adverb, preposition, or particle: $\lambda_i \wedge \lambda_j \big|_{j=i-2}^{i+2}$

Part-of-speech Features

13. context POS: $pos_j \big|_{j=i-2}^{i+2}$
14. context POS bigram: $\mathbf{pos}_j^{j+1} \big|_{j=i-2}^{i+1}$
15. word + context POS: $w_i \wedge pos_{i\pm1}$
16. context word + POS: $w_{i\pm1} \wedge pos_i$

Lexicon Features (unlexicalized)

WordNet only

17. OOV: λ_i is not in WordNet as a unigram lemma $\wedge$ pos_i
18. compound: non-punctuation lemma λ_i and the {previous, next} lemma in the sentence (if it is non-punctuation; an intervening hyphen is allowed) form an entry in WordNet, possibly separated by a hyphen or space
19. compound-hyphen: pos_i = `HYPH` $\wedge$ previous and next tokens form an entry in WordNet, possibly separated by a hyphen or space
20. ambiguity class: if content word unigram λ_i is in WordNet, the set of POS categories it can belong to; else pos_i if not a content POS $\wedge$ the POS of the longest MW match to which λ_i belongs (if any) $\wedge$ the position in that match (B or I)

For each multiword lexicon

21. lexicon name $\wedge$ status of token i in the shortest path segmentation (O, B, or I) $\wedge$ subcategory of lexical entry whose match includes token i, if matched $\wedge$ whether the match is gappy
22. the above $\wedge$ POS tags of the first and last matched tokens in the expression

Over all multiword lexicons

23. at least k lexicons contain a match that includes this token (if $n \geq 1$ matches, n active features)
24. at least k lexicons contain a match that includes this token, starts with a given POS, and ends with a given POS

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé and Nancy Ide, editors, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 165–187. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Timothy Baldwin. 2005. Looking for prepositional verbs in corpus data. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 115–126. Colchester, UK.

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proc. of MWE*, pages 1–2. Marrakech, Morocco.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA.

Eduard Bejček, Pavel Straňák, and Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proc. of the 9th Workshop on Multiword Expressions*, pages 106–115. Atlanta, Georgia, USA.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170. Chiang Mai, Thailand.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012a. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012b. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, California, USA.

Phil Blunsom and Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proc. of EMNLP*, pages 164–171. Sydney, Australia.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of NAACL-HLT*, pages 242–245. Los Angeles, California, USA.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical acquisition: exploiting on-line resources to build a lexicon*, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602. Sydney, Australia.

Michael Collins. 2002. Discriminative training methods for Hidden Markov Models: theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8. Philadelphia, Pennsylvania, USA.

Matthieu Constant and Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 49–56. Portland, Oregon, USA.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proc. of ACL*, pages 204–212. Jeju Island, Korea.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens dataset. In *Proc. of MWE*, pages 19–22. Marrakech, Morocco.

Hal Daumé, III. 2006. *Practical structured learning techniques for natural language processing*. Ph.D. dissertation, University of Southern California, Los Angeles, California, USA. URL `http://hal3.name/docs/daume06thesis.pdf`.

Mona Diab and Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proc. of MWE*, pages 17–22. Suntec, Singapore.

Nick C. Ellis, Rita Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, 42(3):375–396.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Massachusetts, USA.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of 'let alone'. *Language*, 64(3):501–538.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Mahmoud Ghoneim and Mona Diab. 2013. Multiword expressions in the context of statistical machine trans-

lation. In *Proc. of IJCNLP*, pages 1181–1187. Nagoya, Japan.

Kevin Gimpel and Noah A. Smith. 2011. Generative models of monolingual and bilingual gappy patterns. In *Proc. of WMT*, pages 512–522. Edinburgh, Scotland, UK.

Adele E. Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago, Illinois, USA.

Adele E. Goldberg. 2006. *Constructions at work: the nature of generalization in language*. Oxford University Press, Oxford, UK.

Edouard Grave, Guillaume Obozinski, and Francis Bach. 2013. Hidden Markov tree models for semantic class induction. In *Proc. of CoNLL*, pages 94–103. Sofia, Bulgaria.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: a parsing tour de force with French. In *Proc. of EMNLP*, pages 725–735. Edinburgh, Scotland, UK.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2012. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Prague Czech-English Dependency Treebank 2.0. Technical Report LDC2012T08, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA. URL `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T10`.

Silvana Hartmann, György Szarvas, and Iryna Gurevych. 2012. Mining multiword terms from Wikipedia. In Maria Teresa Pazienza and Armando Stellato, editors, *Semi-Automatic Ontology Development*. IGI Global, Hershey, Pennsylvania, USA.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proc. of EMNLP*, pages 992–1001. Honolulu, Hawaii, USA.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proc. of ACL-08: HLT*, pages 595–603. Columbus, Ohio.

Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. 2003. SAID. Technical Report LDC2003T10, Linguistic Data Consortium, Philadelphia, Pennsylvania, USA. URL `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T10`.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, pages 282–289.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. URL `http://people.csail.mit.edu/pliang/papers/meng-thesis.pdf`.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of EMNLP*, pages 97–108. Providence, Rhode Island, USA.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proc. of HLT*, pages 303–308. Plainsboro, New Jersey, USA.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proc. of HLT-NAACL*, pages 337–342. Boston, Massachusetts, USA.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proc. of EACL*, pages 162–173. Avignon, France.

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proc. of the EACL 2006 Workshop on Multi-word Expressions in a Multilingual Context*, pages 33–40. Trento, Italy.

Rosamund Moon. 1998. *Fixed expressions and idioms in English: a corpus-based approach*. Oxford Studies in Lexicography and Lexicology. Clarendon Press, Oxford, UK.

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proc. of COLING 2012*, pages 2077–2092. Mumbai, India.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL-HLT*, pages 380–390. Atlanta, Georgia, USA.

Gerhard Paaß and Frank Reichartz. 2009. Exploiting

semantic constraints for estimating supersenses with CRFs. In *Proc. of the Ninth SIAM International Conference on Data Mining*, pages 485–496. Sparks, Nevada, USA.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158.

Carlos Ramisch. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Ph.D. dissertation, University of Grenoble and Federal University of Rio Grande do Sul, Grenoble, France. URL `http://www.inf.ufrgs.br/~ceramisch/download_files/thesis-getalp.pdf`.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proc. of ACL 2012 Student Research Workshop*, pages 1–6. Jeju Island, Korea.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In *Proc. of LREC*, pages 662–669. Valletta, Malta.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proc. of the Third ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge, Massachusetts, USA.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proc. of CoNLL*, pages 147–155. Boulder, Colorado, USA.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proc. of EMNLP*, pages 1524–1534. Edinburgh, Scotland, UK.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: a pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 189–206. Springer, Berlin, Germany.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proc. of LREC*. Reykjavík, Iceland.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proc. of the 9th Workshop on Multiword Expressions*, pages 139–144. Atlanta, Georgia, USA.

James W. Thatcher. 1967. Characterizing derivation trees of context-free grammars through a generalization of finite automata theory. *Journal of Computer and System Sciences*, 1(4):317–322.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*, pages 173–180. Edmonton, Alberta, Canada.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264. Beijing, China.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proc. of EMNLP*, pages 836–845. Edinburgh, Scotland, UK.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: contextual or statistical. In *Proc. of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39. Portland, Oregon, USA.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In *Proc. of *SEM*, pages 65–69. Montréal, Quebec, Canada.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. of ACL*, pages 384–394. Uppsala, Sweden.

Martin Čmejrek, Jan Cuřín, Jan Hajič, and Jiří Havelka. 2005. Prague Czech-English Dependency Treebank: resource for structure-based MT. In *Proc. of EAMT*, pages 73–78. Budapest, Hungary.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of MUC-6*, pages 45–52. Columbia, Maryland, USA.

Veronika Vincze. 2012. Light verb constructions in the SzegedParalellFX English-Hungarian parallel corpus. In *Proc. of LREC*. Istanbul, Turkey.

Veronika Vincze, István Nagy T., and János Zsibrita. 2013. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing*, 10(2):6:1–6:25.