Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an n-gram Lattice

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

We present a new model for acquiring comprehensive multiword lexicons from large corpora based on competition among n-gram candidates. In contrast to the standard approach of simple ranking by association measure, in our model n-grams are arranged in a lattice structure based on subsumption and overlap relationships, with nodes inhibiting other nodes in their vicinity when they are selected as a lexical item. We show how the configuration of such a lattice can be optimized tractably, and demonstrate using annotations of sampled n-grams that our method consistently outperforms alternatives by at least 0.05 F-score across several corpora and languages.

1 Introduction

Despite over 25 years of research in computational linguistics aimed at acquiring multiword lexicons using corpora statistics, and growing evidence that speakers process language primarily in terms of memorized sequences (Wray, 2008), the individual word nonetheless stubbornly remains the *de facto* standard processing unit for most research in modern NLP. The potential of multiword knowledge to improve both the automatic processing of language as well as offer new understanding of human acquisition and usage of language is the primary motivator of this work. We present an effective, expandable, and above all tractable new approach to comprehensive multiword lexicon acquisition that aims to find a middle ground between standard MWE acquisition approaches based on association measures (Ramisch, 2014), and more sophisticated statistical models (Newman et al., 2012) which fail to scale to the large corpora which are the main sources of the distributional information in modern NLP systems.

A central challenge in building comprehensive multiword lexicons is pairing down the huge space of possibilities without imposing restrictions which disregard a major portion of the multiword vocabulary of a language: allowing for diversity creates significant redundancy among statistically promising candidates. The lattice model proposed here addresses this primarily by having the candidates—contiguous and non-contiguous n-gram types—compete with each other based on subsumption and overlap relations to be selected as the best (i.e., most parsimonious) explanation for statistical irregularities due to lexical affinity. We test this approach across four large corpora in three languages, including two relatively free-word-order languages (Croatian and Japanese), and find that this approach consistency outperforms alternatives, offering scalability and many avenues for future enhancement.

2 Background and Related Work

In this paper we will refer to the targets of our lexicon creation efforts as **formulaic sequences**, following the terminology of Wray (2002; 2008), wherein a formulaic sequence (FS) is defined as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." That is, a FS shows signs of being part of a mental lexi-

con. Though by this definition individuals or small groups may have their own FS, here we are only interested in FS that are shared by a recognizable language community.

In computational linguistics, the most common term used to describe multiword lexical units is *multiword expression* ("MWE": Sag et al. (2002), Baldwin and Kim (2010)), but here we wish to make a principled distinction between at least somewhat non-compositional, strongly lexicalized MWEs and FS, a near superset which includes many MWEs but also compositional linguistic formulas. This distinction is not a new one; it exists, for example, in the original paper of Sag et al. (2002) in the distinction between lexicalized and institutionalized phrases, and also to some extent in the MWE annotation of Schneider et al. (2014b), who distinguish between weak (collocational)[1] and strong (non-compositional) MWEs. It is our contention, however, that separate, precise terminology is useful for research targeted at either class: we need not strain the concept of MWE to include items which do not require special semantics, nor are we inclined to disregard the larger formulaticity of language simply because it is not the dominant focus of MWE research. Many MWE researchers might defensibly balk at including in their MWE lexicons and corpus annotations (English) FS such as *there is something going on*, *it is more important than ever to ...*, *...do not know what it is like to ...*, *there is no shortage of...*, *the rise and fall of...*, *now is not the time to...*, etc. as well as tens of thousands of other such phrases which, along with less compositional MWEs like *be worth ...'s weight in gold*, fall under the FS umbrella. Another reason to introduce a different terminology is that there are classes of what is typically considered MWE that do not fit well into an FS framework, for instance novel compound nouns whose semantics are accessible by analogy (e.g., *glass limb* or *government ambiguity*). Also, we exclude from the definition of both FS and MWE those named entities which refer to people or places which are little-known and/or whose surface form appears derived (e.g., *Mrs. Barbara W. Smith* or *Smith Garden Supplies Ltd*). Figure 1 shows the conception of the relationship between FS, MWE, and (multiword) named entities that we assume for this paper.

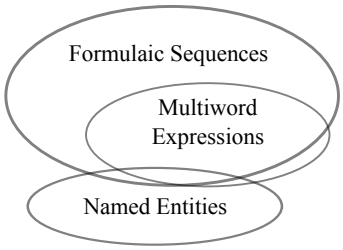

Figure 1: Multiword Terminology

Regardless of the terminology used to describe them, the starting point for multiword lexicon creation has typically been lexical association measures (Church and Hanks, 1990; Dunning, 1993; Schone and Jurafsky, 2001; Evert, 2004; Pecina, 2010; Araujo et al., 2011; Kulkarni and Finlayson, 2011; Ramisch, 2014). When these methods are used to build a lexicon, particular binary syntactic patterns are typically chosen. Only some of these measures generalize tractably beyond two words, for example PMI (Church and Hanks, 1990), i.e., the log ratio of the joint probability to the product of the marginal probabilities of the individual words. Other measures specifically designed to address sequences of larger than two words include: the c-value (Frantzi et al., 2000), a metric designed for term extraction which weights term frequency by the log length of the n-gram while penalizing n-grams that appear in frequent larger ones; and mutual expectation (Dias et al., 1999), which produces a normalized statistic that reflects how much a candidate phrase resists the omission of any particular word. Similarly, the lexical predictability ratio (LPR) of Brooke et al. (2015) is an association measure intended for any possible syntactic pattern which is calculated by discounting syntactic predictability from the overall conditional probability for each word given the other words in the phrase. Though most association measures involve only usage statistics of the phrase and its subparts, the DRUID measure is an exception which uses distributional semantics around the phrase to

[1]Here we avoid the term *collocation* entirely due to confusion with respect to its interpretation. Though some define it similarly to our definition of FS, it can be applied to any words that show a statistical tendency to appear in the vicinity of one another for any reason: for instance, the pair of words *doctor/nurse* might be considered a collocation (Ramisch, 2014).

identify how easily an n-gram could be replaced by a single word (Riedl and Biemann, 2015).

Typically multiword lexicons are created by ranking n-grams according to an association measure and applying a threshold. The algorithm of da Silva and Lopes (1999) is somewhat more sophisticated, in that it identifies the local maxima of association measures across subsuming n-grams within a sentence to identify MWEs of unrestricted length and syntactic composition; its effectiveness beyond noun phrases, however, seems relatively limited (Ramisch et al., 2012). Brooke et al. (2014; 2015) developed a heuristic method intended for general FS extraction in larger corpora, first using conditional probably statistics to do an initial (single pass) coarse-grained segmentation of the corpus, followed by a pass through the resulting vocabulary, breaking larger units into smaller ones based on a tradeoff between marginal and conditional statistics. Beyond association measures, other general unsupervised approaches to the multiword unit identification include that of Newman et al. (2012), who used a generative Dirichlet Process model which jointly creates a linear segmentation of the corpus and a multiword vocabulary of keyphrases.

Other research in MWEs has tended to be rather focused on particular syntactic patterns such as verb-noun combinations (Fazly et al., 2009). The work of Schneider et al. (2014a) is a rare example of a comprehensive token-level MWE identification system which distinguishes a full range of MWE sequences in the English Web Treebank, including those involving gaps, using a supervised sequence tagging model. Schneider et al. make use of existing manual lexical resources and note that an (unsupervised) automatic lexical resource could be useful addition to the model, although attempts to do so have achieved mixed success (Riedl and Biemann, 2016).

The motivation for building lexicons of FS naturally overlaps with those for MWE: models of distributional semantics, in particular, can benefit from sensitivity to multiword units (Cohen and Widdows, 2009), as can parsing (Constant and Nivre, 2016) and topic models (Lau et al., 2013). One major motivation for looking beyond MWEs is the ability to carry out broader linguistic analyses. Within corpus linguistics, multiword sequences have been studied in the form of *lexical bundles* (Biber et al., 2004), which are simply n-grams that occur above a certain frequency threshold. Like FS, Lexical bundles generally involve larger phrasal chunks that would be missed by traditional MWE extraction, and so research in this area has tended to focus on how particular formulaic phrases (e.g., *if you look at*) are indicative of particular genres (e.g., university lectures). Lexical bundles have been applied, in particular, to learner language: for example, Chen and Baker (2010) show that non-native student writers use a severely restricted range of lexical bundle types, and tend to overuse those types, while Granger and Bestgen (2014) investigate the role of proficiency, demonstrating that intermediate learners underuse lower-frequency bigrams and overuse high-frequency bigrams relative to advanced learners. Sakaguchi et al. (2016) demonstrate that improving fluency (closely linked to the use of linguistic formulas) is more important than improving strict grammaticality with respect to native speaker judgments of non-native productions; Brooke et al. (2015) explicitly argue for FS lexicons as a way to identify, track, and improve learner proficiency.

3 Method

Our approach to FS identification involves optimization of the total explanatory power of a lattice, where each node corresponds to an n-gram type. The explanatory power of the whole lattice is defined simply as a product of the **explainedness** of the individual nodes. Each node can be considered either "on" (is an FS) or "off" (is not an FS). The basis of the calculation of explainedness is the syntax-sensitive LPR association measure of Brooke et al. (2015), but it is calculated differently depending on the on/off status of the node as well as the status of the nodes in the vicinity: nodes are linked based on n-gram subsumption and corpus overlap relationships (see Figure 2), with "on" nodes typically explaining other nodes. Given these relationships, we iterate over the nodes and greedily optimize the on/off choice relative to explainedness in the local neighborhood of each node, until convergence.

3.1 Collecting statistics

The first step in the process is to derive a set of n-grams and related statistics from a large, unlabeled

corpus of text. Since our primary association measure is an adaption of LPR, our approach in this section mostly follows Brooke et al. (2015) up until the last stage. An initial requirement of any such method is an n-gram frequency threshold, which we set to 1 instance per 10 million words, following Brooke et al. (2015).[2]

We include gapped or non-contiguous n-grams in our analysis, in acknowledgment of the fact that many languages have MWEs where the components can be "separated", including verb particle constructions in English (Dehé, 2002), and noun–verb idioms in Japanese (Hashimoto and Kawahara, 2008). Having said this, there are generally strong syntactic and length (Wasow, 2002) restrictions on what can constitute a gap, which we capture in the form of a language-specific POS-based regular expression (see Section 4 for details). This greatly lowers the number of potentially gapped n-gram types, increasing precision and efficiency for negligible loss of recall. We also exclude punctuation and lemmatize the corpus, and enforce an n-gram count threshold. As long as the count threshold is substantially above 1, efficient extraction of all n-grams can be done iteratively: in iteration i, i-grams are filtered by the frequency threshold, and then pairs of instances of these i-grams with $(i-1)$ words of overlap are found, which derives a set of $(i+1)$-grams which necessarily includes all those over the frequency threshold.

Once a set of relevant n-grams is identified and counted, other statistics required to calculate the **Lexical Predictability Ratio** ("LPR") for each word in the n-gram are collected. LPR is a measure of how predictable a word is in a lexical context, as compared to how predictable it is given only syntactic context (over the same span of words). Formally, the LPR for word w_i in the context of a word sequence $w_1, ..., w_i, ..., w_n$ with POS tag sequence $t_1, ..., t_n$, is given by:

$$\text{LPR}(w_i, w_{1,n}) = \max_{1 \le j < k \le n} \frac{p(w_i|w_{j,k})}{p(w_i|t_{j,k})}$$

[2]Based on manual analysis using the MWE corpus of Schneider et al. (2014b), this achieves very good (over 90%) type-level MWE coverage using the frequency filtered n-gram statistics from the ICWSM blog corpus (see Section 4) after filtering out proper names.

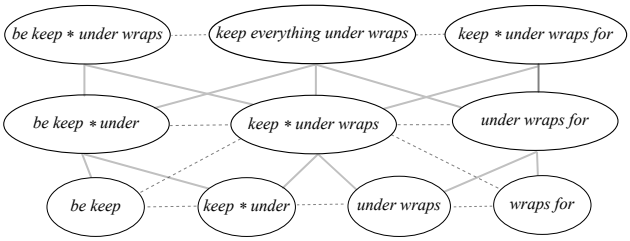

Figure 2: A portion of an n-gram lattice. Solid lines indicate subsumption, dotted lines overlaps

where $w_{j,k}$ denotes the word sequence $w_j, ..., w_{i-1}, w_{i+1}, ..., w_k$ excluding w_i (similarly for $t_{j,k}$). Note that the lower bound of LPR is 1, since the ratio for a word with no context is trivially 1. We use the same equation for gapped n-grams, with the caveat that quantities involving sequences which include the location where the gap occurs are derived from special gapped n-gram statistics.

In the segmentation approach of Brooke et al. (2015), LPR for an entire span is calculated as a product of the individual LPRs, but here we will use the minimum LPR across the words in the sequence:

$$\text{minLPR}(w_{1,n}) = \min_{1 \le i \le n} \text{LPR}(w_i, w_{1,n})$$

Here, minLPR for a particular n-gram does not reflect the *overall* degree to which it holds together, but rather focuses on the word which is its weakest link. For example, in the case of *be keep ∗ under*, a general statistical metric might assign it a high score due to the strong association between *keep* and *under*, but minLPR is focused on the weaker relationship between *be* and *keep ∗ under*.

3.2 Node interactions

The n-gram nodes in the lattice are directionally connected to nodes consisting of $(n+1)$-grams which subsume them and $(n-1)$-grams which they subsume. For example, as detailed in Figure 2, the (gapped) n-gram *keep ∗ under wraps* would be connected "upwards" to the node *keep everything under wraps* and connected "downwards" to *under wraps*. These directional relationships allow for two basic interactions between nodes in the lattice when a node is turned on: **covering**, which inhibits nodes below (subsumed by) a turned-on node (e.g., if *keep ∗ under wraps* is on, the model will tend not to

choose *under wraps* as an FS); and **clearing**, which inhibits nodes above a turned-on node (e.g., if *keep* $*$ *under wraps* is on, the model would avoid selecting *keep everything under wraps* as an FS). A third, undirected mechanism is **overlapping**, where nodes inhibit each other due to overlaps in the corpus (e.g., having both *keep* $*$ *under wraps* and *be keep* $*$ *under* as FS will be avoided).

3.2.1 Covering

The most important node interaction is **covering**, which corresponds to discounting or entirely excluding a node due to a node higher in the lattice. Our model includes two types of covering: hard and soft.

Hard covering is based on the idea that, due to very similar counts, we can reasonably conclude that the presence of an n-gram in our statistics is a direct result of the other. In Figure 2, e.g., if we have 143 counts of *keep* $*$ *under wraps* and 152 counts of *under wraps*, the presence of *keep* $*$ *under wraps* almost completely explains *under wraps*, and we should consider these two n-grams as one. We do this by permanently disabling any hard covered node, and setting the minLPR of the covering node to the maximum minLPR among all the nodes it covers (including itself); this means that longer n-grams with function words (which often have lower minLPR) can benefit from the strong statistical relationships between open-class lexical features in n-grams that they cover. This is done as a preprocessing step, and greatly improves the tractability of the iterative optimization of the lattice. Of course, a threshold for hard covering must be chosen: empirically, we have found that a ratio of $2/3$ (corresponding to a significant majority of the counts of a lower node corresponding to the higher node) works well. We also use the concept of hard covering to address the issue of pronouns, based on the observation that pronouns often have high LPR values (Brooke et al., 2015). In the lattice, n-grams with pronouns are considered covered (inactive) unless they cover at least one other node which does not have a pronoun, which allows us to limit FS with pronouns without excluding them entirely: they are included only in cases where they are definitively formulaic.

Soft covering is used in cases when a single n-gram does not entirely account for another, but a turned-on n-gram to some extent may explain some of the statistical irregularity of one lower in the lattice. For instance, in Figure 2 *keep* $*$ *under* is not hard-covered by *keep* $*$ *under wraps* (since there are FS such as *keep* $*$ *under surveillance*, *keep it under your hat*, etc.), but if *keep* $*$ *under wraps* is tagged as an FS, we nevertheless want to discount the portion of the *keep* $*$ *under* counts that correspond to occurrences of *keep* $*$ *under wraps*. This is accomplished by increasing the turned-off explainedness of *keep* $*$ *under* (and thus making turning on less desirable) in the following manner: let $c(\cdot)$ be the count function, and $s_1 ... s_m$ be any turned-on nodes which are above t in the lattice (covering nodes). Then, the cover(t) score for a covered node t is:

$$\text{cover}(t) = \max\left(0, \frac{c(t) - \sum_{i=1}^{m} c(s_i)}{c(t)}\right)$$

This is intended as a simple, quick-to-calculate approximation of the result of recalculating minLPR with the counts corresponding to the covering nodes actually removed. cover takes on values in the range 0 to 1, with 1 being the default when no covering is occurring.

3.2.2 Clearing

In general, covering prefers turning on longer, covering n-grams since doing so explains nodes lower in the lattice. Not surprisingly, it is generally desirable to have a mechanism working in opposition, i.e., one which views shorter FS as helping to explain the presence of longer n-grams which contain them, beyond the FS-neutral syntactic explanation provided by minLPR. **Clearing** does this by changing the E_0 of nodes higher in the lattice when a lower node is turned-on. The basic mechanism is similar to covering, except that counts cannot be made use of in the same way—whereas it makes sense to explain covered nodes in proportion to the counts of their covering nodes (since the counts of the covered n-grams are from the covering n-gram), in the reverse direction this logic fails.

A simple but effective solution which avoids introducing extra hyperparameters is to make use of the minLPR values of the relevant nodes. In the most common two-node situation, we increase the explainedness of the cleared node based on the ratio of the minLPR of two nodes, though only if the minLPR of the lower node is higher. Generalized to

the (rare) case of multiple clearing nodes, we define clear(t) as:

$$\text{clear}(t) = \prod_{i=1}^{m} \min\left(1, \frac{\text{minLPR}(t)}{\text{minLPR}(u_i)}\right)$$

where u_i is the ith clearing node of t, i.e., turned-on nodes below t in the lattice. We refer to this mechanism as "clearing" because it tends to clear away a variety of trivial uses of common FS which may have higher LPR due to the lexical and syntactic specificity of the FS. For instance, in Figure 2 if the node *keep* $*$ *under wraps* is turned on and has a minLPR of 8, then, if the minLPR of a node such as *keep* $*$ *under wraps for* is 4, clear(t) will be 0.5. Like cover, clear takes on values in the range 0 to 1, with 1 being the default when no clearing occurs.

3.2.3 Overlap

The third mechanism of node interaction involves n-grams which overlap in the corpus. In general, independent FS do not consistently overlap. For example, given that *be keep* $*$ *under* and *keep* $*$ *under wraps* often appear together (overlapping on the tokens *keep* $*$ *under*), we do not want both being selected as an FS, even in the case that both have high minLPR. To address this problem, we use a mechanism somewhat different than above: rather than increasing the explainedness of turned-off nodes, we decrease the explainedness of the overlapping turned-on nodes—a penalty rather than an incentive which expresses the model's confusion at having overlapping FS. Let $oc(x, y)$ refer to number of times the n-grams corresponding to nodes x and y overlapped in the corpus, and $o_1, ..., o_m$ refer to overlapping nodes, i.e., turned-on nodes which overlap with our target node $t$in the corpus. We define overlap (t) as:

$$\text{overlap}(t) = \frac{c(t)}{c(t) - \sum_{i=1}^{m} oc(t, o_i)}$$

Overlap takes on values in the range 1 to $+\infty$, also defaulting to 1 when no overlaps exist. The effect of overlap is hyperbolic: small amounts of overlap have little effect, but nodes with significant overlap will effectively be forced to turn off.

3.3 Explainedness

In order to capture the current state of the model and provide an overall explainedness store, we introduce two indicator functions: $I_{on}(t)$ is 1 if t is on and 0 if t is off ; $I_{\mathit{off}}(t)$ is the reverse. The objective function maximized by the model is then the explainedness ($expl$) across all the nodes of the lattice, which can be defined in terms of minLPR and the node interaction functions:

$$expl(t_1, ..., t_N) = \prod_{i=1}^{N} I_{on}(t_i) C^{-\text{overlap}(t_1)} + I_{\mathit{off}}(t_i) minLPR(t_i)^{-\text{cover}(t_i)\cdot\text{clear}(t_i)} \quad (1)$$

When a node is off, its explainedness is the inverse of its minLPR, except if there are covering or clearing nodes which explain it by pushing the exponent of minLPR towards zero. When the node is on, its explainedness is the inverse of a fixed cost hyperparameter C, though this cost is increased if it overlaps with other active nodes. All else being equal, when minLPR$(t) > C$, a node will be selected as an FS, and so, independent of the node interactions, C can be viewed as the threshold for the minLPR association measure under a traditional approach to MWE identification. There is no upper bound on C, but empirically, we have found values in the range $[3, 6]$ give reasonable results for the languages presented in this paper.

3.4 Optimization

The dependence of the explainedness of nodes on their neighbors effectively prohibits a global optimization of the lattice. Fortunately, though most of the nodes in the lattice are part of a single main connected component, most of the effects of nodes on each other are relatively local, and effective local optimizations can be made tractable by applying some simple restrictions. The main optimization loop consists of iterations over the lattice until complete convergence (no changes in the final iteration). For each iteration over the main loop, each potentially active node is examined in order to evaluate whether its current status is optimal given the current state of the lattice. The order that we perform this has an effect on the result: among the obvious

options, good results are obtained through ordering nodes by frequency, which gives an implicit advantage to relatively common n-grams.

Given the relationships between nodes, it is obviously not sufficient to consider switching only the present node. If, for instance, one or more of *be keep* $*$ *under wraps*, *under wraps*, or *be keep* $*$ *under* has been turned on, the covering, blocking, or overlapping effects of these other nodes will likely prevent a competing node like *keep* $*$ *under wraps* from being correctly activated. Instead, the algorithm identifies a small set of "relevant" nodes which are the most important to the status of the node under consideration. Since turned-off nodes have no direct effect on each other, only turned-on nodes above, below, or overlapping with the current node need be considered. Once the relevant nodes have been identified, all nodes (including turned-off nodes) whose explainedness is affected by one or more of the relevant nodes are identified, and then a greedy search is carried out for the optimal configuration of the relevant nodes, starting from an 'all-on' state and at each step turning off the node (if any) which most increases overall explainedness.

In practice, we apply the following efficiency restrictions, which significantly reduce the runtime without sacrificing quality (based on development set testing):

- We limit the total number of relevant nodes to 5. When there are more than 5 nodes turned on in the vicinity of the target node, the most relevant nodes are selected by ranking candidates by the change in explainedness across possible configurations of the target and candidate node considered in isolation;
- To avoid having to deal with storing and processing trivial overlaps, we exclude overlaps with a count of less than 5 from our lattice;
- Many nodes have a minLPR which is slighlty larger than 1. There is very little chance these nodes will be activated by the algorithm, and so after applying hard covering, we do not consider activating nodes with minLPR < 2.

4 Evaluation

We evaluate our approach across three different languages including evaluation sets derived from four different corpora. In English, we follow Brooke et al. (2015) in using a 890M token filtered portion of the ICWSM blog corpus (Burton et al., 2009) tagged with the Tree Tagger (Schmid, 1995). To facilitate a comparison with Newman et al. (2012), which does not scale up to a corpus as large as the ICWSM, we also build a lexicon using the 100M token British National Corpus (Burnard, 2000), using the standard CLAWS-derived POS tags for the corpus. Lemmatization included removing all inflectional marking from both words and POS tags. For English, gaps are identified using the same POS regex used in Brooke et al. (2015), which includes simple nouns and portions thereof, up to a maximum of 4 words.

The other two languages we include in our evaluation are Croatian and Japanese. Relative to English, both languages have freer word order: we were interested in probing the challenges associated with using an n-gram approach to FS identification in such languages. For Croatian, we used the fhrWaC corpus (Šnajder et al., 2013), a filtered version of the Croatian web corpus hrWaC (Ljubešić and Klubička, 2014), which is POS-tagged and lemmatized using the tools of Agić et al. (2013). Similar to English, the POS regex for Croatian includes simple nouns, adjectives and pronouns, but also other elements that regularly appear inside FS, including both adverbs and copulas. For Japanese, we used a subset of the 100M-page web corpus of Shinzato et al. (2008), which was roughly the same size (in terms of token count) as the English corpus. We segmented and POS-tagged the corpus with MeCab (Kudo, 2008) using the UNIDIC morphological dictionary (Den et al., 2007). The POS regex for Japanese covers the same basic nominal structures as English, but also includes case markers and adverbials. Though our processing of Japanese includes basic lemmatization related to superficial elements like the choice of writing script and politeness markers, many elements (such as case marking) which are removed by lemmatization in Croatian are segmented into independent morphological units in the MeCab output, making the task somewhat different for the two languages.

Brooke et al. (2015) introduced a method for evaluating FS extraction without a reference lexicon or direct annotation of the output of a model. Instead, n-grams are sampled after applying the frequency

	Contiguous		Gapped		κ
	FS	non-FS	FS	non-FS	
ICWSM	169	702	29	916	0.84
BNC	49	403	8	475	0.84
Croatian	64	382	11	456	0.87
Japanese	124	286	36	341	0.81

Table 1: Statistics for test sets

threshold and annotated as being either an FS or not, allowing for calculation of a true F-score for any model. We use the annotation of 2000 n-grams in the ICWSM corpus from that earlier work, and applied the same annotation methodology to the other corpora discussed above: after training and based on written guidelines derived from the definitions of Wray (2008), three native-speaker, educated annotators judged 500 contiguous n-grams and another 500 gapped n-grams for each of the new corpora.

Other than the inclusion of new languages, our test sets differ from Brooke et al. (2015) in two ways. First, instead of relying on strict majority annotation, we entirely excluded n-grams which just one annotator marked as FS, improving the reliability of evaluation at the cost of some representativeness. Second, for the main evaluation we collapsed gapped and contiguous n-grams into a single test set. The rationale is that the number of positive gapped examples was too low (particularly for the BNC and the Croatian corpus) to provide a reliable independent F-score (see further discussion in Section 6). Statistics for the four test sets are given in Table 1.

Our primary comparison is with the heuristic LPR model from Brooke et al. (2015), which is scalable to large corpora and includes gapped n-grams. For the BNC, we also compare, separately, the DP-seg model from Newman et al. (2012) with recommended settings, and the LocalMaxs algorithm of da Silva and Lopes (1999) using SCP; because these other approaches only generate sequential multiword units, we use only the sequential part of the BNC test set for this evaluation. All comparison approaches have themselves been previously compared against a wide range of association measures. As such, we do not repeat all these comparisons here, but we do consider a lexicon built from ranking n-grams according to the measure used in our lattice (minLPR) as well as PMI. For each of these two association measures we build a lexicon equal to the size of the lexicon produced by our model.

We created small development sets for each corpus and used them to do a thorough testing of parameter settings. Although it is generally possible to increase precision by increasing C, we found that across corpora we were always obtained near-optimal results using a C of 4, so to demonstrate the usefulness of the lattice technique as an entirely off-the-shelf tool, we present the results using identical settings for all four corpora. We treat covering as a fundamental part of the Lattice model, but to investigate the efficacy of other node interactions within the model we present results with each of overlap and clearing node interactions turned off.

5 Results

The main results for FS acquisition across all four corpora are shown in Table 2. As noted in previous work, simple statistical association measures like PMI do fairly poorly when faced with syntactically-unrestricted n-grams of variable length: minLPR is clearly a much better statistic for this purpose. The LPRseg method of Brooke et al. (2015) consistently outperforms simple ranking, and the lattice method proposed here does better still, with a margin that is fairly consistent across languages. Generally, clearing and overlap node interactions provide a relatively large increase in precision at the cost of a smaller drop in recall, though the change is fairly symmetrical in Croatian. The Japanese and ICWSM corpus had relatively high precision and low recall, whereas both the BNC and Croatian corpus have low precision and high recall.

In the contiguous FS test set for the BNC (Table 3), we found that both the LocalMaxs algorithm and the DP-seg method of Newman (Newman et al., 2012) were able to beat our other baseline methods with roughly similar F-scores, though both are well below our Lattice method. Some of the difference seems attributable to fairly severe precision/recall imbalance, though we were unable to improve the F-score by changing the parameters from recommended settings for either model.

	English											
	ICWSM			BNC			Croatian			Japanese		
Source	P	R	F	P	R	F	P	R	F	P	R	F
PMIrank	0.24	0.15	0.18	0.12	0.25	0.16	0.21	0.32	0.26	0.23	0.09	0.15
minLPRrank	0.49	0.31	0.39	0.25	0.44	0.32	0.36	0.45	0.40	0.51	0.14	0.22
LPR-seg	0.53	0.42	0.47	0.37	0.44	0.40	0.41	0.47	0.43	0.77	0.34	0.47
Lattice $-cl$	0.57	0.42	0.49	0.33	0.58	0.42	0.39	0.56	0.46	0.74	0.39	0.51
Lattice $-ovr$	0.52	**0.51**	0.51	0.34	**0.60**	0.44	0.36	**0.67**	0.47	0.69	**0.51**	**0.58**
Lattice	**0.67**	0.43	**0.52**	**0.40**	0.60	**0.48**	**0.44**	0.56	**0.49**	**0.87**	0.39	0.53

Table 2: Results of FS identification in various test sets; PMIrank = lexicon created by ranking pointwise mutual information, minLPRrank = lexicon created by ranking by minLPR, LPRseg = lexicon created by method of Brooke at al. (2015), -cl = no clearing, -ovr = no penalization of overlaps, P = Precision, R = Recall, F = F-score. Bold is best in column.

	P	R	F
PMIrank	0.20	0.29	0.23
minLPRrank	0.34	0.45	0.39
LPR-seg	0.42	0.45	0.43
LocalMaxs	**0.56**	0.39	0.46
DP-seg	0.35	**0.71**	0.47
Lattice	0.47	0.63	**0.54**

Table 3: Results of FS identification in contiguous BNC test set; LocalMaxs = method of da Silva and Lopes (1999); DP-seg = method of Newman et al. (2012)

6 Discussion

Though the results across the four corpora are reasonably similar with respect to overall F-score, there are significant discrepancies. By using the UNIDIC morpheme representation as the base unit for Japanese, the model is doing an extra layer of FS identification, one which is provided by word boundaries in the other languages. The result is that there are many more FS for Japanese: precision is high, and recall is relatively low. Under these conditions, methods which increase precision at the cost of recall will have worse F-score, which explains why the full lattice model is not the best in that case. Importantly, the initial n-gram statistics derived from corpus reflect that Japanese is different: the number of n-gram types over length 4 is almost twice the number in the ICWSM corpus. One idea for future work is to automatically adapt to the input language/corpus in order to ensure a good balance between precision and recall.

At the opposite extreme, the low precision of the BNC is almost certainly at least somewhat a reflection of its relatively small size: whereas the n-gram threshold we have used here results in minimum counts of roughly 100 for the other three corpora used here, the BNC statistics includes n-grams with counts less than 10. This might be resolved by increasing the n-gram threshold, or simply avoiding small corpora, but for some applications it may be useful to build comprehensive FS lexicons even in relatively low resource situations. One idea we are pursing is modifying the calculation of the LPR metric to integrate uncertainty due to low counts.

There is one other more general explanation for the precision/recall imbalances we see across the four corpora: both the BNC and the Croatian corpus are composed primarily of published texts written by professional writers, particularly news. The ICWSM corpus and the Japanese corpus, on the other hand, are mostly blogs and web pages. Though all genres have FS, it is our observation that there is much more diversity in less controlled genres: many English gapped expressions, in particular, appear almost exclusively in relatively informal genres.

We were interested in addressing Croatian and Japanese in part because of their relatively free word order, and whether the potential gaps in FS would help with these languages. We discovered, however,

that free word order actually results in *more* of a tendency towards contiguous FS, not less. Strikingly rare in Croatian, in particular, are expressions where the content of a gap is an argument which must be filled to syntactically complete an expression: it is English whose fixed-word-order constraints often keep elements of an FS distant from each other. The gaps that do happen in Croatian are mostly prosody-driven insertions of other elements into already complete FS. This phenomena highlights a problem with the current model, in that gapped and contiguous versions of the same n-gram sequence (e.g., *take away* and *take* $*$ *away*) are, at present, considered entirely independently. Alternatives for dealing with this include collapsing statistics to create a single node in the lattice, creating a promoting link between contiguous and gapped versions of the same n-grams sequence in the lattice model, or switching to a dependency representation (which, we note, requires very little change to the basic model presented here, but would narrow its applicability).

As it stands, connections in the lattice are based entirely on explicit n-gram subsumption and overlap relations. One of the major benefits of this model as compared to alternatives is the relative ease with which additional kinds of interactions between nodes could be implemented. Another connection we have considered is based on identical or similar syntactic patterns (e.g., POS sequences), which could serve to encourage the model to make generalizations: in English, e.g., learning that verb-particle combinations are generally likely to be FS, whereas verb-determiner combinations are not. Our initial investigations suggest, however, it may be difficult to apply this idea without merely amplifying existing undesirable biases in the LPR measure. Bringing in other information such as simple distributional statistics might help the model identify non-compositional semantics, and could, in combination with the existing lattice competition, focus the model on detecting MWEs, which could in turn serve as a reliable basis for generalization when the raw LPR statistics alone are an unreliable indicator of formulaticity.

With respect to speed, although the optimization of the lattice is several orders of magnitude more time consuming than the decomposition heuristics of Brooke et al. (2015), the time needed to build and optimize the lattice was still only a fraction of the time required to collect the statistics necessary to calculate LPR, a necessary first step for both methods: therefore, the overall effect of adding the lattice optimization is fairly minimal. We note that for all four corpora, the lattice optimization algorithm converged within 10 iterations.

Finally, though in general using the sampling methodology to build test sets has significant benefits in terms of providing a robust, replicable evaluation, F-scores become unreliable when there are very few positive examples, such as with most of our gapped n-gram test sets. To do a reliable, focused evaluation of FS with gaps, we would need many more examples than can be produced with straightforward sampling. One option we might consider is adding additional filtering to the sampling processes, for instance using the LPR metric to exclude large numbers of unpromising n-grams.

7 Conclusion

We have presented here a new methodology for acquiring comprehensive multiword lexicons from large corpora, using competition in an n-gram lattice. Our evaluation using annotations of sampled n-grams shows that it is consistently outperforms alternatives across several corpora and languages. A tool which implements the method, as well as the acquired lexicons and test sets, are available at: `http://ANON.YMO.US`.[3]

References

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57.

Vitor De Araujo, Carlos Ramisch, and Aline Villavicencio. 2011. Fast and flexible MWE candidate generation with the mwetoolkit. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 134–136, Portland, USA.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing,*

[3] Removed for anonymity

Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, USA.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. *If you look at. . .*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25:371–405.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*, pages 753–761, Dublin, Ireland.

Julian Brooke, Adam Hammond, David Jacob, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2015. Building a lexicon of formulaic language for language learners. In *Proceedings of the NAACL 15 Workshop on Multiword Expressions*, pages pp. 96–104, Denver, USA.

Lou Burnard. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM '09)*, San Jose, USA.

Yu-Hua Chen and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2):30–49.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*, pages 161–171, Berlin, Germany.

Joaquim da Silva and Gabriel Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, Orlando, USA.

Nicole Dehé. 2002. *Particle Verbs in English: Syntax, Information, Structure and Intonation*. John Benjamins, Amsterdam, Netherlands/Philadelphia, USA.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.

Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of Conférence Traitement Automatique des Langues Naturelles (TALN) 1999*, Cargèse, France.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert. 2004. *The statistics of word cooccurrences–word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3:115–130.

Sylviane Granger and Yves Bestgen. 2014. The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52:229–252.

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 *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 992–1001, Honolulu, USA.

Taku Kudo. 2008. MeCab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`.

Nidhi Kulkarni and Mark Finlayson. 2011. jMWE: A Java toolkit for detecting multi-word expressions. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pages 122–124, Portland, USA.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On collocations and topic models. *ACM Transactions on Speech and Language Processing*, 10(3):10:1–10:14.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden.

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of the 24th International Conference*

on Computational Linguistics (COLING '12), pages 2077–2092, Mumbai, India.

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

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

Carlos Ramisch. 2014. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer, Dordrecht, Netherlands.

Martin Riedl and Chris Biemann. 2015. A single word is not enough: Ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2430–2440, Lisbon, Portugal.

Martin Riedl and Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111, Berlin, Germany.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '02)*, pages 1–15, Mexico City, Mexico.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Assocation for Computational Linguistics*, 4:169–182.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50, Dublin, Ireland.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 455–461, Reykjavík, Iceland.

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP '01)*, pages 100–108, Pittsburgh, USA.

Keiji Shinzato, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. 2008. A large-scale web data collection as a natural language processing infrastructure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2236–2241, Marrakech, Morocco.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Sofia, Bulgaria.

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford, USA.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge, UK.

Alison Wray. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press, Oxford, UK.