Submission # 1137
Action Editor: Noah Smith

We would like to thank all of our reviewers for their detailed comments. We have tried to address the important points as best we can, but were limited to 3 extra pages (beyond the 10 pages of the original submission). As all the reviews had numerous suggestions for improvement, with only one obvious area of overlap, we have had to pick and choose what we could add. In cases where we did not change the paper, we provide a justification below.

We should mention one major change that wasn't prompted by individual reviewer comments: there were clear misinterpretations by our original Japanese annotators
which led to overannotation of FS, leading us to do another round of annotation since the original submission. The Japanese results have therefore changed, and are in fact more consistent with the other languages.

REVIEWER A

> In the introduction, the proposed approach is claimed to be "effective,
> expandable, and tractable", however, it is not clear how each of these
> claims is evaluated in the paper (effective, expandable and tractable). The
> approach is also claimed to have the candidates compete for "the best (most
> parsimonious) explanation for statistical irregularities due to lexical
> affinity", but concepts like parsimony and lexical affinity need to be
> defined (what are they, how candidates display them, and how they are
> measured, ...).

Our intent is that effectiveness is demonstrated by our main evaluation. Tractability is demonstrated implicitly by the fact that we apply our method to corpora much larger than many of our competitors can handle. We have added more details about this to the paper, though we do not want to belabour the point, since it is not the primary purpose. Our claim of expandability relates to our discussion of future work, however we have added a bit more discussion about exactly how expansions would be added without major changes to the model.

> The paper should explicitly state what the hypotheses are in relation to the
> utility functions, particularly in relation to the approach by Brooke et al.
> (2015), which is simpler than the proposed approach. Moreover, the
> state-of-the-art has several simpler formulations for the identification of
> general formulaic sequences of flexible sizes (e.g. Silva et al. 1999 for
> contiguous & non-contiguous cases and Villavicencio et al. 2007 for
> contiguous cases), and the text should justify why this particular approach
> is appealing, and advantageous in relation to the others. The addition of an
> error analysis would help highligh these points, helping to clarify the
> advantages of the particular approach especially in relation to Brooke et
> al.

We have added a new paragraph contrasting our method to other algorithms, and a qualitative discussion which doubles as an error analysis.

Note that we looked at Villavicencio et al. 2007, but it is very limited in scope (they only deal with trigrams) and doesn't provide any workable comparison in the context of unrestricted FS extraction beyond ranking with simple association measures, which, as we mention, have already been shown to be inferior to the more sophisticated models we are comparing against here.

> The paper should also discuss the motivation for each of the node

> interaction functions for rewarding/penalizing competitive candidates in the
> lattice, including how they are expected to contribute to the overall
> performance of the approach and if the authors considered alternative
> functions or why these in particular, which linguistic intuition they wanted
> to capture, why define E0 in terms of the minLPR and E1 of a cost parameter,
> why they are introduced in the explainedness as exponents (why not using a
> linear dependence, for example. A suggestion for helping the presentation is
> for the authors to rewrite E0 and E1 into a single formula that combines the
> three measures in a utility function, as it would be easier to explain (how
> E0 and E1 compete, under which circumstances each of them wins),
> particularly with the addition of an example. In addition, when presentating
> these functions, one problem is that although d0 and d1 are used to define
> E0 and E1 in section 3.2, they are only defined in later sections (sections
> 3.3.2 and 3.3.3). As a result the discussion appears too abstract as the
> functions have not been completely defined yet. There should be some
> explanation for their motivation/intuition, as it's not clear what they are
> trying to capture, how E0 and E1 would vary for FS and non-FS, etc. My
> suggestion is to first introduce and motivate the ideas behind them, and
> later define them together. For d1, there is also a difference in the number
> of arguments of "oc" in the text (oc(x,y)) and in the definition of d1
> (oc(oi)). d0 is defined as a uniform combination of clear and cover, but it
> would be helpful to have a discussion of the effects of different weights
> for each of them for the different languages.

We have substantially reorganized the presentation of the model, and are now providing an explicit objective function, at the same time getting rid of the confusing forward pointers and many of the extra symbols.

We hope that the general purpose of each interaction in terms of the way that it inhibits other nodes is clear. We have added a bit more detail about the choice of the clearing function to both that section as well as the discussion.

> In terms of the methodology, a discussion is needed about the motivation for
> adopting Brooke et al.'s lexical predictability ratio (LPR), and for the
> modifications proposed, such as the use of minLPR instead of the product of
> the LPRs and why the focus is on the weakest link (was there any comparison
> between the use of minLPR and the product of LPRs, is it done for efficiency
> reasons, etc). Moreover, the construction of the lattice needs to be
> explained, including how the nodes relate to one another. For instance, in
> figure 2, the solid lines indicate subsumption, but "keep everything under
> wraps" is linked with "be keep * under" via a subsumption relation, even
> though the latter contains elements that are not in the former (shouldn't it
> be an overlap link?).

We have added a bit more to the discussions of LPR, to be more explicit why LPR is a good metric for our lattice. We've also fixed the incorrect relations you mention, which should hopefully resolve the confusion.

> In terms of the evaluation, the distinction between contiguous and gapped
> test sets, which are later merged into a single test set per language when
> presenting the results, is not helpful. Instead, maintaining the distinction
> would help clarify the advantages of the proposed approach in each language,
> particularly in relation to the error analysis.

As we note in the paper, adding the distinction between continguous and gapped simply isn't defensible given the lack of positive gapped examples across some of the datasets (since both precision and recall rely on primarily on true positives). We would very much like to present this distinction, but the scores depend primarily on a handful of examples for some of datasets and the results are erratic, they would confuse rather than elucidate.

> Additionally, for the contiguous n-grams in these languages a comparison
> with LocalMaxs could be done.

Actually that's not possible. LocalMaxs (like DP-seg) does not scale, due to the fact that it needs to calculate counts for every n-gram in the corpus; it is quicker than DP-Seg, but the memory usage is immense for large corpora. We have added a mention of this fact in the paper.

> Moreover, the authors need to add the statistical significance of the
> results in tables 2 and 3 as the approaches often differ only in the second
> decimal place.

We have added some statistical significance information.

> In Table 2, it would be helpful to have also the results for only the cover
> function (Lattice-cl-ovr), to determine how much the other functions
> contribute.

Thank you for the suggestion. This has been added to the revised version.

> A figure with the top/bottom candidates proposed by the approach be helpful.

There are no explicit top/bottom candidates in our model, as it's not a ranking model and it would be incorrect to present it as such. N-grams could be ranked by the explainedness of their node, of course, but that doesn't properly reflect their overall influence in the model. We do have a new qualitative analysis with random samples from our lexicon.

> Add the list in an appendix for English and explicitly explain the lists for
> Croatian and Japanese. How many expressions and how many cases they cover.

We have added some samples for each language and some general statistics for the lexicons of each language. As should be now clear, the lists of expressions generated by the model are far larger than can be included in a paper. We will be separately releasing all the lexicons.

> The paper also needs to add a list of existing MWE lexica or lexical
> resources with MWEs available to the community (e.g. those listed in the MWE
> community website, or available in WordNet or other resources).

We have to demurr on this. Our paper is not about MWEs specifically, we do not use MWE resources to build our model nor to evaluate it, so a focus on available MWE resources seems a significant detour, especially given the space constraints. It would also be very difficult to do the topic justice, especially given our cross-linguistic interest.

> Stylistically the text would also benefit from a revision that splits
> sentences that are too long (there are several throughout the text). For
> instance, the last sentence in the first paragraph of the introduction is 9
> lines long: 'We present an effective, expand- able, and above all tractable
> new approach to com- prehensive multiword lexicon acquisition that aims to
> find a middle ground between standard MWE ac- quisition 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.').

Thank you for your suggestion. We have split up some of the longer sentences, hopefully improving the readability.

> Moreover, some passages are copied from Brooke et al. (2015), but lack the
> citation, such as:
> *"Other mea- sures 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." (page 2) and
> *"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" (page 3)
> There are also other instances, and these need to be rephrased.

We have rephrased the relevant parts of the literature review, though the basic content is unchanged.

> For the annotations, the text mentions 3 annotators, but doesn't specify
> whether they were different for each of the 3 languages.

Yes, they were different for each of the 3 languages. Since they were native (first language) speakers of each of the relevant languages, we believe this implies that they were different across languages.

> When comparing the languages, the paper also mentions that "free word order
> actually results in more of a tendency towards contiguous FS, not less.",
> but it is unclear where this is shown.

We've tried to explain this a bit better.

> * “accessible by analogy (e.g., glass limb or government ambiguity)” by
> analogy with what? Specify and add the original expression

We have made the analogy explicit.

> * “define the explainedness of a node in terms of two functions” Which two
> functions?

Hopefully this is clearer in the new formulation.

> * In section 3.2 can the authors explain how two valid but similar FSs would
> be treated (e.g. keep * under wraps/surveillance)?

The only kind of interaction between these nodes involves their shared n-gram subsequence (keep * under), as described in the example in the paper. We have expanded the example a bit.

> * "We also use the concept of hard covering to ad- dress the issue of
> pronouns, based on the observation that pronouns often have high LPR
> values". This sentence is too vague. Explain using an example.

We've added an example.

> * How was the 2/3 threshold for hard covering determined? Why also have soft
> covering?

Using performance in the (English) development set, we have made that clearer. Our intention is that the "keep * under wraps/surveillance" example shows why soft covering is also important, we have added a bit more explanation.

> * "has been turned on, the covering, blocking, or overlap- ping effects of
> these other nodes" --> blocking shoud be clearing

Yes, thanks.

> * For the efficiency restrictions explain if other values have been tested
> for them. How did they affect the overall performance? A table detailing the
> behavior with different values would be helpful.

We've added a bit more detail, though perhaps not as much as you're asking for (due to space limitations).

> * "which was roughly the same size (in terms of token count) as the English
> corpus" specify which English corpus

The ICWSM. We've added that.

> * In Table 1 the values are less than 2000 (for ICWSM) and 1000 for the
> other corpora. Add also the initial numbers for each type of FS/non-FS.

As we mention in the text, to improve the reliability of our evaluation we drop instances where just one annotator tagged as FS. This obviously results in test sets that are less than the original sample size. This should now be clearer.

REVIEWER B

> Assuming that the solution is indeed scalable and empirically viable -- I am
> wondering if the same idea could be used in the context of other tasks
> (morphological composition in morphological lattices, preferred expressions
> in lattices of proposed machine translations, etc). It would be nice if the
> author could comment on the applicability of the technical solution in other
> domains (if there is any).

This is interesting, and so we've added a discussion of the morphology idea to the paper (The short answer is yes).

> In the case of syntax, much research in theoretical linguistics and in
> language technology has proposed the idea that idiomatic expressions (of
> which MWEs are a special case) are the rule, rather than the exception, in
> NL grammar. Among the theoretical proposal one can find versions of
> Construction Grammar (Goldberg 1996) where language knowledge is composed of
> complete constructions, rather than primitive units (words) and in
> technological proposals one can find the idea of Data-Oriented Parsing

> (Remko Scha, Rens Bod), wherein each syntactic subtree (with potential slots
> for substitution) is an FS-like element that can potentially be used and
> reused for analysis and generation. And so: while the relation of the
> proposed FS to existing theories of strongly lexicalized fixed expressions
> on the one end is clear, it is unclear what is the relation larger idiomatic
> syntactic constructions with lexicalized or unlexicalized slots on the other
> end -- and this relation should be made clear, for the work to be properly
> situated.

Note that we are not ourselves proposing FS, we are borrowing the theory behind FS from the work of Wray (2002;2008), who includes an extremely thorough discussion of its relationship to other linguistic theories (Chapter 7 of the 2008 book in particular), including Construction Grammar. We have added a bit more detail to our introduction of FS to make it clear that, yes, most "constructions" are FS, and we have added constructions to Figure 1. However, ultimately this is a technical paper, and we are limited in the space we can dedicate to theoretical concerns.

> The case of morphology is far less clear, however it is equally important,
> and it has to do with the basic question of what is the unit that constructs
> the FS (or the lattices themselves in empirical terms). Are these words?
> morphemes? lemmas? POS? inflections? In the case of words, what is the
> relation of inflected sequences forms / inflected idioms (kicked his bucket
> vs kicked her bucket, for instance) -- are they the same? related?
> overalapping? subsumed?. As it stands, the discussion of morphology comes
> out only later as an afterthought -- where the construction of the lattice
> in Japanese (in the evaluation section) involves some morphological
> segmentation whereas the lattice construction for the other languages
> doesn't. It is thus unclear if the algorithm finds sequences that are of the
> same type, or are comparable for that matter, across those languages. The
> selected FS may be incomparable in terms of effective length (morpheme
> sequences tend to be longer), the status of function words vs. inflection in
> FS, and so on. The authors should put forth clearly what is the status of
> morphology inside their general theory (and in terms of their regular
> expressions), and then later revisit the empirical ramifications of their
> decision -- for instance what happens when moving between typologically
> different languages.

We have added a bit more to both the introduction of FS as well as a paragraph in the discussion that addresses this point. Ultimately, in all three languages we are tokenizing and lemmatizing using standards for each language, and building our FS from the result. However, this is not a theoretical choice, but a practical one; it is undeniable that FSes often involve morphological components that we are overlooking. With Japanese, due to the lack of morpheme boundaries, morphological segmentation is a more standard form of tokenization, and so we have taken advantage of this to bring Japanese somewhat closer to our theoretical ideal. Given that many of the case markers of Japanese correspond to prepositions in European languages (or the possesive marker in English), the actual difference in generated FS is fairly minor, except for the extra overhead of word segmentation in Japanese that is provided by word boundaries in English (in Wray's theory, most words would be FS by default).

> The author discuss at length the method, the construction of the lattice,
> and possible relation between arcs (covering clearing and overlapping) and
> their related scores. The objective function however and the respective
> optimisation algorithm are discussed informally in passing. I would much
> prefer a more precise formalisation of the objective function and greedy
> algorithm, possibly with a running example of several states of progress of
> the algorithm over a sample lattice. This would also help to clarify what

> "locality properties" are at play here, which justify a greedy solution that
> is locally optimised. As a more general note, I think it should be made
> clear if the target of the lattice construction is a single lattice for the
> entire text, optimized at once, or multiple lattices, reflecting different
> aspects of the vocabulary, optimized separately. As a minor note, there are
> two many forward references in the text, IMO, which hinder
> understandability. Example "where d1 and d2 are ... as detailed in the next
> section". I would prefer the details closer to where they are relevant.

Thank you for your suggestions. We have reorganized the expanation of the method to remove forward references, and have provided an explict objective function and algorithm. We decided, though, that a full example of the optimization algorithm involved too much overhead with only modest explanatory benefits.

By "text", we assume you mean corpus. It is hopefully clear that the algorithm works at the level of all n-grams for the corpus, and not texts within the corpus (this is what makes the model practical relative to token-level models like DP-seg); the nodes of the lattice are n-gram types, not tokens. As it happens, most nodes are connected to most other nodes by some pathway in the lattice, but there are nodes (mostly individual bigrams) which do not, and thus are not part of the main lattice.

> It is unclear to me why comparing the collected lexicon using the proposed
> method to a sample set of ngrams that respond to a certain frequency
> threashold is a sound comparison-- isnt the frequency threashold just a
> simpler and less sophisticated baseline for MWE extraction? It appears that
> you use this method not as a baseline, but as an actual method to extract
> gold standard for comparison with your method. I understand that others have
> used such evaluation methods before -- but I would need much more convincing
> in order to accept the claim of efficacy. If I misunderstood, and you do not
> use it as a gold for comparison, or you use a different gold, please
> clarify.

We are not entirely sure we are intepreting your concern correctly. It is certainly not the case that we are using high-frequency n-grams directly as a gold standard (i.e. we assume that high frequency n-grams are FS simply by virtue of their high frequency): this should be clear given the discussion of the annotation process, where we taken the n-grams and annotated them as being FS or not. To make this absolutely clear, we have included frequency as a baseline association score. It is poor, but obviously would be unbeatable if our gold standard was built based simply on frequency.

A more subtle issue is that using frequency-filtered n-grams as the sampling pool for our FS gold standard means that there are FS in the corpus that cannot possibilty be considered in the evaluation because they do not appear sufficiently. This is true, and unavoidable in this evaluation setup. As discussed in Brooke et al. 2015, sampling from frequent n-grams in the corpus brings with it a number of benefits: in particular, the test set reflects the variety of FS that appear throughout the entire corpus (many of which would not appear in any existing lexicon or in a small sample of texts), annotators can look at variety of examples to make an judgement, and (most importantly) it is possible to calculate a proper F-score. The trade-off, however, is that there needs to be some initial filtering of the sample so that annotation process is not "finding a needle in the haystack" as it would be if you considered all n-grams in the corpus. As we discuss in 3.1, we investigated coverage in an existing English MWE annotated corpus using n-grams from the ICWSM after applying frequency cutoffs and found that type coverage is quite good (token coverage is even better, of course); that is, we are not losing very many FS with our filtering, at least not ones that occur across corpora of the similar genre (both are social media corpora). With a large corpus and an appropriately low frequency cutoff, we believe that annotations of

frequency filtered n-gram samples is a reasonable way to solve what is without a doubt a very challenging evaluation problem. We have added a slightly longer justification of the evaluation method to the paper.

REVIEWER C:

> Where I had trouble was with the details of the model. First of all, much
> ofit seemed to be ad hoc; even accepting the notion of LPR from prior work,
> there are many heuristic (or at least not obviously natural) choices made to
> operationalize the general lattice idea mathematically. Second, even if the
> model is reasonable, it is not entirely clear how the different parts fit
> together into an optimization problem. There is mention of a parameter C;
> and there is a notion of nodes being activated or not (I don’t think
> notation is introduced for these binary variables, though it would be
> appropriate). Are these the only things that have to be tuned? Is the
> calculation of explainedness deterministic based on these? There is a greedy
> search to optimize explainedness of the nodes in the lattice, subject to
> various heuristics. Presenting a formal algorithm for the optimization may
> clarify things considerably.

We have reorganized the method section and added an explicit objective function, which should help with some of the problems you identified. In terms of the central optimization problem, the only variables are exactly the on/off status of each node, and, if we exclude the parameters used in the creation of the lattice (e.g. the n-gram frequency threshold) and those involved in the efficiency restrictions, the only (hyper)parameter in the optimization is C; we have a included psuedocode for optimization algorithm. With regards to the ad hoc parts of the model, we have added a discussion separating the core aspects of the model from some of the more ad hoc choices made in the particular instantions of component functions (such as clearing), making it clear that we believe these can potentially be improved on (both in terms of performance as well as mathematical elegance), and that our primary purpose here is to show that overall architecture of the model is a promising one.

> I think it would also be worth discussing how the lattice model relates back
> to the cognitive/psychological literature. Would you predict (or are you
> aware of any evidence) that node interactions such as covering, clearing,
> and overlapping have an analogue in human language processing? Do any of the
> design decisions intentionally sacrifice cognitive plausibility for
> engineering reasons?

In general, we did not consider cognitive plausibility when creating the model, and we are certainly making no claims as such. For starters, our model makes use of more data than a human being could possibly read in a lifetime, and it is clear humans identify FS with far less evidence than our model requires. With space, we could discuss our model in the context of the more cognitive aspects of the theory of formulaic sequences as proposed by Wray, but given that this wasn't a focus in the development of the model, we think it is really too much of a digression to justify.

> - After the equation for LPR, it would be good to give the intuition:
> choosing a context span such that w_i is more likely given its context words
> than its context POS tags.

This was removed from an earlier draft due to space, and we've added back in.

> - Inter-annotator agreement is surprisingly high. Could you elaborate on
> Wray’s criteria for FS that were applied by annotators?

Your surprise about the Kappa being high is based on a misunderstanding: the kappa number reported in the table is after filtering to remove borderline cases (i.e. it is the kappa for the final test set, not the original sampling). To avoid confusion, we've added the original kappa score which is quite a bit lower. Your intution is correct that there is significant subjectivity to FS identification (more so than MWE annotation), and as such we would never rely on a single annotator for building a FS test set. We have not added more information about Wray's criteria; though it provide a starting point for annotation, more relevant to achieving reasonable agreement are the specific annotation guidelines which we created. Presenting them in the paper would be too much of a digression, but we promise to make them public.

> - I would like to see qualitative analysis of the sequences that were
> identified as formulaic (or not) by annotators/the method. E.g.: How many of
> them are syntactic constituents? Are there a wide variety of grammatical
> kinds of expressions? Are high-frequency as well as low-frequency sequences
> detected? What is the distribution in terms of length? What kinds of gapped
> sequences are there? It would be nice to see some examples.

In the discussion, we have added a thorough analysis of specific examples and some statistics, including most of the information you ask for (implicitly if not explicitly).

> - Related to the previous point: While noting the distinction between FS and
> MWEs, is it perhaps worth measuring recall against MWE datasets? Most MWEs
> apart from proper names should count as FS, right?

Do you mean as a way of evaluating different systems built on the same corpus, or just to get a sense of coverage more generally? One central problem with using recall for evaluation is that in order to make it fair, there must be some control of the size of the lexicon. Neither our system nor our best performing competitors are ranking models, none of them offer an easy or fair way to simply take the best N words and look at recall. And even if they did, there's other kinds of biases that would come into play: for instance, PMI and similar metrics do well with rare terms that correspond well to the MWE lexicon of WordNet, for instance, which (at the type level) is dominated by certain kinds of terminology (biomedical terms, particularly) which are not really used outside a very specific community. This relates to a more fundamental problem with the idea of using recall in external lexicons as any kind of useful measure: we could not expect any particular corpus (even a very large one) to do justice to the full range of FS/MWE for a language, since most of which appear regularly only within specific genres. For example, we could not expect to build a good lexicon of English verb-particle constructions from a biomedical corpus, because many are too informal to be appropriate for that genre (e.g. "mess up"). With the qualitative analysis we have added, the genre-specificity of the FS we extract is hopefully more clear.

Though it perhaps does not directly address your comment (since it doesn't involve an external MWE lexicon), we were nonetheless inspired by your comments to include an analysis of the recall of MWE vs. non-MWE FS in the context of our existing ICWSM test set, and a small investigation of the recall (also in the ICWSM) of extremely common MWE from another MWE-annotated social media corpus.

> Typographical issues:

<snip>

We have addressed these typos, thanks.