Original Decision Letter for Submission 1561

As  TACL action editor for submission 1561, "Tabula nearly rasa: Probing the  
linguistic knowledge of character-level neural language models trained on  
unsegmented text",   I am writing to tell you that I am not accepting your  
paper in its current form, but due to its current strengths and potential, I  
encourage you to revise and submit it within 3-6 months.  
  
You can find the detailed reviews below. As you will see, the  
recommendations were mixed, with an (a), a (b), and a (c), although reviewer  
A, who recommended a (b), later revised this to (c) in a private discussion.  
Despite the differences in the reviews, all of the reviewers were very  
enthusiastic about the research direction and excited about some of the  
individual experiments presented in a paper. (I am too!) The main objections  
are that the many small insights in the individual experiments do not add up  
to a concrete claim about what these models learn, and they definitely are  
not strong enough on their own to hold up the broad claims that frame the  
paper, which encompass language acquisition, multilinguality, phonology,  
morphology, syntax, and semantics! See especially reviewer C's comments,  
which suggest that the paper may actually be clearer with less material,  
more precisely described; reviewer A's comments, which suggest that the  
paper should tone down its claims and make them more concrete; and the paper  
itself, which acknowledges that "our results are preliminary in many ways"  
(line 967).  
  
In light of its strengths, and considering that the objections are largely  
presentational, I considered giving this paper a (b), but that would require  
me to give you a specific prescription to make the paper publication-ready.  
In this case, my prescription is simply to present a concrete claim that is  
carefully supported by a coherent set of experiments. But this prescription  
is in fact vague: many different subsets of these results could be framed in  
different ways, possibly requiring different additional sets of supporting  
experiments. I don't feel it's my place to make that choice for you, so I've  
given you a (c). But for what it's worth, I suspect that making this paper  
TACL-worthy would require an amount of work on the short side of the 3-6  
month period suggested for a (c) review. TACL would be very happy to  
reconsider a revised version that presents a more focused story. The second  
half of section 2 cites many good examples of papers in this mold.  
  
If you do choose to revise and resubmit, please make use a \*new\* submission  
number, and follow the instructions in section "Revision and Resubmission  
Policy for TACL Submissions" at  
[https://transacl.org/ojs/index](_blank).  
I am allowing you one to two additional pages in the revised version for  
addressing the referees' concerns.  
  
Please understand that while we have endeavored to provide some guidance on  
how to revise the manuscript, we have NOT provided a complete list of  
modifications that guarantee acceptance; this is the distinguishing  
characteristic between the decision we have given your submission --- (c),  
rejection, but with encouragement to resubmit --- and the next higher level  
of evaluation, which is conditional acceptance ("(b)", in TACL terminology).  
 The paper will be \*\*reviewed afresh\*\* should you choose to resubmit  
(possibly involving a change of action editor and reviewers), with \*\*no  
guarantee of acceptance\*\*, even if you make all the changes suggested.    
  
Again, just to prevent misunderstandings, we repeat: \*\*making all the  
changes suggested here does not guarantee subsequent acceptance\*\*.  A  
resubmission is treated as a new submission, and the subsequent review may  
identify different problems with the paper.  
  
Please also note that if you do choose to revise and resubmit, TACL policy  
is, generally, to try not to give a (c) resubmission another (c), but  
rather, if the second revision does not meet the acceptance bar, to impose a  
rejection with a 1-year moratorium on resubmission.  Thus, please be very  
thorough in revising any resubmission.  
  
Thank you for considering TACL for your work, and, although you should take  
careful note of the caveats above, I do encourage you to revise and resubmit  
within the specified timeframe.  
  
  
Adam Lopez  
University of Edinburgh  
[alopez@inf.ed.ac.uk](mailto:alopez@inf.ed.ac.uk)  
------------------------------  
------------------------------  
....THE REVIEWS....  
------------------------------  
------------------------------  
Reviewer A:  
  
CLARITY: For the reasonably well-prepared reader, is it clear what was done  
and why? Is the paper well-written and well-structured?:   
        3. Mostly understandable to me (a qualified reviewer) with some effort.  
  
  
INNOVATIVENESS: How original is the approach? Does this paper break new  
ground in topic, methodology, or content? How exciting and innovative is the  
research it describes?  
  
Note that a paper can score high for innovativeness even if its impact will  
be limited.  
:   
        4. Creative: An intriguing problem, technique, or approach that is  
substantially different from previous research.  
  
SOUNDNESS/CORRECTNESS: First, is the technical approach sound and  
well-chosen? Second, can one trust the claims of the paper -- are they  
supported by proper experiments and are the results of the experiments  
correctly interpreted?:   
        3. Fairly reasonable work. The approach is not bad, and at least the main  
claims are probably correct, but I am not entirely ready to accept them  
(based on the material in the paper).  
  
  
RELATED WORK: Does the submission make clear where the presented system sits  
with respect to existing literature? Are the references adequate?  
  
Note that the existing literature includes preprints, but in the case of  
preprints:  
• Authors should be informed of but not penalized for missing very recent  
and/or not widely known work.  
• If a refereed version exists, authors should cite it in addition to or  
instead of the preprint.  
:   
        3. Bibliography and comparison are somewhat helpful, but it could be hard  
for a reader to determine exactly how this work relates to previous work or  
what its benefits and limitations are.  
  
  
SUBSTANCE: Does this paper have enough substance (in terms of the amount of  
work), or would it benefit from more ideas or analysis?  
  
Note that papers or preprints appearing less than three months before a  
paper is submitted to TACL are considered contemporaneous with the  
submission. This relieves authors from the obligation to make detailed  
comparisons that require additional experiments and/or in-depth analysis,  
although authors should still cite and discuss contemporaneous work to the  
degree feasible.  
:   
        4. Represents an appropriate amount of work for a publication in this  
journal. (most submissions)  
  
IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the  
ideas are novel, will they also be useful or inspirational? If the results  
are sound, are they also important? Does the paper bring new insights into  
the nature of the problem?:   
        4. Some of the ideas or results will substantially help other people's  
ongoing research.  
  
REPLICABILITY: Will members of the ACL community be able to reproduce or  
verify the results in this paper?:   
        3. They could reproduce the results with some difficulty. The settings of  
parameters are underspecified or subjectively determined, and/or the  
training/evaluation data are not widely available.  
  
IMPACT OF PROMISED SOFTWARE:  If the authors state (in anonymous fashion)  
that their software will be available, what is the expected impact of the  
software package?:   
        2. Documentary: The new software will be useful to study or replicate the  
reported research, although for other purposes it may have limited interest  
or limited usability. (Still a positive rating)  
  
IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)  
that datasets will be released, how valuable will they be to others?:   
        2. Documentary: The new datasets will be useful to study or replicate the  
reported research, although for other purposes they may have limited  
interest or limited usability. (Still a positive rating)  
  
  
TACL-WORTHY AS IS? In answering, think over all your scores above. If a  
paper has some weaknesses, but you really got a lot out of it, feel free to  
recommend it. If a paper is solid but you could live without it, let us know  
that you're ambivalent.  
  
Reviewers: after you save this review form, you'll have to make a  
confidential recommendation to the editors via pull-down menu as to: what  
degree of revision would be needed to make the submission eventually  
TACL-worthy?  
:   
        3. Ambivalent: OK but does not seem up to the standards of TACL.  
  
  
Detailed Comments for the Authors  
  
Reviewers, please draft your comments on your own filesystem and then copy  
the results into the text-entry box.  You will thus have a saved copy in  
case of system glitches.  
:   
        This paper aims to explore what RNNs trained in a language modeling task  
are learning about linguistic structure by testing them on a range of  
probing tasks related to phonology, morphology, syntax and semantics in  
English, German and Italian. I think these are very interesting questions to  
be asking, and the methodology is for the most part rigorous. I think the  
study is worthwhile, but I think the authors need to be far more cautious in  
the claims they are making about what these models learn. It would be more  
beneficial to reflect on how these tasks \*begin\* to inform us about what  
kinds of linguistic structure language-model trained neural nets can  
"learn".  
  
  
Major concerns:  
  
(1) The choice of languages should be motivated up front. Why English,  
German and Italian, which are all closely related? Why only three?   
  
(2) The very first evaluation ("Discovering phonological classes") is oddly  
imprecise and impressionistic. Why should the reader take the authors' word  
for it that "it definitely suggests that the CNLM has discovered a fair deal  
about the features organizing the phonological system of the language." This  
should be replaced with something quantitative or at least more objective,  
or dropped.  
  
(3) The authors claim to be testing whether the CNLM develops an implicit  
notion of words, but the testing methodology involves a supervised training  
step. The paper needs to be much clearer about how this is actually testing  
whether the unsupervised system has an implicit notion of "word". (Similar  
remarks hold for the morphology tests.)  
  
(4) The results of the pluralization study seem quite equivocal. In  
particular, the fact that the Umlaut plurals aren't properly modeled  
suggests that it's \*not\* picking up on an abstract notion of "plural". The  
paper doesn't seem to acknowledge this sufficiently, either here or  
especially in the conclusion.  
  
(5) That "case subcategorization" is represented by testing exactly one  
preposition in one language seems very narrow. Also, unlike German verbs  
which can be separated from their objects, P-NP sequences are not likely to  
be broken up, so this seems like something pretty surfacy/sequential and not  
really convincingly "syntax".  
  
(6) The conclusion seems to over-claim compared to what the paper is  
actually showing. Most egregiously, I don't think that the sentence  
completion task establishes knowledge of "basic semantics". The syntactic  
agreement phenomena results are also somewhat equivocal (see detailed  
comments below) and the word units results rely on a supervised training  
step.  
  
  
More detailed comments:  
  
Sec 2: How does this related work inform the questions you are asking? (The  
literature review reads as 'defensive', i.e. trying to prove that the work  
in the paper is novel, rather than situating the work with respect to  
existing literature.)  
  
Sec 2: This paper may also be relevant:  
Ettinger et al 2018 `Assessing Composition in Sentence Vector  
Representations'  
[https://aclanthology.coli.uni-](https://aclanthology.coli.uni-saarland.de/papers/C18-1152/c18-1152)   
  
ln 209 It's not clear to me what "in a localist fashion" means.  
  
ln 240 Does "We used LSTM cells for WordNLMs" mean something different from  
"We only tested a word-level LSTM and not a word-level RNN"? If so, what?  
Also, why not do the word-level RNN?  
  
ln 325 "The LSTM assigns higher probability to the acceptable bi-grams in  
all but two cases." Are the ratios of "~1" being counted as "higher"? Why?  
Similarly the caption to Table 2 says "Values > 1 in bold", but "~1" is in  
bold (in two places).  
  
ln 385 What would be the linguistic basis for wider contexts helping with  
phoneme classes? (Long-distance phonological phenomena are relatively rare,  
and none---things like vowel harmony--immediately come to mind for the  
languages tested.)  
  
ln 417 Why 20 characters? Isn't that way longer than most words, even in  
German?  
  
ln 475 If you're working from phonological properties, why would fixed  
expressions turn up? Is there any reason to believe that in their  
orthographic form the internal word boundaries of fixed expressions are less  
like other word boundaries?  
  
ln 516 What was the training set used for the Berkeley Parser to be able to  
parse German?   
  
ln 546 "unambiguously tagged in the corpus":  I think it would be useful to  
remind the reader here that these aren't gold tags but come from TreeTagger  
(right?)  
  
Table 5 I don't understand what the last two lines are. Is WordNLM\_subs.  
without OOV and WordNLM the full test set? If so, then ln 578 "the  
word-based model fares better" doesn't seem to make any sense---WordNLM  
scores \*lowest\*.  
  
ln 582 "We study German as it possesses nominal classes that form plural  
through different morphological processes" This is also true in Italian!   
  
ln 589 Both of the cites given for "German UD treebank" seem to be about the  
UD project in general. Surely there's a specific citation for the German UD  
treebank that should be included to give those researchers credit for their  
work.  
  
ln 661 "To avoid phrase segmentation ambiguities, we present phrases  
surrounded by full stops." I'm not sure what this means. What is the system  
presented with at test time? Just a phrase like in (1) (with only one  
article)? Why would not having full stops (before and after??) lead to  
ambiguity?  
  
ln 744 "as these often reflect lemmatiziation problems": Are these problems  
with TreeTagger, your system, or something else?  
  
ln 750 When would German ever have discontinuous NPs?  
  
ln 752 Is it well established that RNNs & LSTMs have the same probabilistic  
bias for shorter sequences that e.g. HMMs do?  
  
ln 774-776 I found this too terse. What is the n-gram count model? Why omit  
the sentence environment?  
  
ln 778 What stimuli not including the preposition? Where are these  
described?  
  
4.4.2 If the words occur in the corpus, they presumably occur with their  
article, so it's not immediately clear to me that the stimuli don't occur in  
the corpus. Perhaps the unattested n-grams are the adj+N combination?  
  
ln 835 What does "strong semantic anomaly" mean and how is it checked for?  
  
ln 890 Threshold for what? (I couldn't quickly figure out what the 500  
occurrence were \*of\*, nor what to compare to "above").  
  
ln 919ff I'm extremely skeptical of the claims about the sentence completion  
task. In particular, no language model has information about "syntax,  
lexical semantics, world knowledge, and pragmatics" beyond what can be  
characterized in purely distributional terms --- i.e. what words share what  
kind of distributional similarity with what other words. That will be a  
partial reflection of part of speech (syntax-ish) and lexical semantics, but  
it is no way "world knowledge". Furthermore, models don't "realize" anything  
let alone "that [friend and mistress] are human beings".  
  
ln 965 "somewhat deeper linguistic templates" seems like an overclaim.   
  
ln 990 Why didn't you include polysynthetic and agglutinative languages in  
your testing? There are pretty good resources available for Inuktitut and  
Turkish, respectively, for example.  
  
ln 991 "the common view that": This should come with citations. Places to  
look are work on Construction Grammar (authors such as Chuck Fillmore and  
Paul Kay) and also work by Ray Jackendoff.  
  
  
  
Typos/stylistic points:  
  
ln 13-14 recently reached -> has recently reached  
ln 096 as it goes -> as it gets  
ln 149 model -> models?  
ln 431 ad hoc doesn't need a hyphen  
ln 531 can discover about -> can discover -or- can discover information  
about  
ln 622 I'm not sure what "the latter" is supposed to refer back to.  
ln 720 the Universal Dependencies -> the German UD treebank  
ln 996 capable to flexibly store -> capable of flexibly storing  
  
REVIEWER CONFIDENCE:   
        4. Quite sure. I tried to check the important points carefully. It's  
unlikely, though conceivable, that I missed something that should affect my  
ratings.  
  
------------------------------  
  
------------------------------  
Reviewer B:  
  
CLARITY: For the reasonably well-prepared reader, is it clear what was done  
and why? Is the paper well-written and well-structured?:   
        5. Very clear.  
  
  
INNOVATIVENESS: How original is the approach? Does this paper break new  
ground in topic, methodology, or content? How exciting and innovative is the  
research it describes?  
  
Note that a paper can score high for innovativeness even if its impact will  
be limited.  
:   
        3. Respectable: A nice research contribution that represents a notable  
extension of prior approaches or methodologies.  
  
SOUNDNESS/CORRECTNESS: First, is the technical approach sound and  
well-chosen? Second, can one trust the claims of the paper -- are they  
supported by proper experiments and are the results of the experiments  
correctly interpreted?:   
        4. Generally solid work, although there are some aspects of the approach or  
evaluation I am not sure about.  
  
  
RELATED WORK: Does the submission make clear where the presented system sits  
with respect to existing literature? Are the references adequate?  
  
Note that the existing literature includes preprints, but in the case of  
preprints:  
• Authors should be informed of but not penalized for missing very recent  
and/or not widely known work.  
• If a refereed version exists, authors should cite it in addition to or  
instead of the preprint.  
:   
        5. Precise and complete comparison with related work. Benefits and  
limitations are fully described and supported.  
  
  
SUBSTANCE: Does this paper have enough substance (in terms of the amount of  
work), or would it benefit from more ideas or analysis?  
  
Note that papers or preprints appearing less than three months before a  
paper is submitted to TACL are considered contemporaneous with the  
submission. This relieves authors from the obligation to make detailed  
comparisons that require additional experiments and/or in-depth analysis,  
although authors should still cite and discuss contemporaneous work to the  
degree feasible.  
:   
        4. Represents an appropriate amount of work for a publication in this  
journal. (most submissions)  
  
IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the  
ideas are novel, will they also be useful or inspirational? If the results  
are sound, are they also important? Does the paper bring new insights into  
the nature of the problem?:   
        4. Some of the ideas or results will substantially help other people's  
ongoing research.  
  
REPLICABILITY: Will members of the ACL community be able to reproduce or  
verify the results in this paper?:   
        4. They could mostly reproduce the results, but there may be some  
variation because of sample variance or minor variations in their  
interpretation of the protocol or method.  
  
IMPACT OF PROMISED SOFTWARE:  If the authors state (in anonymous fashion)  
that their software will be available, what is the expected impact of the  
software package?:   
        2. Documentary: The new software will be useful to study or replicate the  
reported research, although for other purposes it may have limited interest  
or limited usability. (Still a positive rating)  
  
IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)  
that datasets will be released, how valuable will they be to others?:   
        4. Useful: I would recommend the new datasets to other researchers or  
developers for their ongoing work.  
  
  
TACL-WORTHY AS IS? In answering, think over all your scores above. If a  
paper has some weaknesses, but you really got a lot out of it, feel free to  
recommend it. If a paper is solid but you could live without it, let us know  
that you're ambivalent.  
  
Reviewers: after you save this review form, you'll have to make a  
confidential recommendation to the editors via pull-down menu as to: what  
degree of revision would be needed to make the submission eventually  
TACL-worthy?  
:   
        5. Strong: I'd like to see it accepted; it will be one of the better papers  
in TACL.  
  
  
Detailed Comments for the Authors  
  
Reviewers, please draft your comments on your own filesystem and then copy  
the results into the text-entry box.  You will thus have a saved copy in  
case of system glitches.  
:   
        The paper presents an analysis of RNN-based character-based neural  
language models (CNLMs). An interesting take is to train the RNNs on  
raw untokenized input, and subsequently analyze (or probe) the models  
across the levels of the linguistic hierarchy (see details  
below). Multiple languages are considered (English, German and  
Italian). The probing tasks include:  
  
- phonological properties (phonological classes as induced via  
  agglomerative clustering; acceptability of bigrams phonotactically  
  acceptable in one language, while not so in the other language)  
  
- word segmentation (here, the paper performs experiments on two  
  datasets - Wikipedia and Brent's child-directed speech corpus; the  
  latter to compare to a Bayesian model)  
  
- syntactic properties (mostly derived from UD data, e.g., verb-noun  
  distinctions; gender, case and sub-categorization properties with  
  increasing number of intervening elements)  
  
- a semantic task (sentence completion task - 5-word multiple choice  
  test)  
  
The paper is very well written, and presents itself well in light of  
the (at times very recent) literature. The experimental evaluation is  
sound and extensive, with carefully constructed setups across the  
linguistic spectrum.  
  
I found it a pleasure to read this paper. I have a couple of  
suggestions for improvements.  
  
1. Section 4.2 presents results on word segmentation. The paragraph  
   starting on line 464 qualitatively investigates the errors made by  
   the CNLM trained on Wikipedia test (note: it would be beneficial to  
   state Wikipedia right at the beginning of the paragraph, rather  
   than at its end). It would though be more interesting if this were  
   a comparison between the Bayesian and the CNLM model, rather than  
   just analyzing the CNLM.  Because, albeit the fact that "CNLM  
   performance is comparable" (ref. to Table 4), a close look reveals  
   that there is quiet a gap of the two models in terms of precision  
   on inducing lexical word types. A comparative analysis would shed  
   some light here, it might be that the LSTM gets frequent types  
   right but misses other types, compared to the Bayesian method  
   constructed with a lexical bias in mind.  
  
2. For the first analysis (phonological classes induced by the output  
   embeddings) results for German only are provided in Figure 1. The  
   paper should include plots for all three languages, as there is no  
   clear motivation why one was selected. There should be space to  
   include all three plots.  
  
3. What really surprised me is the bad performance of the vanilla RNN  
   compared to the LSTM on the bigram acceptability judgment task  
   (lines 382-383). This is in fact dramatic, as the model only needs  
   to consider adjacent characters. At first it seems the model is  
   underfit, but then the RNN performs reasonably well on other tasks,  
   sometimes even being close to the LSTM (e.g., adj-gender agreement  
   on Italian, Table 7) and perplexity scores are reasonable as  
   well. Maybe a further discussion in light of training data  
   properties and locality of the task might shed some light here (how  
   long are the paragraphs the models are trained on?). Finally, what  
   is also surprising is that the RNN does not improve with in-domain  
   training data for the last task (sentence completion, see line 2 in  
   Table 8). Why is the vanilla RNN not improving? Would it help to  
   fine-tune on the in-domain data?  
  
4. The paper does a great job in discussing related work. I though  
   kept wondering about the difference with Kementchedjhieva & Lopez  
   (2018). While overall results are in line (RNN-LMs do capture  
   morphological properties), the paper is very brief on reporting an  
   interesting divergence: "we could not replicate the result with our  
   model" (on a single neuron tracking morpheme boundaries). It would  
   be interesting to know if this is due to the different modeling  
   setup (e.g., would this also hold for the model trained with white-space,  
   footnote 6?) or what other reasons there could be at play.  
  
Smaller, possible typos and stylistic suggestions:  
  
- Table 3: check F1 score for Italian (should be 59 rather than 60)  
- Presentation of results in Table 3 and 4: use of different decimal places.  
- Colored figures are unreadable in b/w printing.  
- line 936: in Figure 8 > in Table 8  
  
REVIEWER CONFIDENCE:   
        4. Quite sure. I tried to check the important points carefully. It's  
unlikely, though conceivable, that I missed something that should affect my  
ratings.  
  
------------------------------  
  
------------------------------  
Reviewer C:  
  
CLARITY: For the reasonably well-prepared reader, is it clear what was done  
and why? Is the paper well-written and well-structured?:   
        2. Important questions were hard to resolve even with effort.  
  
  
INNOVATIVENESS: How original is the approach? Does this paper break new  
ground in topic, methodology, or content? How exciting and innovative is the  
research it describes?  
  
Note that a paper can score high for innovativeness even if its impact will  
be limited.  
:   
        3. Respectable: A nice research contribution that represents a notable  
extension of prior approaches or methodologies.  
  
SOUNDNESS/CORRECTNESS: First, is the technical approach sound and  
well-chosen? Second, can one trust the claims of the paper -- are they  
supported by proper experiments and are the results of the experiments  
correctly interpreted?:   
        2. Troublesome. There are some ideas worth salvaging here, but the work  
should really have been done or evaluated differently.  
  
  
RELATED WORK: Does the submission make clear where the presented system sits  
with respect to existing literature? Are the references adequate?  
  
Note that the existing literature includes preprints, but in the case of  
preprints:  
• Authors should be informed of but not penalized for missing very recent  
and/or not widely known work.  
• If a refereed version exists, authors should cite it in addition to or  
instead of the preprint.  
:   
        4. Mostly solid bibliography and comparison, but there are a few additional  
references that should be included. Discussion of benefits and limitations  
is acceptable but not enlightening.  
  
  
SUBSTANCE: Does this paper have enough substance (in terms of the amount of  
work), or would it benefit from more ideas or analysis?  
  
Note that papers or preprints appearing less than three months before a  
paper is submitted to TACL are considered contemporaneous with the  
submission. This relieves authors from the obligation to make detailed  
comparisons that require additional experiments and/or in-depth analysis,  
although authors should still cite and discuss contemporaneous work to the  
degree feasible.  
:   
        2. Work in progress. There are enough good ideas, but perhaps not enough  
results yet.  
  
IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the  
ideas are novel, will they also be useful or inspirational? If the results  
are sound, are they also important? Does the paper bring new insights into  
the nature of the problem?:   
        3. Interesting but not too influential. The work will be cited, but mainly  
for comparison or as a source of minor contributions.  
  
REPLICABILITY: Will members of the ACL community be able to reproduce or  
verify the results in this paper?:   
        1. They would not be able to reproduce the results here no matter how hard  
they tried.  
  
IMPACT OF PROMISED SOFTWARE:  If the authors state (in anonymous fashion)  
that their software will be available, what is the expected impact of the  
software package?:   
        1. No usable software released.  
  
IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)  
that datasets will be released, how valuable will they be to others?:   
        1. No usable datasets submitted.  
  
  
TACL-WORTHY AS IS? In answering, think over all your scores above. If a  
paper has some weaknesses, but you really got a lot out of it, feel free to  
recommend it. If a paper is solid but you could live without it, let us know  
that you're ambivalent.  
  
Reviewers: after you save this review form, you'll have to make a  
confidential recommendation to the editors via pull-down menu as to: what  
degree of revision would be needed to make the submission eventually  
TACL-worthy?  
:   
        2. Leaning against: I'd rather not see it appear in TACL.  
  
  
Detailed Comments for the Authors  
  
Reviewers, please draft your comments on your own filesystem and then copy  
the results into the text-entry box.  You will thus have a saved copy in  
case of system glitches.  
:   
        This paper tests the conjecture that LSTMs can learn more than just  
spelling from streams of text, but also things like word boundaries (when  
spaces are removed) and the phonetic categories of characters.  The authors  
postulate that this is more similar to the task infants face when learning  
to parse utterances, and is a truer test of what an LSTM can learn.  
  
I think this is an interesting area of inquiry.  The experiments in this  
paper are extensive, but sometimes don't seem to fit the intent of the  
authors and/or are not clearly explained.  The abstract really focuses on  
the idea of removing spaces and still being able to recover words and  
morphology, but the experiments veer away from that pretty quickly (starting  
with experiment 5 below).  In general, there are too many experiments  
crammed into this paper, and not enough explanation of the experimental set  
up, or careful consideration of results.  This paper is right at the page  
limit, so I think the authors should reconsider which experiments are most  
telling, and move some of the extraneous ones to supplementary material.  I  
can't figure out from the TACL page if TACL allows supplementary material,  
but in any case, there's too much in these 10 pages to cover in the detail  
required for a reader to understand and be able to reproduce any of these  
results.  
  
  
Here's a list of some of the experiments, and my questions for each  
experiment  
  
1. Remove spaces, how does that affect perplexity/bits-per-char?  
I'm not convinced that removing spaces is a good proxy to the word  
segmentation problem infants and young children encounter, since they are  
exposed to much simpler language (single words, very simple sentences).  
  
2. Cluster characters by their embeddings.  Do the cluster represent  
phonetics?  
This experiment is not repeated (or results are not shown) for the RNN.  No  
details are given for how the clustering was run (distance metric?) and the  
cutoff for clusters appears to be chosen arbitrarily.  
  
3. Identify some acceptable and unacceptable bigrams in each language.   
Train on data with both sets removed, and then test if the held out bigrams  
are assigned probabilities that are consistent with the  
acceptable/unacceptable categorization.  
Here, I am very surprised that the RNN did so terribly, to the point where I  
wonder if there is a bug in the analysis or code.  If there is no bug, I  
think a better explanation for this behavior needs to be brought forward.   
For example, perhaps the clustering as in Fig 1 would show that the  
phonological categories are not learned by the RNN, which would help to  
explain the lack of generalization we're seeing in this experiment (which  
requires learning phonological categories).  
  
4a. Word segmentation  
This experiment is not fully explained.  In particular the context PMI is  
unclear to me here, and needs more explanation.  But somehow they are  
creating features which they use to predict which characters start words  
  
4b. A small little experiment with a LDA word segmenting algorithm is  
included here, but so little detail is given that we can't draw much of a  
conclusion.  It's also trained on a different corpus, so it sticks out a  
bit.  Suggest this be put into a supplementary material section with more  
details.  
  
4c. Error analysis   
This is actually fairly interesting and I appreciate this qualitative  
account  
  
4d. Compare PMI to hierarchical distance  
This experiment is really light on details and the accompanying figure 2 HAS  
NO LABELS WHATSOEVER.  No axis labels and no legend labels!  There is only  
one paragraph actually explaining this experiment, and it's not nearly  
enough to understand the results.  
  
At this point we begin to veer off course, and the models seem to be trained  
and/or tested on single words, which makes a bit of sense sometimes (e.g.  
experiment 5 below which uses the models trained in previous sections) but  
not always.  
  
5. Nouns vs verbs: can they be classified using the final hidden state of a  
pre-trained model after reading the last char?  
I don't speak German, but this sentence doesn't make any sense to me  
"requiring that they end in -en (German) or -re (Italian) (so that models  
can’t rely on the affix for classification), " how would restricting the  
suffix (en, re) also restrict the affix?  The baseline here is an  
autoencoder LSTM trained on words in isolation.  This seems like a straw  
man, if only shown words in isolation this model is missing much of the  
context information that help the context-full LSTM tell verbs from nouns.  
  
6. Can the model detect number  
Here I'm unclear what this has to do with the model trained on space-free  
text.  The authors seem to be training on single words? "For the training  
set, we randomly selected 15 singulars and plurals from each training  
class."  The results show that the CNLM can't generalize to umlaut, but the  
explanation is lacking (suffix vs internal root vowel change).  Why? is the  
interesting question here.  
  
There are many more experiments after this point, and the main themes of my  
critiques are the same.  There is not enough information given to fully  
understand these experiments (and thus replicating would be impossible).   
The figures have NO labels.  There is no careful consideration of results.  
  
---- Minor comments ----  
line 242, the models were not trained until validation accuracy plateaued?   
That does not seem standard.  How can we know if these models are fit to  
compare against each other if we're not sure they're done training?  
  
The citations for the figures/tables are missing a lot of information.  It's  
nice to not have to scan through the text to figure out what each figure is  
showing, and many of the important details are left out of the  
figures+captions (e.g. the acceptable/unacceptable order in Table 2, what  
model is used to do the clustering in figure 1).  This is a little of  
personal preference (which is why I include it here in the minor comments),  
but it makes for an annoying first skim of the paper if you can't figure out  
any of the figures without.  E.g. the caption for Table 5: "word class  
accuracy with standard errors. ..."  For what task???  
  
Tables with 9 cells, and 3 numbers per cell are pretty hard to parse e.g.  
Table 3/4.  A bar graph with just F1 would be nice, unless the authors  
actually think P/R make any contribution (they don't seem to talk about P/R  
in any detail).  
  
Table 3 just gives P/R/F1, I don't think the claim on line 428 (classify  
half the tokens correctly) can be inferred from P or R, rather, one needs  
accuracy.  
  
Section 4.2 needs some subheaders, there's too much going on here and it's  
hard to keep track of what the point of the current experiment is.  
  
A few tables/figures have STD or bootstrapped CI, but many do not. Would  
like to see them consistently throughout  
  
Line 616: "as above" there's a lot of stuff above at this point, please  
refer to something more concrete  
  
Is table 8 really comparing results across corpora?  The top 3 models are  
trained on wikipedia, and the bottom on Sherlock Holmes?  This is not a  
sound comparison, not sure what we're supposed to take away from this  
experiment  
  
REVIEWER CONFIDENCE:   
        4. Quite sure. I tried to check the important points carefully. It's  
unlikely, though conceivable, that I missed something that should affect my  
ratings.  
  
------------------------------