Dear Authors,

As you can see the two previously assigned reviewers were very positive. The new reviewer not so much.

The previous action editor and one of the previous reviewers were not available, which is unfortunate.

I think reviewer C makes many excellent points, but adequately dealing with their concerns would require a complete reorientation of the research -- which would be unfair given that you seem to have implemented all changes recommended in the last round.

So C's recommendations are not mandatory for you to implement -- however, please carefully read the review and try to address C's issues where possible.

Reviewers A and B have many good comments, which I would like to see reflected in your resubmission.

MANDATORY CHANGE: The reviewers agree that the word boundary finding is the most interesting part of the paper. But you don't show what word recognition rates are on actual text. Please do the following additional experiment: "duplicate" Table 6 and also report word recognition rates on real text. I guess F1 would be a standard and clear measure. (90% for English on the artificial data set doesn't tell me how well your model segments real text.)

Below you find two additional typos/unclarities (over and above those pointed out by the reviewers) and below that the message generated by the TACL machine.

Best - Hinrich

line 995: perhaps add that the error analysis is for the "single unit" classifier (can be inferred only from the result Table 6, where German is the language w/ the single-unit classifier reaching highest accuracy).

link 962: "aditional dataspoints"

Dear (anonymized):

As TACL action editor for submission 1709, "Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text",   I am happy to tell you that I am accepting your paper subject (conditional) to your making specific revisions within two months.

LIST OF MANDATORY REVISIONS:

Generally, your revised version will be handled by the same action editor (me) and the same reviewers (if necessary) in making the final decision --- which, *if* all requested revisions are made, will be final acceptance.

You are allowed one to two extra pages of content to accommodate these revisions. To submit your revised version, follow the instructions in the "Revision and Resubmission Policy for TACL Submissions" section of the Author Guidelines at
https://transacl.org/ojs/index .

Thank you for submitting to TACL, and I look forward to your revised version!

Hinrich Schütze
Ludwig Maximilian University of Munich
inquiries@cislmu.org

....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?:
        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.
:
        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?:

5. The approach is very apt, and the claims are convincingly supported.

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?:

3. Potentially useful: Someone might find the new software useful for their work.

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?
:

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.
:

Title: Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text.

This paper asks whether neural nets, trained on a simple language modeling objective with character-based, non-word delimited text as input nonetheless pick up a sensitivity to words as recognized by speakers (and marked with white space in the orthography). With experiments on German, Italian and English, concerning morhpological categories, agreement, and sentence completion, the authors show that the models are picking up on words as a significant unit. They furthermore follow the methodology of Kementchedjhieva and Lopez (2018) to investigate whether the NNs devote 'units' to boundary detection and find that they do.

This is a resubmission for which I was previously reviewer A. I find that the revised paper is much stronger -- clearer, more focused, and without the problematic overclaims of the original. I have a very contentful remarks + some stylistic points below. I believe this paper should be accepted with minor revisions.

Contentful comments:

ln 070-073: I'd like a citation or two substantiating the claim that this is standard.

ln 192: What is the connection between the prior work in this paragraph and your work?

ln 222: "Radford et al. (2017) focused on CNLMs deployed in the domain of sentiment analysis." The fact that they were working on sentiment analysis doesn't seem relevant to the current discussion. What did they learn about CLNMs? Why is this paper relevant?

ln 242: Here, too, I miss a brief statement of why your questions are the logiacl next step to take.

ln 689 "For the n-gram baseline, we only counted occurrences of the prepositional phrase, omitting sentences.": This use of "omitting" makes it sound like were tested on both PPs and sentences. But surely that's not right?

ln 737 "We required moreover the -a and -o forms of a noun to be reasonably balanced in frequency (neither form is twice more frequent than the other)," Why? How does this decision contribute/relate to the overall research goal?

ln 947 "Again, in left-to-right processing, the unit has a tendency to immediately posit boundaries when frequent function words are encountered." This suggests a way in which NN processing is quite unlike human processing, which can recover from incorrect decisions in light of further information.

ln 1162 "Intriguingly, our CNLMs captured a range of lexical phenomena without anything resembling a word dictionary": Your results seem to suggest they build one internally, though! I think you’re confusing levels of abstraction here. No one argues that there are specific neurons for each word in the human brain.

Stylistic points/typos:

ln 013 reached -> has reached
ln 135 is -> are
ln 178 "that CLNMs hierarchical structure": Missing 'model'?
ln 212 work -> word
ln 559 The sentence final . in the middle of a parenthetical that is itself embedded in a sentence is awkward.
ln 724 "Italian has a relatively extended paradigm": This led me to expect lots of suffixes, not lots of stems. I'd suggest revising this.
ln 822 capable to track -> capable of tracking
ln 1042, 1048: ' -> ` x2
ln 1121 input, -> input
ln 1126 latter ability -> latter's ability

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?:

4. Understandable by most readers.

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?:

5. The approach is very apt, and the claims are convincingly supported.

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.
:

5. Contains more ideas or analysis than most publications in this journal;
goes the extra mile.

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?:

5. They could easily reproduce the results.

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?:

4. Useful: I would recommend the new software to other researchers or
developers for their ongoing work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion)
that datasets will be released, how valuable will they be to others?:

5. Enabling: The newly released datasets should affect other people's
choice of research or development projects to undertake.

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?
:

6. Exciting: I'd fight to get it accepted; probably would be one of the
best papers in TACL this year.

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 in-depth analysis of a character-based recurrent neural network trained on unsegmented input, to investigate (probe) whether such RNN-based models learn genuine linguistic (morphosyntactic) information.

The experiments span, as a first step, a motivated set of three languages (German, Italian and English) and the following tasks:

- syntactic categories (verb vs nouns)
- agreement (number, gender, cases)
- sentence completion (5-word prediction task)
- word segmentation

In comparison to the earlier version, the paper is more coherent.

I agree that the dropped phonological investigation makes it clearer, resulting in a more coherent story that is targeted around the notation of 'word-hood'. In this light, the question of what constitutes a word (and segmentation section) has been largely extended, and it includes a more in-depth analysis of the word boundaries that are induced, complementing (rather than contradicting, i.e., not being able to reproduce) earlier findings (Kementchedjhieva and Lopez, 2018). I find the overall story and flow very pleasant, the paper more focused, and I am glad to my co-reviewers to have raised the detailed concerns about the phonological parts, because the new version is more convincing and focused, and the empirical results are stronger aligned to the theoretical question.

The authors acknowledge and explain the low results of the RNN, and add a footnote that this might be due to resource availability. However, I am still puzzled, as the authors. While I agree that leaving a further investigation on the low RNN performance to future work, this makes me think that the paper might be better off dropping the RNN model completely. What's the real benefit of keeping it, other than showing it does not work so well as the LSTM?
In fact, it is still pretty bad, particularly in Table 5 in-domain vs cross-domain giving the same performance? Also, by keeping the RNN, one could argue why not instead investigate GRUs as well, which is a bit besides the point of the paper. On the word class prediction task, the RNN has a huge standard deviation, it is very unstable. I'd say drop the RNN, and instead, add a random baseline throughout the paper, and explain the autoencoder baseline (regarding random: even though for most tasks the chance-level is just 50%) -- see comment on missing autoencoder setup below.

Regarding point 4. in the answer to reviewers, reviewer B (difference to Kementchedjhieva and Lopez, 2018):

I am very glad that my comment encouraged this additional work on word boundary detection, which I find particularly pleasing. Thank YOU! This is clearly going an extra mile.
As looking 'beyond' the threshold meant getting confirmation and additional very interesting results, I would appreciate if the paper could openly say so (e.g., in a footnote? appendix?) I think this *is* an educative aspect, which is now only visible to reviewers. Without lowering the threshold this interesting analysis (including the quantitative analysis in Figure 2) would not be part of the paper. Is there a way to 'keep' this?

I found it a pleasure to read this paper (already in its original format, but even more so this stronger version).

I have a couple of minor suggestions/comments/questions for improvements:

- "soft" in abstract: why "soft" word boundaries? Evaluation (Table 6) assumes a hard word boundary prediction tasks, hence consider to drop "soft", which is not explained.

- experimental setup: the paper now introduces an autoencoder as baseline. As mentioned above, I would propose to explicitly add an even simpler baselines (random or majority). Moreover, there is a large space of possibilities for autoencoders (sparse, overcomplete...) and the current version entirely misses to describe the autoencoder exp. setup. it could be added as a paragraph at the end of Section 4.

- "the very notion of what counts as a word" - this is in fact an important general question, there must be a less recent reference, before 2017?

- Table 2 + 3 presentation of results: I typically find it easier if the baseline is at the start of the table, then models, then other (like the "subs" model, which is not strictly comparable, could be be moved to the bottom of the table separated by a hline). This "teasing apart" would make the tables more readable, I believe.

- The plural noun number classification experiment (page 5) has a well-motivated setup, the test sets (-r, and umlaut change) seem to have been set up in a way to consist of the more difficult tasks (while training on the more regular forms -n/-s/-e), which the last part seems to hint at "generalization is not completely abstract". If this is in line of what the authors had in mind when designing the experiments, this motivation could be made overt in favor of strengthening the motivation of the setup (because otherwise once could be inclined to ask for more permutations of these experiments..). ;)

- line 167: than predicting -> than prediction
- line 178: missing verb in CNLMs sentence
- line 467: "successful outperforming in most cases" - split in multiple sentences. "... It outperforms.."

- line 473: "near-random" -> add random baseline to table
- line 514: "controlled for character length" - I might have missed how this has been done
- line 570: comma in number (in a few other places as well)
- line 936: nice example with Hauptaufgabe!
- line 1042: ' -> `

- discussion section: consider adding a few new lines (first paragraph) to make it more readable, e.g., in line 1124 and 1131

- consider adding a reference; an earlier study that proposes the use of segmentation-free NLP (Schütze, 2017): http://aclweb.org/anthology/E/

- "and a dummy variable coding word-final position" - this part remains unclear to me. How exactly twas this per-neuron correlation done, how was this dummy-variable derived?

Thank you for a very thorough and insightful response to the editors and reviewers.

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?:
 4. Understandable by most readers.

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.
:
 2. Pedestrian: Obvious, or a minor improvement on familiar techniques.

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?:

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?:

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?:

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?:

3. Potentially useful: Someone might find the new datasets useful for their 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?
:

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.
:

The paper presents an investigation into what information character-level LSTM language models learn, through different probing tasks.
In particular, the language models are applied in a setting where word boundaries are removed, in order to make the task more challenging and cognitively realistic.
A number of different properties are investigated, including word class and number detection, gender and case agreement, and sentence completion.
The final section investigates how these models, though trained on unsegmented character streams, still specialise individual neurons for the detection of word boundaries.

The paper is very well written and is pleasant to read. There are also a large number of different experiments to probe various properties.

However, I did not feel that the overall setting of removing word boundaries is sufficiently argumented or justified. The motivation seems to be that this is how humans experience language. However, written text is already quite different from human speech and I wouldn't say it becomes substantially more similar by removing the word boundaries. If punctuation is kept (even dashes in words), under the argument of encoding prosody information, then surely word boundaries also indicate certain prosodic features. Instead, the whole setting in the paper, in which neural character-based language models are investigated, does not match any of the normal settings in which a neural character model would be applied. If the goal was to investigate LSTM performance on human speech, then the experiments could have been performed on actual audio input, using aligned transcriptions as the targets. If the goal was to investigate LSTM performance on text, then I would suggest including the results from a word-delimited character LM as well.

I would have expected the probing tasks to either:
a) compare different language model architectures and conclude which ones are better at which tasks, or
b) analyse the performance of one language model architecture across different phenomena and conclude what is it good at, where are its weaknesses and what does it mean for future work.
Unfortunately, I did not see either of these in the current paper. The main conclusions seem to be that LSTMs still perform reasonably well after introducing the artificial constraint of removing word boundaries, and I'm not sure what this shows or how this will be useful.

It is fairly expected that language models would learn to encode information such as word class, number or gender. They are specifically trained to predict the surrounding words/characters, which requires this information for agreement. However, it is less clear to which extent the supervised experiments actually show this property - given that a separate supervised component is trained on top of the LM representations, it could theoretically be picking up on useful feature correlations instead of the desired property directly.

Several of the chosen baselines seem to be particularly poor in this paper. Plain RNNs are not used in practice any more, as their lower performance is very well established. And the ngram model is not really a proper ngram language model - L577 says it's basically just picking the most frequent case in the data, based on one word of context. Giving the model only one word of context when several of the tasks are specifically constructed to require 2 or more words of context seems very unfair. I do not see why a proper ngram language model could not have been used (e.g. a 5-gram model with Kneser-Ney smoothing). A truly useful investigation would be to include some of the more recent state-of-the-art language model architectures into the comparison.

The finding of the word boundary neurons is probably the strongest part of the paper.
However, as pointed out in the paper as well, this finding has already been shown in previous work (Kementchedjhieva and Lopez, 2018).
It is unclear what the differences are and what is the novel contribution in this section.

The argument that word boundaries are not necessary in language is somewhat weakened by the finding that even without boundary information LSTMs still adapt by explicitly learning to detect word boundaries internally.

Many of the experiments seem to be very heuristically constructed with not much motivation provided. This includes the limited choice of languages for investigation, the words that are selected for training or testing in various experiments, various filters based on frequency and suffixes, etc. Making the experiments more general and providing better explanations for the remaining choices would make the paper stronger.

Spelling errors:
L178: I think a word is missing
L180: Unnecessary comma
L963: dataspoints

REVIEWER CONFIDENCE:

3. Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.