The LAMBADA dataset: Supplementary Material

We provide here further details on the construction of the LAMBADA dataset and on model implementation and training.

Criteria for passage construction. In addition to those specified in the Section 3.1 of the paper: 1) target sentence contains at least 10 tokens; 2) target word occurs at least 5 times in the training section of the corpus OR occurs in the passage (this is to account for cases in which the target word is rare but it can nonetheless be guessed given the broad context, for instance the name of a character); 3) context is the minimum number of complete sentences before the target sentence such that they cumulatively contain at least 50 tokens; 4) we randomly sampled no more than 200 passages from each novel in the development+test partition (to avoid the use of knowledge about the particular novel when solving the task).

Further information on the in-house-trained language models used for passage filtering (see Section 3.1 of main paper). Due to various practical considerations and their different purpose, the language models used for filtering are different in architecture, training method, vocabulary and tuning from those we used for testing. The whole set of 5,325 novels was used to train the language models employed for filtering (for the feed-forward neural network and recurrent neural network, a 100K-sentence validation set was randomly separated from the training data). Note that, to maximize their performance on the data of interest at the cost of overfitting, these language models were built on the same corpus on which they were applied. The vocabulary included the 50K most frequent words (plus the UNKNOWN symbol).

The 4-gram language model was trained with the CMU toolkit (Clarkson and Rosenfeld, 1997) and used Witten-Bell discounting (Bell et al., 1990). The feed-forward network was trained with the SOUL Toolkit (Le et al., 2011), with 10-gram input, word embedding size and hidden layer sizes of 300, 500, and 300 respectively, learning rate 0.02 (it converged after 10 iterations), and hierarchical softmax in the output randomly partitioning the vocabulary into 2,000 equally-sized word classes. For the RNN, the hidden layer size was 256, the learning rate was initially set to 1, decreasing when the validation perplexity increased, and we used hierarchical softmax with 1,200 classes.

The CrowdFlower survey. The instructions given to participants that were asked to predict words in the full-passage condition (steps 1 and 2 in Section 3.1 of the paper) are shown in Figure 1. A screenshot showing how example items were presented to subjects in the full-passage condition is shown in Figure 2. Analogous data for the sentence-level condition (step 3 in Section 3.1 of the paper) are presented in figures 3 and 4, respectively.

Quality checks in the crowdsourcing experiment. Only subjects that reside in the following English-speaking countries could participate: UK, Ireland, Canada, USA, Australia, and New Zealand. We asked subjects to participate only if they were native speakers of English. We allowed highest-quality (Level 3) performance level contributors only (as measured by CrowdFlower), manually examined the answers of under- and over-performers (subjects below 5% and above 50% hit rate in the first step were subsequently removed from further data collection), and included test questions. The latter were sampled from data points obtained in a pilot study. For the test questions, we accepted a larger pool of plausible alternative answers in addition to the original missing words from the passages. Subjects below a 60% hit rate on test questions were excluded from

Important Notice

By participating in this survey you declare that you have read and understood the contents of the following documents and consequently agree to take part in the study, and that you agree with the processing of your personal data described in the second document.

`http://clic.cimec.unitn.it/composes/materials/informed_consent.html`
`http://clic.cimec.unitn.it/composes/materials/data_protection.html`

Instructions

Please accept this job only if you are a native speaker of English.

You will be presented with short text passages missing the last word (one word only!), and your job is to guess what the last word might be.

For example, you might read:

Context: Rosa-Lee and Roberto drew their swords, ready. The Falcon walked closer, straight to Rosa-Lee, looked at her, and said, “Who have we here?” Reaching toward her, he lifted the torch to her face and smirked. “Ah, Señorita, lovelier than ever, I almost did not recognize you in men ’s clothing.” He cupped her cheeks again. Her heart was racing as she slapped his hand away, lifting her ______ .

Guess: sword

You might decide to type “sword” (as appropriate given the example), since this is a plausible ending of the passage. Sometimes, it might be very difficult to find an appropriate word, but please always make a guess, even if you’re not fully convinced by your answer.

If, as in the example below, you find several guesses to be possible (e.g. “lunch” or “meal”), select the one you find the most likely.

Context: He was hardly recognizable, the suit replaced by baggy jeans, tee shirt and faded leather jacket, on his feet a well scuffed pair of trainers that had started out life white. On his head the blue of a Chelsea FC cap, over his shoulder a gym bag that had seen better days. He walked through the door to the restaurant. It was half full with the lunchtime crowd. Waving to acquaintances he sat down and ordered a bowl of soft noodles but instead of waiting he quickly walked through the kitchen area, nodded to the owner and let himself out through the back door to the alley behind. The young colored guy, in the grey hoody, settled in a doorway on the opposite side of the street and waited for Tony to have his ______ .

Guess: lunch

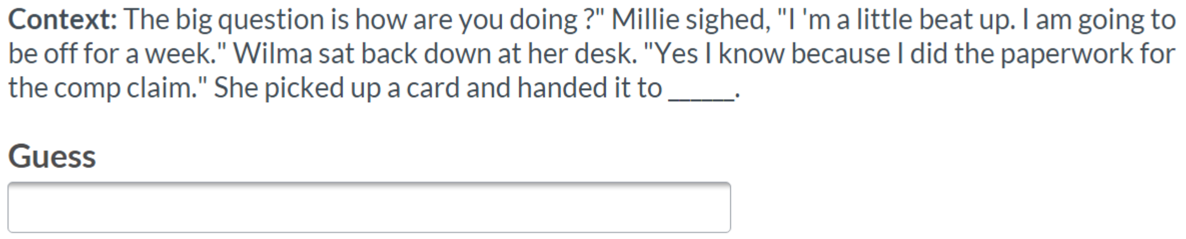
Each guess should consist of a single word only (e.g. “lunch” but not “quick lunch”)

We are interested in your linguistic intuition - what word *you* would guess in the context. Please do not search for the passage online: that is not helpful to us, as we already know how these passages end! **Searching online may lead you to be flagged.**

To ensure that your guess does indeed reflect your linguistic intuitions, we have included some test questions where guessing the final word after reading the passage is very easy (given our experience in previous experiments). If you read the passage carefully, you should have no problem guessing the final word of a test question.

Thanks for your participation, and have fun!

Figure 1: Instructions given to participants for passage-level data collection.



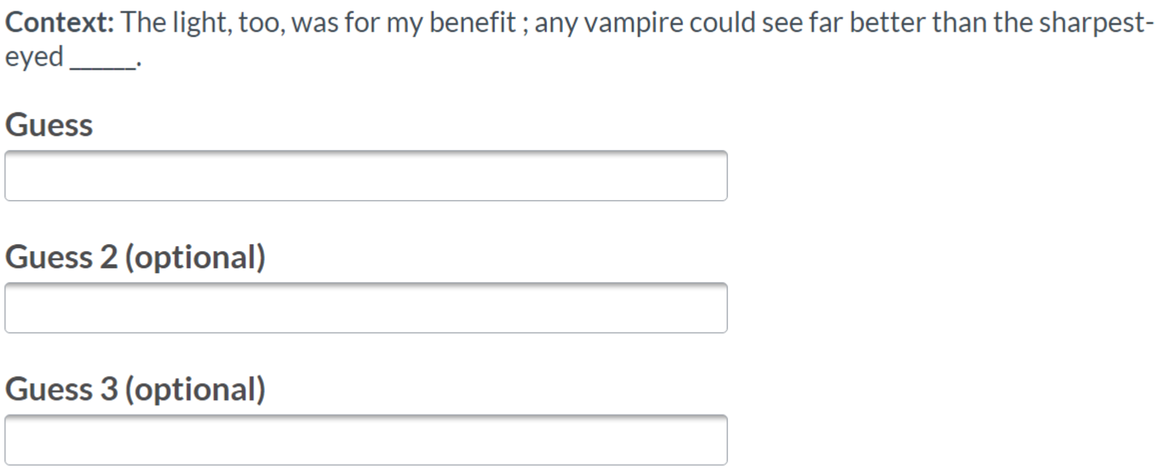
Context: The big question is how are you doing ?" Millie sighed, "I 'm a little beat up. I am going to be off for a week." Wilma sat back down at her desk. "Yes I know because I did the paperwork for the comp claim." She picked up a card and handed it to _____.

Guess

Figure 2: Passage-level data collection interface.

Important Notice
By participating in this survey you declare that you have read and understood the contents of the following documents and consequently agree to take part in the study, and that you agree with the processing of your personal data described in the second document.
`http://clic.cimec.unitn.it/composes/materials/informed_consent.html`
`http://clic.cimec.unitn.it/composes/materials/data_protection.html`
Instructions
Please accept this job only if you are a native speaker of English.
You will be presented with sentences missing the last word (one word only!), and your job is to guess what the last word might be.
For example, you might read:
Context: "Let 's get fizzical, fizzical," Taylor intonated the century old tune that he had heard on the Lilly-Book, swaying his hips and grabbing Lilly Grace's arms for a little ______ .
Guess: dance
You might decide to type "dance" (as appropriate given the example), since this is a plausible ending of the sentence. Sometimes, it might be very difficult to find an appropriate word, but please always make a guess, even if you're not fully convinced by your answer. On the other hand, in some cases you might think of more than one appropriate word: in that case, you may pick up to three words that seem most plausible, without trying to choose the best one. If you choose more than one word, type them in separate fields, for example:
Context: "Tom has always been a heck of a good person Mrs. Palmer, I can't see why your husband would've had a problem ______ ? "
Guess: there **Guess 2:** here
While you are welcome to enter multiple guesses, **each guess should consist of a single word only** (e.g. "dance" but not "quick dance")
We are interested in your linguistic intuition - what word *you* would guess in the context. Please do not search for the sentence online: that is not helpful to us, as we already know how these sentences end!
Thanks for your participation, and have fun!

Figure 3: Instructions given to participants for sentence-level data collection.



Context: The light, too, was for my benefit ; any vampire could see far better than the sharpest-eyed _____.

Guess

Guess 2 (optional)

Guess 3 (optional)

Figure 4: Sentence-level data collection interface.

the data collection. We limited to 300 the number of judgments that could be provided by each subject in each CrowdFlower job (there were 107 total jobs at the passage level, 47 jobs at the sentence level). We also set a minimum time of 1 minute to complete a page, in order to exclude hurried, thoughtless answers.

To ensure that it would not be possible for the same subject to judge the same item at the passage level and at the sentence level, thereby using discourse information when carrying out the sentence-level task, we sent the respective jobs to complementary countries. The countries were divided into two groups: USA, Ireland and Australia in Group 1 and United Kingdom, Canada and New Zealand in Group 2. For a given batch of input data, if the passage-level job was sent to Group 1, then the sentence-level job was sent to Group 2 (or vice-versa, in a counterbalanced fashion). As a curiosity, we observed (on average) higher and more uniform performance of Group 2 subjects compared to Group 1, in which we had to remove many under- and over-performers.

Further information on the language models tested on LAMBADA (see Section 3.1 of of main paper). Here we present the details of the models that we tested on LAMBADA (as opposed to the models used for pre-processing, discussed above).

We implemented our own RNN/LSTM code.[1] For these models, we used the development data to explore the following hyper-parameters: hidden layer size (128,256,512 – picked 512 for both) and network depth (1,2 – picked 1 for both). The initial hidden state was initialized to 0 and the models weights to random values in $(-0.1, 0.1)$. RNNs were unrolled for 10 time steps and LSTMs for 35. For the 2-layered models, Dropout was used for regularization, following Zaremba et al. (2014). We used hierarchical softmax with the vocabulary words randomly split into 1,200 equally-sized classes.

The N-Gram models were constructed using the SRILM toolkit (Stolcke, 2002). Other than the parameters described next, we set the `-unk`, `-no-eos` and `no-sos` flags. For the simple N-Gram model, we calibrated smoothing method (Original Kneser-Ney, Chen and Goodman's Kneser-Ney, Witten-Bell – picked Witten-Bell), size (4,5 – picked 5) and interpolation (yes or no – picked yes). For the cache model we took the selected N-Gram model and explored interpolation parameters (9 options between 0.1 and 0.9 – picked 0.1).

The Memory Network was trained using the code available at `https://github.com/facebook/MemNN`. We used the default except for the following parameters: α (0.008, 0.009, 0.01 – picked 0.01), memory size (50, 100 – picked 50) and hidden layer size (150, 250 – picked 250).We used a full-softmax output layer.

For CBOW-Sup we tuned the learning rate (12 values between 0.2 and 0.004 – picked 0.2), and also used full softmax.

Note, finally, that for Unsup-CBOW we had to use random vectors for nearly 8K vocabulary items for which we did not have semantic vectors.

References

Timothy C Bell, John G Cleary, and Ian H Witten. 1990. *Text compression*. Prentice-Hall, Inc.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech*, pages 2707–2710.

Hai Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. 2011. Structured output layer neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5524–5527. IEEE.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, volume 2002, page 2002.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv*.

[1] Available on `https://github.com/quanpn90/lambada`