A Experimental Details

In all of our experiments, we train embeddings on the Westbury Wikipedia Corpus (WWC) (Shaoul and Westbury, 2010). For skipgram, we use Gensim (Řehůřek and Sojka, 2010) and its default settings with two exceptions:

- We leave the minimum word count at 50, but we explicitly include all words that occur in the test set of our evaluation tasks, even if they occur less than 50 times in the WWC.
- We increase the dimensionality d of the embedding space; the values of d chosen for each experiment are mentioned below.

For experiments in which we use fastText, we use the default parameters of the implementation by Bojanowski et al. (2017). To evaluate the Mimick model by Pinter et al. (2017), we use their implementation and keep the default settings.

To obtain training instances for the attentive mimicking model, we use the same setup as Schick and Schütze (2019): we use only words occurring at least 100 times in the WWC and if a word w has a total of $f(w)$ occurrences, we train on it $n(w)$ times for each epoch, where

$$n(w) = \min(\lfloor \frac{f(w)}{100} \rfloor, 5) \, .$$

We restrict each context of a word to at most 25 words on its left and right, respectively. While Schick and Schütze (2019) use a fixed number of 20 contexts per word during training, we instead randomly sample between 1 and 64 contexts. We do so for both the form-context model and the attentive mimicking model as we found this modification to generally improve results for both models. For all experiments, we train both the form-context model and the attentive mimicking model for 5 epochs using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.01 and a batch size of 64.

VecMap

The test set for the VecMap evaluation was created using the following steps:

1. We sample 1000 words from the lowercased and tokenized WWC that occur at least 1000 times therein, contain only alphabetic characters and at least two characters.
2. We evenly distribute the 1000 words into 8 buckets $B_0, \ldots, B_7$ such that each bucket contains 125 words.
3. We downsample each word w in bucket B_i to exactly 2^i randomly chosen occurrences.

For the variants of AM and FCM where the downsampled words are included in the training set, in every epoch we construct 5 training pairs $(w, \mathcal{C}_1), \ldots, (w, \mathcal{C}_5)$ for each downsampled word w. For training of both skipgram and fastText, we use 400-dimensional embeddings.

Sentiment Dictionary

To obtain the training set for the Sentiment Dictionary evaluation, we fuse Opinion lexicon (Hu and Liu, 2004) and the NRC Emotion lexicons (Mohammad and Turney, 2013) and remove all words that occur less than 100 times in the WWC corpus. From the SemEval2015 Task 10E data set, we remove all non-alphanumeric characters and all words that have less than 2 letters. We do so as the test set contains many hashtags, giving an unfair disadvantage to our baseline skipgram model as it makes no use of surface-form information.

We use 300-dimensional embeddings and train the logistic regression model for 5 epochs using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.01.

Name Typing

We use the same setup as for the Sentiment Dictionary experiment. That is, we use 300-dimensional embeddings and train the logistic regression model for 5 epochs using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.01.

Chimeras

Following Herbelot and Baroni (2017), we use 400-dimensional embeddings for the Chimeras task.

B Significance Tests

We perform significance tests for the results obtained on both the VecMap and the Name Typing dataset.

For VecMap, given two models m_1 and m_2, we count the number of times that the embedding assigned to a word w by m_1 is closer to the gold embedding of w than the embedding assigned by

model	skipgram	fastText	Mimick	FCM	AM	FCM†	AM†
skipgram	–	64,128	2,4,8,16,32,64,128	32,64,128	32,64,128	–	–
fastText	1,2,4,8,16	–	1,2,4,8,16,32,64,128	1,128	1,128	1,2,4	1,2,4
Mimick	–	–	–	–	–	–	–
FCM	1,2,4,8,16	8	1,2,4,8,16,32,64,128	–	–	1,2,4,8	1,2,4,8
AM	1,2,4,8,16	8	1,2,4,8,16,32,64,128	1,4	–	1,2,4,8,16	1,2,4,8,16
FCM†	1,2,4,8,16	32,64,128	1,2,4,8,16,32,64,128	32,64,128	32,64,128	–	–
AM†	1,2,4,8,16,64	32,64,128	1,2,4,8,16,32,64,128	32,64,128	32,64,128	2,4,8,32,64,128	–

Table 1: Significance results for the VecMap evaluation. Each cell lists the numbers of word occurrences for which the model of the row performs significantly better than the model of the column ($p < 0.05$). For example, FCM is significantly better than skipgram for 1, 2, 4, 8 and 16 contexts.

model	skipgram	fastText	Mimick	FCM	AM	AM+skip
skipgram	–	f_4,f_5,f_6	f_2,f_3,f_4,f_5,f_6	f_5,f_6	f_5,f_6	–
fastText	f_0,f_1,f_2	–	f_0,f_1,f_2,f_3,f_4,f_5,f_6	f_5,f_6	–	–
Mimick	–	–	–	–	–	–
FCM	f_0,f_1,f_2,f_3	f_0,f_1,f_2,f_3	f_0,f_1,f_2,f_3,f_4,f_5,f_6	–	–	–
AM	f_0,f_1,f_2,f_3	f_0,f_1,f_2,f_3,f_4	f_0,f_1,f_2,f_3,f_4,f_5,f_6	f_4,f_5,f_6	–	–
AM+skip	f_0,f_1,f_2,f_3,f_4,f_6	f_0,f_1,f_2,f_3,f_4,f_5,f_6	f_0,f_1,f_2,f_3,f_4,f_5,f_6	f_4,f_5,f_6	f_4,f_5,f_6	–

Table 2: Significance results for the Name Typing task. Each cell lists the frequency intervals for which the model of the row performs significantly better than the model of the column ($p < 0.05$) with regards to micro accuracy. We use abbreviations $f_i = [2^i, 2^{i+1})$ for $0 \leq i \leq 5$ and $f_6 = [1, 100]$.

m_2; we do so for each number of occurrences separately. Based on the so-obtained counts, we perform a binomial test whose results are shown in Table 1. As can be seen, both FCM and AM perform significantly better than the original skipgram embeddings for up to 16 contexts, but the difference between FCM and AM is only significant given one or four contexts. However, for the variants that include downsampled words during training, AM† (using attention) is significantly better than FCM† (without attention) given more than one context.

For the Name Typing dataset, we compare models based on their micro accuracy, ignoring all dataset entries for which both models perform equally well. Again, we consider all frequency ranges separately. Results of the binomial test for significance can be seen in Table 2. The best-performing method, AM+skip, is significantly better than skipgram, fastText and Mimick for almost all frequency ranges. AM is significantly better than FCM only when there is a sufficient number of contexts.