Disambiguated skip-gram model
Appendix

A Training details

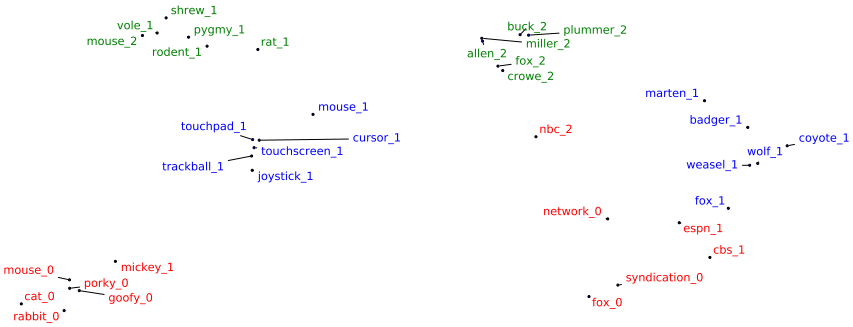
We trained disambiguated skip-gram models on the Westbury Lab Wikipedia corpus (Shaoul and Westbury, 2010). We used standard text preprocessing, i.e., converted all words to lower case, removed stop words and converted all numbers to a unique token. We then removed all words with less than 100 occurrences in the text. This gives a vocabulary with approximately 130 thousand words. We define the context of a word w as the 5 words preceding w and the 5 words following w in X. However, we do not allow the context to cross sentence boundaries, but instead pad the sentences with a unique token.

We optimize our models using mini-batch stochastic gradient descend with momentum. All models are trained with a mini-batch size of 128 examples. We perform three epochs of training, with a learning rate of 0.1, 0.05 and 0.01, respectively. During the first epoch we linearly decrease the temperature τ (Eq. 7) from 1.0 to 0.5. In the subsequent epochs the temperature is kept at $\tau = 0.5$.

To facilitate training of disambiguated skip-gram over large vocabularies, we approximate the log-likelihood gradient in Eq. 10 with importance sampling (Cho et al., 2015). Two sets of parameters in disambiguated skip-gram perform similar functions: the output embedding vectors $\mathbf{u}_d$ represent the context words in the word prediction model and the context embedding vectors $\mathbf{r}_d$ represent the context words in the sense disambiguation model. In practice, we tie the context embedding vectors $\mathbf{r}_d$ to the output embedding vectors $\mathbf{u}_d$. We found that this simplifies training of our models. Finally, to speed up the fitting of embeddings for rare words, we initially tie sense disambiguation vectors $\mathbf{q}_{ds}$ between all words $d \in D$. In order to obtain a balanced initial distribution of senses we use a small negative entropy cost ($\gamma = -0.1$) in this phase. After the initialization, we untie sense disambiguation vectors and let each word fit its own disambiguation model.

B Visualization of sense embedding vectors

In Fig. 1 we visualize nearest neighbours of sense embedding vectors of two ambiguous words. Note that the sense representations learned by disambiguated skip-gram are well separated in the embedding space.



(a) mouse (b) fox

Figure 1: Two-dimensional projections of nearest neighbors of sense embedding vectors learned by disambiguated skip-gram for two ambiguous words. We used principal component analysis to project the embedding vectors onto a plane.

C Word-similarity experiments

For completeness, we also evaluated disambiguated skip-gram in a contextual word similarity task. To this end, we carried out experiments on the Word Similarity in Context (SCWS) dataset (Huang et al., 2012). The dataset consist of 2,003 word pairs. Each word in a pair comes with a context in which it occurred and each word pair comes with a human similarity rating. The goal of this task is to estimate the word similarity given the word pair and the corresponding contexts.

For this evaluation we trained a 300-dimensional disambiguated skip-gram model with three sense embedding vectors allocated to each word and no entropy cost. Using the SCWS contexts we then inferred senses for all words in the test pairs. Finally, for each word pair we calculated the cosine similarity between the inferred word sense embedding vectors. For the performance metric in this experiment we use the Spearman's rank correlation coefficient between the human word similarity ratings and the word similarities estimated from multi-sense word embeddings. In previous works this metric is called *MaxSimC*. Results for disambiguated skip-gram and several baseline algorithms are reported in Tab. 1. Results for all baseline algorithms are taken from (Bartunov et al., 2016).

Disambiguated skip-gram achieves second best result among multi-sense word embedding methods, outperforming AdaGram, MSSG and NP-MSSG. That said, as pointed out by Bartunov et al. (2016), the task is dominated by a well trained single-sense skip-gram model. Note that the performance of multi-sense models under the *MaxSimC* metric depends on both the accuracy of the sense disambiguation, which is a difficult task, and the quality of the learned sense vectors. Skip-gram, on the other hand, is directly optimizing the average case, i.e. the semantic similarity of word vectors irrespective of their senses. To illustrate this point, we took the sense embedding vectors learned by disambiguated skip-gram and calcu-

Model	MaxSimC [%]
Skip-gram	**65.2**
MSSG	57.3
NP-MSSG	59.8
MPSG	63.6
AdaGram	53.8
Disambiguated skip-gram	62.0

Table 1: Spearman's rank correlation coefficients between the human word similarity ratings and the word similarities estimated from multi-sense word embeddings. Results for all models except disambiguated skip-gram are taken from (Bartunov et al., 2016).

lated mean vector for each word. These mean vectors outperform vanilla skip-gram in the SCWS benchmark, achieving rank correlation coefficient of 67.5. This suggests that the similarity in context task favors single-sense models, and therefore it may be not the best choice for evaluating multi-sense word embedding methods.

References

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 130–138.

Sébastien Jean Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1–10. Association for Computational Linguistics.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Cyrus Shaoul and Chris Westbury. 2010. The Westbury lab Wikipedia corpus.