Breaking Sticks and Ambiguities with Adaptive Skip-gram (supplementary material)

Sergey Bartunov
National Research University
Higher School of Economics[†]
Moscow, Russia

Dmitry Kondrashkin
Yandex
Moscow, Russia

Anton Osokin
INRIA – Sierra Project-Team,
École Normale Supérieure
Paris, France

Dmitry P. Vetrov[†]
Skolkovo Institute of
Science and Technology
Moscow, Russia

1 Full evaluation on WSI task

We report V-measure and F-score values in tables 1 and 2. The results are rather contradicting due to the reasons we described in the paper: while V-measure prefers larger number of meanings F-score encourages small number of meanings. We report these numbers in order to make the values comparable with other results.

2 Full evaluation of SemEval-2013 Task-13

This task evaluates Word Sense Induction systems by performing fuzzy clustering comparison, i.e. in the gold standard each context could be assigned to several meanings with some score indicating confidence of the assignment. Two metrics were used for comparing such fuzzy clusterings: Fuzzy Normalized Mutual Information and Fuzzy-B-Cubed which are introduced in (Jurgens & Klapaftis, 2013). Fuzzy-NMI measures the alignment of two clustering and it is independent of the cluster sizes. It is suitable to measure how well the model captures rare senses. On the contrary Fuzzy-B-Cubed is sensitive to the cluster sizes. So it reflects the performance of the system on a dataset where the clusters have almost the same frequency. Results of MSSG, NP-MSSG and AdaGram models are shown in Table 3.

However these measures have similar drawbacks as V-Measure and F-score described above. Trivial solution like assigning one sense per each context obtains high value of Fuzzy-NMI while treating each word as single-sense one performs well in terms of Fuzzy-B-Cubed. All WSI systems participated in this task failed to completely surpass these baselines according to the Table 3 in (Jurgens & Klapaftis, 2013). Hence we consider ARI comparison as more reliable. Note that since we excluded multi-token words from the evaluation the numbers we report are not comparable with other results made on the dataset.

The ARI comparison we report in the paper was done by transforming fuzzy clusterings into hard ones, i.e. each context was assigned to most probable meaning.

3 WWSI Dataset construction details

Similarly to (Navigli & Vannella, 2013) we considered Wikipedia's disambiguation pages as a list of ambiguous words. From that list we have selected target single-term words which had occurred in the text at least 5000 times to ensure there is enough training contexts in Wikipedia to capture different meanings of a word (note, however, that all models were trained on earlier snapshot of Wikipedia). We also did not consider pages belonging to some categories such as "Letter-number_combination_disambiguation_pages" as they did not contain meaningful words. Then we prepared the sense inventory for each word in the list using Wikipedia pages with names matching to the pattern "WORD_(*)" which is used as convenient naming of specific word meanings. Again, we applied some automatic filtering to remove names of people and geographical places in order to obtain more coarse-grained meanings. Finally for each page selected on the previous step we find all occurrences of the target word on it and use its 5-word neighbourhood (5 words on the left and 5 words on the right) as a context. Such size of the context was chosen to minimize the intersection between adjacent contexts but still provide enough words for disambiguation. 10-word context results into average intersection of 1.115 words.

The list of the categories pages belonging to which were excluded during target word selection is following:

- Place_name_disambiguation_pages
- Disambiguation_pages_with_surname-holder_lists
- Human_name_disambiguation_pages
- Lists_of_ambiguous_numbers
- Disambiguation_pages_with_given-name-holder_lists
- Letter-number_combination_disambiguation_pages
- Two-letter_disambiguation_pages
- Transport_route_disambiguation_pages

- Temple_name_disambiguation_pages
- and also those from the categories which name contains one of the substrings: “cleanup”, “people”, “surnames”

During the sense inventory collection we do not consider pages which name contains one of the following substrings: “tv_”, “series”, “movie”, “film”, “song”, “album”, “band”, “singer”, “musical", “comics"; and also those from the categories with names containing geography terms “countries”, “people”, “province”, “provinces”.

4 Experiments on contextual word similarity

In this section we compare AdaGram to other multi-prototype models on the contextual word similarity task using the SCWS dataset proposed in (Huang et al., 2012). The dataset consists of 2003 pairs of words each assigned with 10 human judgements on their semantic similarity. The common evaluation methodology is to average these 10 values for each pair and measure Spearman’s rank correlation of the result and the similarities obtained using word representations learned by a model, i.e. by a cosine similarity of corresponding vectors.

There are two measures of word similarity based on context: expected similarity of prototypes with respect to posterior distributions given contexts

$$AvgSimC(w_1, w_2) = \frac{1}{K_1}\frac{1}{K_2}\cdot$$
$$\sum_{k_1}\sum_{k_2} p(k_1|w_1, C_1)p(k_2|w_2, C_2)\cos(vec(w_1, k_1), vec(w_2, k_2)),$$

and similarity of the most probable prototypes given contexts

$$MaxSimC(w_1, w_2) = \cos(vec(w_1, k_1), vec(w_2, k_2)),$$

where $k1 = \arg\max_k p(k|w_1, C_1)$ and $k2 = \arg\max_k p(k|w_2, C_2)$, correspondingly. Here we define K_1 and K_2 as the number learned prototypes for each of the words and C_1, C_2 as their corresponding contexts. In AdaGram $vec(w, k) = In_{wk}$ and the posterior distribution over word senses is computed according to sec. 3.2. For word disambiguation AdaGram uses 4 nearest words in a context. Results for NP-MSSG and MSSG are taken from (Neelakantan et al., 2014) and results for MPSG – from (Tian et al., 2014).

We also consider the original Skip-gram model as a baseline. We train two models: the first one with prototypes of dimensionality 300 based on hierarchical soft-max and the second one of dimensionality 900 trained using negative sampling (Mikolov et al., 2013) (number of negative samples is set to 5, three iterations over training data are made). The training data and other parameters are identical to the training of AdaGram used in main experiments. Note that for Skip-gram measures $AvgSimC$ and $MaxSimC$ coincide because the model learns only one representation per word.

The results on the experiment are provided in table 4. NP-MSSG model of Neelakantan et al. (2014) outperforms other models in terms of $AvgSimC$, however, one may see that the improvement over 900-dimensional Skip-Gram baseline is only marginal, moreover, the latter is the second best model despite ignoring the contextual information and hence being unable to distinguish between different word meanings. This may suggest that SCWS is of limited use for evaluating multi-prototype word representation models as the ability of differentiating between word senses is not necessary to achieve a good score. One may consider another example of an undesirable model which will not be penalized by the target metric in the world similarity task. That is, if a model learned too many prototypes for a word, e.g. with very close vector representations it is hardly usable in practice, but as long as averaged similarities between prototypes correlate with human judgements such non-interpretability will not be accounted during evaluation. We thus consider word-sense induction as a more natural task for evaluation since it explicitly accounts for proper and interpretable mapping from contexts into discovered word meanings.

5 Discussion on hyperparameter α

As mentioned by the anonymous reviewer, generally setting the hyperparamter α equal for all the words may lead to poor results, especially at extreme values of α. Ideally the hyperparameter should be learned for each word independently. One of the reasonable solutions would be to put a mixture prior on it which makes a word either have strongly one sense or to allow more of them to be learned. One may also imagine a more complex prior which somehow takes into account word frequency statistics or models the semantic resolution hierarchically.

Our preliminary experiments in which we simply assigned individual α_w to each word w and optimized the variational lower bound with respect to these parameters have always led to increasing α_w over training, so a more sophisticated approach such as mentioned above should be considered.

References

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

Jurgens, David and Klapaftis, Ioannis. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second joint conference on lexical and computational semantics (* SEM)*, volume 2, pp. 290–299, 2013.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and

Table 1: V-Measure for word sense induction task for different datasets. Here we use the test subset of WWSI dataset.

MODEL	SEMEVAL-2007	SEMEVAL-2010	SEMEVAL-2013	WWSI
MSSG.300D.30K	0.067	0.144	0.033	0.215
NP-MSSG.50D.30K	0.057	0.119	0.023	0.188
NP-MSSG.300D.6K	0.073	0.089	0.033	0.128
AdaGram.300D $\alpha = 0.15$	**0.114**	**0.200**	**0.192**	**0.326**

Table 2: F-Score for word sense induction task for different datasets. Here we use the test subset of WWSI dataset.

MODEL	SEMEVAL-2007	SEMEVAL-2010	SEMEVAL-2013	WWSI
MSSG.300D.30K	0.528	0.492	**0.437**	0.632
NP-MSSG.50D.30K	0.496	0.488	0.392	0.621
NP-MSSG.300D.6K	**0.557**	**0.531**	0.419	**0.660**
AdaGram.300D $\alpha = 0.15$	0.448	0.439	0.342	0.588

Table 3: Fuzzy Normalized Mutual Information and Fuzzy B-Cubed metric values for task-13 of Semeval-2013 competition. See text for details.

MODEL	FUZZY-NMI	FUZZY-B-CUBED
MSSG.300D.30K	0.070	0.287
NP-MSSG.50D.30K	0.064	0.273
NP-MSSG.300D.6K	0.063	**0.290**
AdaGram.300D $\alpha = 0.15$	**0.089**	0.132

phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

Navigli, R. and Vannella, D. SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *SemEval*, pp. 193–201, 2013.

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *EMNLP*, 2014.

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., and Liu, T.-Y. A probabilistic model for learning multi-prototype word embeddings. In *COLING*, pp. 151–160, 2014.

Table 4: Spearman's rank correlation results for contextual similarity task on SCWS dataset. Numbers are multiplied with 100.

MODEL	$AvgSimC$	$MaxSimC$
MSSG.300D.30K	**69.3**	57.26
NP-MSSG.50D.30K	66.1	50.27
NP-MSSG.300D.6K	69.1	59.8
MPSG.300D	65.4	63.6
Skip-Gram.300D	65.2	65.2
Skip-Gram.900D	68.4	**68.4**
AdaGram.300D $\alpha = 0.15$	61.2	53.8