Non-Linear Relational Information Probing in Word Embeddings

Anonymous ACL submission

Abstract

Pre-trained word embeddings such as SkipGram and GloVe are known to contain a myriad of useful information about words. In this work, we use multilayer perceptrons (MLP) to probe the relational information contained in these word embeddings. Previous studies that use linear models on the analogy and relation induction tasks have shown that SkipGram generally outperforms GloVe, suggesting that SkipGram embeddings contain more relational information than GloVe embeddings. However, by using non-linear probe like MLP, our results instead suggest that GloVe embeddings contain more relational information than SkipGram embeddings, but a good amount of that is stored in a non-linear form and thus previous linear models failed to reveal that. Interpreting our relation probes using post-hoc analysis provides us with an explanation for this difference.[1]

1 Introduction

Word embeddings (Mikolov et al., 2013a; Pennington et al., 2014; Gladkova et al., 2016; Vylomova et al., 2016) obtained from pre-trained language models (LMs) have completely transformed the face of NLP research. When trained on a large corpus, the resultant embeddings have been shown to capture various degrees of semantic and syntactic information from the corpus. Such information has remarkably benefited a wide range of NLP tasks (Wang et al., 2019) such as dependency parsing (Chen et al., 2014; Ouchi et al., 2016; Shen et al., 2014), sentiment analysis (Yu et al., 2017; Sharma et al., 2017), question answering (Zhou et al., 2016; Hao et al., 2017), tagging (Wang et al., 2016), and text summarization (Rossiello et al., 2017; Nallapati et al., 2016; Daðason et al., 2021).

A popular method to gauge the quality of non-contextualized word embeddings is the analogy task (Mikolov et al., 2013b; Drozd et al., 2016; Levy and Goldberg, 2014; Levy et al., 2015). We define an ordered pair of words (s, o) (s:subject and o:object) which have a relation r between them. Given (s, o) and another word s', the analogy task is to correctly identify o' such that (s', o') also have relation r. Recently, a variant of this task, relation induction, has been explored (Vylomova et al., 2016; Bouraoui et al., 2018), where we predict whether an ordered pair (s, o) has the relation r or not. Existing solutions for analogy and relation induction tasks (Mikolov et al., 2013a; Drozd et al., 2016; Bouraoui et al., 2018) rely on linear features like vector offset $\vec{s} - \vec{o}$. In addition, existing methods either involve no training or are linear in nature. We believe substantial information may be encoded in non-linear form in word embeddings and that the linear nature of existing methods limits their analysis. We therefore propose an MLP (multilayer perceptron) based model as a probe for the task of relation induction on non-contextualized word embeddings SkipGram (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), and conduct comprehensive investigation for the relational information contained in these embeddings.

Our first contribution is showing that non-linear supervised training using MLPs leads to higher relation induction performance for non-contextualized word embeddings, suggesting that a good portion of relational information is stored in non-linear form. Contrary to existing state-of-the-art methods for analogy task (Drozd et al., 2016) and relation induction task (Bouraoui et al., 2018), our results show that GloVe embeddings contain more relational information than SkipGram. Using a fine-grained analysis we find that GloVe embeddings are particularly richer in Encyclopedic relations that require factual knowledge, and this additional knowledge is stored in non-linear form. Finally, a post-hoc analysis on the learned probes suggests that relational information may be contained in different forms for different relations.

[1] Both code and data will be released upon acceptance.

2 Model

2.1 Baselines

We choose four state-of-the-art relation induction models to compare with our proposed probe. 3CosAverage (Drozd et al., 2016) (3CA) produces the following score, where (s_i, o_i) are training examples for a relation:

$$score_{3CA}(s,o) = cos(\vec{s} + \frac{\sum_{i=1}^{N}(\vec{o}_i - \vec{s}_i)}{N}, \vec{o}).$$

LRCos (Drozd et al., 2016) extends the 3CA model by checking how well s and o belong to their respective word-groups using logistic regression classifier. Trans and Regr (Bouraoui et al., 2018) are probabilistic models. Trans learns a Gaussian distribution over $\vec{s} - \vec{o}$, and Regr uses Bayesian linear regression to learn linear mappings between $\vec{s}$ & $\vec{o}$.

2.2 MLP-Based Relation Probe

Existing relation induction models (Vylomova et al., 2016; Bouraoui et al., 2018) have two potential restrictions that limit their applicability as relation probes. First, they are designed around the vector difference between s and o embeddings. We speculate that alternate vector operations could also contain complementary relational information. Second, existing approaches are either linear (Weeds et al., 2014; Bouraoui et al., 2018; Vylomova et al., 2016) or require no supervised training at all (Bouraoui et al., 2018). This limits the models' ability to decode the relational information contained in word embeddings, which may be encoded in non-linear ways.

These limitations motivated the design of our relation probe. Given a relation r and word pair (s,o), our model generates the confidence score by:

$$score_{s,o,r} = \sigma(ReLU(x_{s,o}W_1^r + b_1^r)W_2^r + b_2^r),$$

where W_1^r, b_1^r, W_2^r and b_2^r are learnable parameters for relation r.[2] The input to the MLP, $x_{s,o}$, is a concatenation of features derived from the word embeddings $\vec{s}$ and $\vec{o}$. We create two models that differ in the input features. The first model (RLProbe) consists of three features $\vec{s}$, $\vec{o}$ and $\vec{s} - \vec{o}$. Since the baseline models rely on these three features, RLProbe lets us compare the performance gains that our MLP achieves over the baselines. The second model (RLProbe+) contains two additional features $\vec{s} + \vec{o}$ and $\vec{s} \odot \vec{o}$.[3] These additional features act as inductive bias during training, and lead to performance gains as seen in our experiments.

3 Datasets

We use two popular datasets, Google Analogy Test Set (Mikolov et al., 2013a) and Bigger Analogy Test Set (BATS) (Gladkova et al., 2016), for English language. The Google Test Set contains 14 relations (5 semantic and 9 syntactic). BATS contains 40 relations with around 50 word pairs per relation and is generally considered more comprehensive and challenging than the Google set. The BATS relations are further divided into four categories: Lexicographical (e.g. "antonyms"), Encyclopedic (e.g. "country-capital"), Derivational (e.g. "verb+er") and Inflectional (e.g. "singular-plural"). We use the standard public version of two pre-trained LMs, GloVe and SkipGram. The SkipGram model is trained on the Google News Corpus (100B tokens) and the GloVe model is trained on the Common Crawl (840B tokens). Since the relation induction datasets only contain positive instances, we generate negative instances for each positive instance using the same negative sampling strategy as in Bouraoui et al. (2018). We conduct 10-fold cross-validation on each relation and report macro F_1 score averaged over the relations for evaluation (details in Appendix).

4 Experimental Results

4.1 Relation Probing

Table 1 shows the macro F_1 scores obtained on different dataset and embedding configurations. Our probes outperform the other approaches on three of the four configurations, setting a new state of the art on these datasets for non-contextualized word embeddings. This shows the effectiveness of non-linear probes at detecting the relational knowledge contained in word embeddings, suggesting that a good portion of this information might be stored in non-linear form. We note that RLProbe+ performs better than RLProbe. This supports our hypothesis that additional features provide complementary information required for decoding the relational information in word embeddings. For the rest of this section, we discuss our results on RLProbe+.

Comparing the probe's performance on GloVe and SkipGram embeddings, we find that GloVe sig-

[2] We provide the training details in the Appendix.

[3] $\vec{s} \odot \vec{o}$ denotes element-wise or Hadamard product.

	Google-SkipGram	BATS-SkipGram	Google-GloVe	BATS-GloVe
3CA	52.2	46.9	65.3	50.3
LRCos	56.5	55.5	51.8	46.6
Regr	64.6	46.2	62.9	41.9
Trans	**75.6**	68.2	72.8	62.4
RLProbe	63.4 ±1.5	68.9 ±0.3	74.1 ±1.5	71.3 ±0.5
RLProbe+	68.0 ±2.5	**71.7** ±0.6	**80.0** ±1.3	**76.0** ±0.6

Table 1: Macro F_1 scores for all the 4 configurations. 3CA, LRCos, Trans and Regr scores are copied from (Bouraoui et al., 2018). RLProbe and RLProbe+ are averaged over 5 runs with different random seeds

.

	Google Dataset		BATS			
	Semantic	Syntactic	Encyclopedic	Lexicographical	Derivational	Inflectional
SkipGram	55.8	74.2	55.8	73.9	71.6	82.3
GloVe	**74.8**	**82.8**	**65.8**	**74.7**	**77.7**	**84.9**

Table 2: RLProbe+ macro F_1 scores for SkipGram and GloVe embeddings on 4 BATS and 2 Google Set categories.

nificantly outperforms SkipGram on both Google and BATS. Table 2 further shows the macro F_1 scores obtained on fine-grained relation categories. GloVe achieves higher macro F_1 scores on all the 4 BATS groups, but the major performance gain comes from Encyclopedic relations. On Google dataset, GloVe embeddings perform better for the semantic relations by a large margin. We find that the top-performing semantic relations are again encyclopedic relations such as "country-capital" and "city-state". This suggests that GloVe embeddings are especially richer than SkipGram in their knowledge about such encyclopedic relations. Contrary to this, the probabilistic models (Bouraoui et al., 2018) show the opposite trend, that GloVe embeddings generally perform worse than the SkipGram embeddings. We attribute this to our novel observation, as further demonstrated in the feature occlusion analysis later, that more relational information is encoded in non-linear ways in GloVe which is hard for linear models to detect.

4.2 Feature Occlusion Analysis

Now we investigate how sensitive RLProbe+ is to each of the five features. We use occlusion analysis (Bastings and Filippova, 2020) to compute the sensitivity of RLProbe+'s performance to the features. For a relation, each feature is occluded (individually) and the new macro F_1 is computed on the test set. To occlude a feature, the feature vector is replaced with the zero vector (leaving the other 4 features untouched). Thus we obtain 5 new macro F_1 scores ($F_1^{occlusion}$) for each relation. The sensitivity of a relation on a feature is defined as:

$$\Delta = \frac{F_1^{original} - F_1^{occlusion}}{F_1^{original}}.$$

Figure 1 shows the pie charts obtained for RLProbe+ trained on BATS data using GloVe embeddings.[4] Each pie chart corresponds to a relation in the BATS dataset. The size of the pie charts corresponds to the maximum Δ for that relation, and arc angles correspond to the normalized Δ values for the features. Each relation can be seen to have its characteristic sensitivity pattern. We notice that the majority of relations are most sensitive to feature $\vec{s} - \vec{o}$ (■), which is expected since the offset-based methods (3CA, LRCos, Trans) perform reasonably well. The second most salient feature is $\vec{s} \odot \vec{o}$ (■). We find this to be a unique characteristic of GloVe embeddings and is missing in the SkipGram embeddings. We also find that the Encyclopedic and Derivational Morphology relations (rows 1 & 2 in Figure 1) are affected most by this feature, suggesting that a significant amount of such relational information is encoded in GloVe in a non-linear way. Using these additional features in downstream relational tasks such as relation extraction could be a promising way to improving performance.

5 Related Work

Learning relational information: Weeds et al. (2014) compare different vector operations in their ability to identify hypernym and co-hyponym relations. Vylomova et al. (2016) use linear SVM

[4] Pie charts for other configurations are in the Appendix.

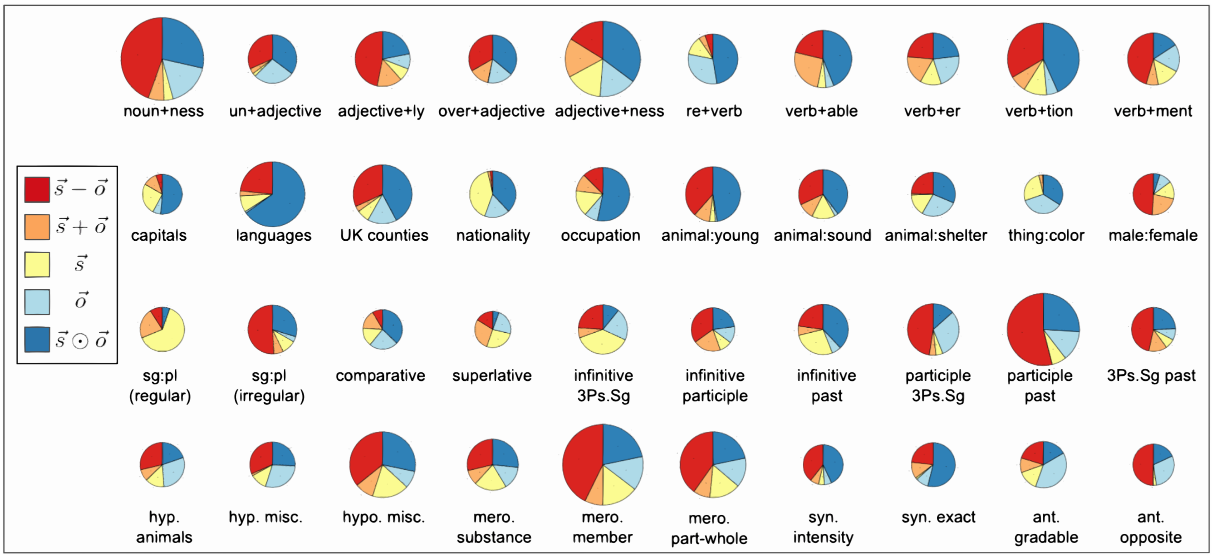

Figure 1: Pie charts for the occlusion analysis. RLProbe+ is trained on BATS relations using Glove embeddings[4].

on vector offset for analyzing limited relations. Bouraoui et al. (2018) propose probabilistic models for relation induction. Their best model uses vector offset. We use relation induction as a probing task and do not limit the model to vector offset. Many works (Jameel et al., 2018; Joshi et al., 2019; Camacho-Collados et al., 2019) train relation-specific embeddings. We focus on publicly available PLMs. Bouraoui et al. (2020) propose relation induction tasks for contextualized embedding models. Our focus is on non-contextualized models. In summary our work uses a learnable probe with additional linear and non-linear features missing in previous studies. The additional features prove to useful (especially for Glove) in our results.
Probing word embeddings: Liu et al. (2019) probe token-pair semantic/syntactic dependency arc relationships. We focus on semantic and morphological relations in word-pairs. Alain and Bengio (2016) use linear classifiers to understand intermediate layers in models. Belinkov et al. (2017b,a) probe different layers of neural machine translation models for linguistic properties target-language specific information. Conneau et al. (2018) propose 10 linguistic probing tasks for sentence embeddings. Hewitt and Manning (2019) proposed a structural probe for parse tree information. We solve relation induction using convectional probing paradigm of supervised learning using MLPs.

6 Discussion and Conclusion

Vector offset-based analogy evaluation has been known to work better on SkipGram when compared to GloVe (Xun et al., 2017), and this has been attributed to the type of information learned by these embedding models. GloVe captures the global context information whereas SkipGram captures the local context. We show that using non-linear relation induction models leads to opposite trend. GloVe outperforms SkipGram on two popular datasets by achieving higher macro F1. Recent works that use word embeddings for downstream tasks like question answering (Kamath et al., 2017), query answering (Frąckowiak et al., 2017), word similarity (Liu et al., 2015), tagging (Wang et al., 2016; Chiu and Nichols, 2016) have found GloVe embeddings to outperform SkipGram. Our investigation provides a plausible explanation for these previous empirical findings, especially for tasks that require relational information between words.

We present non-linear MLPs to probe word embeddings for relational information. Contrary to existing linear relation induction approaches, our probes show that GloVe contains more relational information than SkipGram and a good amount of this relational information is stored in non-linear form. Linear approaches are largely oblivious to this. Our results show that GloVe embeddings are richer in Encyclopedic relation information than SkipGram. Our post-hoc analysis suggests this additional information in GloVe, and can be linked to the non-linear feature $\vec{s}\odot\vec{o}$. Incorporating this knowledge in relational downstream tasks could potentially lead to improvements. For future work, we plan to explore relational information stored in contextualized LMs (Bouraoui et al., 2020).

7 Ethics Statement

Many existing works (Bolukbasi et al., 2016; Zhao et al., 2018; Lauscher et al., 2020) on biases and stereotypes in word embeddings have adopted the analogical reasoning tasks for bias diagnosis. There exist many types of biases and stereotypes, e.g., gender, ethnicity, race, religion, etc. Gender stereotypes, for instance, where certain occupations or adjectives might be associated more with one specific gender. E.g. "engineer" being associated more with a man and "nurse" with a woman, or "delicate" with a woman, and "crude" with a man. Such stereotypes often seep into LMs, as these LMs are trained on data crawled from the internet. When used in applications like "Resume Scoring Algorithms", etc., this can lead to unfair NLP models.

Our repositioning of the relation induction task as a probing task could enable us to use the relation induction for stereotype diagnosis instead. Similar to the datasets discussed in this work, these stereotypes can be formulated as relations between word pairs. For example, word pairs (female, delicate) and (male, aggressive) are examples of the "gender stereotype" relation. Relation probes trained on an LM devoid of such a stereotype must predict these pairs as "Negative" (or no relation), and a "Positive" might be an indication of "gender stereotype" in the LM.

Diagnosing LMs for such stereotypes before their application to the downstream task can potentially make NLP applications fairer, as the experts can compare and choose the most fair PLM. Our relation probes could be used as a tool for such diagnosis.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2018. Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

José Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019. Relational word embeddings. *CoRR*, abs/1906.01373.

Wenliang Chen, Min Zhang, and Yue Zhang. 2014. Distributed feature representations for dependency parsing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):451–460.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jón Daðason, Hrafn Loftsson, Salome Sigurðardóttir, and Þorsteinn Björnsson. 2021. IceSum: An Icelandic text summarization corpus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–14, Online. Association for Computational Linguistics.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics:*

Technical Papers, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Michał Frąckowiak, Jakub Dutkiewicz, Czesław Jędrzejek, Marek Retinger, and Paweł Werda. 2017. Query answering to iq test questions using word embedding. In *Multimedia and Network Information Systems*, pages 283–294. Springer.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. 2018. Unsupervised learning of distributional relation vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23–33, Melbourne, Australia. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Omer Levy, Daniel Weld, and Luke Zettlemoyer. 2019. pair2vec: Compositional word-pair embeddings for cross-sentence inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3597–3608, Minneapolis, Minnesota. Association for Computational Linguistics.

Sanjay Kamath, Brigitte Grau, and Yue Ma. 2017. A study of word embeddings for biomedical question answering. In *4e édition du Symposium sur l'Ingénierie de l'Information Médicale*.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. 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: Long Papers)*, pages 1501–1511.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization.

Hiroki Ouchi, Kevin Duh, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Transition-based dependency parsing exploiting supertags. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2059–2068.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21.

Yash Sharma, Gaurav Agrawal, Pooja Jain, and Tapan Kumar. 2017. Vector representation of words for sentiment analysis using glove. In *2017 international conference on intelligent communication and computational techniques (icct)*, pages 279–284. IEEE.

Mo Shen, Daisuke Kawahara, and Sadao Kurohashi. 2014. Dependency parse reranking with rich subtree features. *IEEE/ACM transactions on audio, speech, and language processing*, 22(7):1208–1218.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2016. Learning distributed word representations for bidirectional lstm recurrent neural network. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 527–533.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Guangxu Xun, Yaliang Li, Jing Gao, and Aidong Zhang. 2017. Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–543.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):671–681.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Guangyou Zhou, Zhiwen Xie, Tingting He, Jun Zhao, and Xiaohua Tony Hu. 2016. Learning the multilingual translation representations for question retrieval in community question answering via non-negative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1305–1314.

A Data Preparation

Given the set of ordered pairs for relation r, we divide this set into train, dev, and test sets. Since we only have factual (positive) data, we generate negative examples following the approach outlined in (Bouraoui et al., 2018). For each positive pair (s,o), we add four types of negative pairs. 1.) (o,s), 2.) we sample two distinct o' from the train set such that $(s, o') \notin R_r$ (provided there exist such o') and add (s, o'), 3.) we randomly sample an ordered pair from some other relation set $R_r^{'}$, and 4.) we generate a randomly ordered pair by sampling two words from the vocabulary of the dataset. We use the same process to generate negative pairs for the dev and test sets. (For symmetric relations like 'Antonyms' and 'Synonyms', (o,s) is added as a positive sample instead of negative).

We use 10-fold cross-validation for the evaluation. The dev sets are obtained by sampling 10% of the pairs from the train splits. For relations with less than 10 pairs, a leave-one-out evaluation is performed. The same 10-fold splits used by (Bouraoui et al., 2018) are employed for a fair comparison. To evaluate the models, the relation induction problem is treated as a binary classification problem (Bouraoui et al., 2018). The performance is measured in terms of macro F_1 scores. Dev set loss is used for early stopping.

B Training Details

Table 3 shows the MLP architecture for both the probes. We use the same architecture across all relations. We add a drop-out of probability 0.2 on the hidden-layer. We experimented with different probability values and 0.2 consistently gave good results. The loss function used is binary cross entropy loss with L2 penalty. Adam optimizer with learning rate of 0.001 and SGD with batch size 16 is used for all the relations and probes. For the $\vec{s}-\vec{o}$ and $\vec{s}+\vec{o}$ features, we find that adding a ReLU layer on these features improved the performance significantly. All models are trained for maximum 50 epochs with early stopping using dev set loss.

	Model-Architecture
RLProbe	$W_r^1 : R^{900\times 75}$ $b_r^1 : R^{75}$ $W_r^2 : R^{75\times 1}$ $b_r^2 : R$
RLProbe+	$W_r^1 : R^{1500\times 75}$ $b_r^1 : R^{75}$ $W_r^2 : R^{75\times 1}$ $b_r^2 : R$

Table 3: MLP architecture details for all RLProbe and RLProbe+ models.

.

	Relation
Encyclopedic	geography: capitals geography: languages geography: UK counties people: nationality people: occupation animals: the young animals: sounds animals: shelter thing:color male:female
Lexicographic	hypernyms: animals hypernyms: miscellaneous hyponyms: miscellaneous meronyms: substance meronyms: member meronyms: part-whole synonyms: intensity synonyms: exact antonyms: gradable antonyms: opposite
Inflectional	noun sg:pl (regular) noun sg:pl (irregular) adjective: comparative adjective: superlative infinitive: 3Ps.Sg infinitive: participle infinitive: past participle: 3Ps.Sg participle: past 3Ps.Sg: past
Derivational	noun+ness un+adjective adjective+ly over+adjective adjective+ness re+verb verb+able verb+er verb+tion verb+ment

Table 4: All BATS relations grouped into the 4 categories.

.

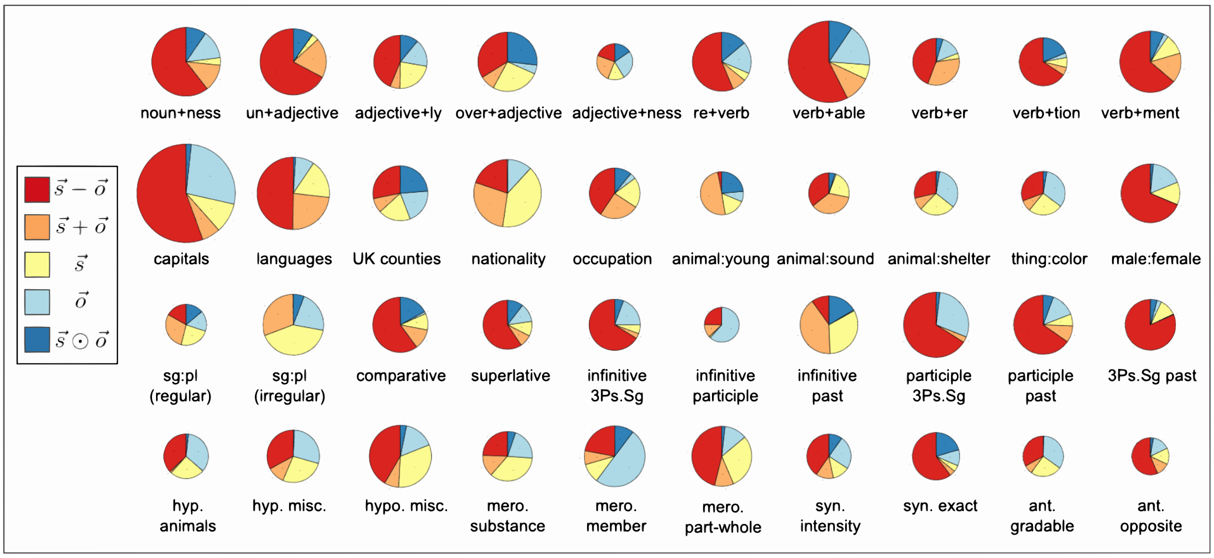

Figure 2: Pie chart visualization for the occlusion analysis. SkipGram - BATS

	Relation
Semantic	common capital city
	all capital cities
	currency
	city in state
	man-woman
Syntactic	adjective to adverb
	opposite
	comparative
	superlative
	present participle
	nationality adjective
	past tense
	plural nouns
	plural verbs

Table 5: All Google dataset relations grouped into semantic and syntactic categories.

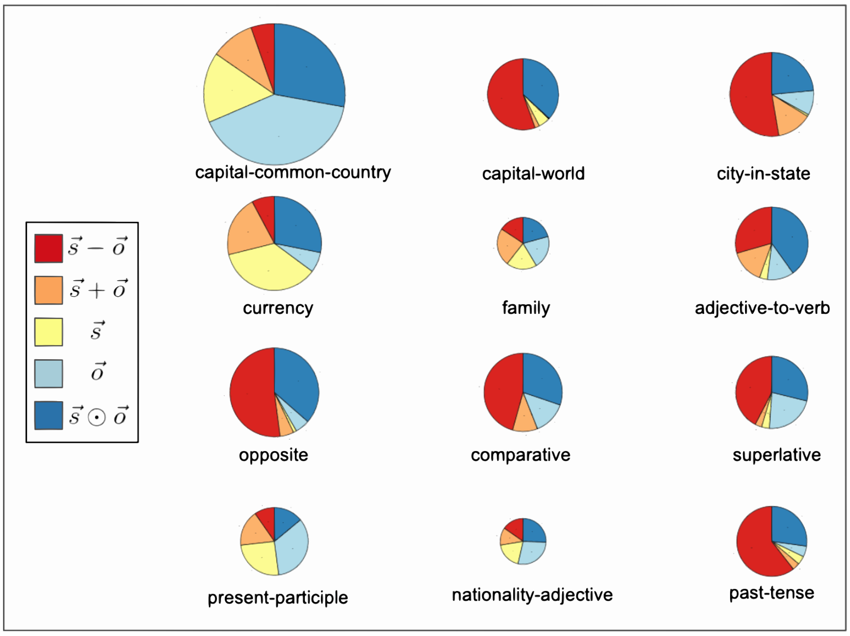

Figure 3: Pie chart visualization for the occlusion analysis. GloVe - Google dataset.

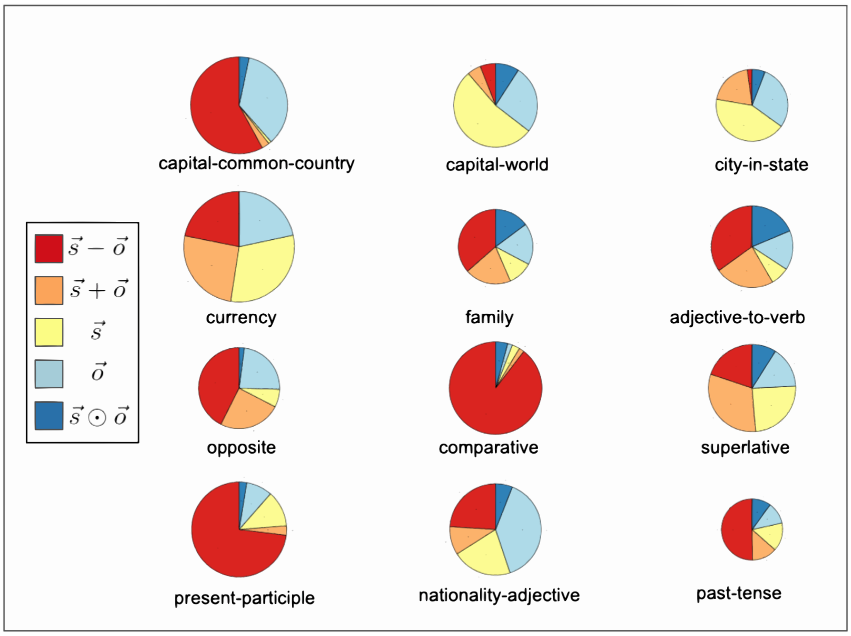

Figure 4: Pie chart visualization for the occlusion analysis. SkipGram - Google dataset.