Retrofitting Sense-Specific Word Vectors Using Parallel Text

Allyson Ettinger[1], **Philip Resnik**[1,3], **Marine Carpuat**[2,3]
[1]Linguistics, [2]Computer Science, [3]Institute for Advanced Computer Studies
University of Maryland, College Park, MD
{aetting, resnik}@umd.edu, marine@cs.umd.edu

Abstract

Jauhar et al. (2015) recently proposed to learn sense-specific word representations by "retrofitting" standard distributional word representations to an existing ontology. We observe that this approach does not require an ontology, and can be generalized to any graph defining word senses and relations between them. We create such a graph using translations learned from parallel corpora. On a set of lexical semantic tasks, representations learned using parallel text perform roughly as well as those derived from WordNet, and combining the two representation types significantly improves performance.

1 Introduction

Vector space models (VSMs) provide a powerful tool for representing word meanings and modeling the relations between them. While these models have demonstrated impressive success in capturing some aspects of word meaning (Landauer and Dumais, 1997; Turney et al., 2010; Mikolov et al., 2013; Baroni et al., 2014; Levy et al., 2014), they generally fail to capture the fact that single word forms often have multiple meanings. This can lead to counterintuitive results—for example, it should be possible for the nearest word to *rock* to be *stone* in everyday usage, *punk* in discussions of music, and *crack* (cocaine) in discussions about drugs.

In a recent paper, Jauhar et al. (2015) introduce a method for "retrofitting" generic word vectors to create *sense-specific* vectors using the WordNet semantic lexicon (Miller, 1995). From WordNet, they create a graph structure comprising two classes of relations: form-based relations between each word form and its respective senses, and meaning-based relations between word senses with similar meanings. This graph structure is then used to transform a traditional VSM into an enriched VSM, where each point in the space represents a word sense, rather than a word form. This approach is appealing as, unlike with prior sense-aware representations, senses are defined categories in a semantic lexicon, rather than clusters induced from raw text (Reisinger and Mooney, 2010; Huang et al., 2012; Neelakantan et al., 2015; Tian et al., 2014), and the method does not require performing word sense disambiguation (Guo et al., 2014).

In this paper, we observe that the crucial meaning relationships in the Jauhar et al. retrofitting process—the word sense graph—can be inferred based on another widely available resource: bilingual parallel text. This observation is grounded in a well-established tradition of using cross-language correspondences as a form of sense annotation (Gale et al., 1992; Diab and Resnik, 2002; Ng et al., 2003; Carpuat and Wu, 2007; Lefever and Hoste, 2010, and others). Using parallel text to define sense distinctions sidesteps the persistent difficulty of identifying a single correct sense partitioning based on human intuition, and avoids large investments in manual curation or annotation.

We use parallel text and word alignment to infer both word sense identities and inter-sense relations required for the sense graph, and apply the approach of Jauhar et al. to retrofit existing word vector representations and create a sense-based vec-

Proceedings of NAACL-HLT 2016, pages 1378–1383,
San Diego, California, June 12-17, 2016.

tor space, using bilingual correspondences to define word senses. When evaluated on semantic judgment tasks, the vector spaces derived from this graph perform comparably to and sometimes better than the WordNet-based space of Jauhar et al., indicating that parallel text is a viable alternative to WordNet for defining graph structure. Combining the output of parallel-data-based and WordNet-based retrofitted VSMs consistently improves performance, suggesting that the different sense graph methods make complementary contributions to this sense-specific retrofitting process.

2 Model

Retrofitting. The technique introduced by Jauhar et al. (2015) is based on what we will call a *sense graph*, which we formulate as follows. Nodes in the sense graph comprise the words w_i in a vocabulary W together with the senses s_{ij} for those words. Labeled, undirected edges include word-sense edges $\langle w_i, s_{i,j} \rangle$, which connect each word to all of its possible senses, and sense-sense edges $\langle s_{ij}, s_{i'j'} \rangle$ labeled with a meaning relationship r that holds between the two senses.

Jauhar et al. use WordNet to define their sense graph. Synsets in the WordNet ontology define the sense nodes, a word-sense edge exists between any word and every synset to which it belongs, and WordNet's synset-to-synset relations of synonymy, hypernymy, and hyponymy define the sense-sense edges. Figure 1 illustrates a fragment of a WordNet-based sense graph, suppressing edge labels.

Adopting Jauhar et al.'s notation, the original vector space to be retrofitted is defined by the original word-form vectors $\hat{u}_i$ for each $w_i \in W$, and the goal is to infer a set V of sense-specific vectors v_{ij} corresponding to each sense s_{ij}. Jauhar et al. use the sense graph to define a Markov network with variables for all word vectors and sense vectors, within which each word's vector $\hat{u}_i$ is connected to all of its sense vectors v_{ij}, and the variables for sense vectors v_{ij} and $v_{i'j'}$ are connected iff the corresponding senses are connected in the sense graph.

Retrofitting then consists in optimizing the following objective, where α is a sense-agnostic weight, and β_r are relation-specific weights for types of relations between senses:

$$C(V) = \arg\min_{V} \sum_{i-ij} \alpha \|\hat{u}_i - v_{ij}\|^2 + \sum_{ij-i'j'} \beta_r \|v_{ij} - v_{i'j'}\|^2 \quad (1)$$

The objective encourages similarity between a word's vector and its senses' vectors (first term), as well as similarity between the vectors for senses that are related in the sense graph (second term).

Defining a sense graph from parallel text. Our key observation is that, although Jauhar et al. (2015) assume their sense graph to be an ontology, this graph can be based on *any* inventory of word-sense and sense-sense relationships. In particular, given a parallel corpus, we can follow the tradition of translation-as-sense-annotation: the senses of an English word type can be defined by different possible translations of that word in another language.

Operationalizing this observation is straightforward, given a word-aligned parallel corpus. If English word form e_i is aligned with Chinese word form c_j, then $e_i(c_j)$ is a sense of e_i in the sense graph, and there is a word-sense edge $\langle e_i, e_i(c_j) \rangle$. Edges signifying a meaning relation are drawn between sense nodes if those senses are defined by the same translation word. For instance, English senses *swear*(发誓) and *vow*(发誓) both arise via alignment to 发誓 (*fashi*), so a sense-sense edge will be drawn between these two sense nodes. See Figure 2 for illustration.

3 Evaluation

Tasks. We evaluate on both the synonym selection and word similarity rating tasks used by Jauhar et al. Synonym selection nicely demonstrates the advantages afforded by sense partitioning: if we believe that *spin* means "make up a story", then we are not likely to perform well on a question in which the correct synonym is *twirl*. Word similarity rating, on the other hand, is a classic test of the extent to which vector representations simulate human intuitions of word relations in general.

For synonym selection, we follow Jauhar et al. in testing with ESL-50 (Turney, 2001), RD-300 (Jarmasz and Szpakowicz, 2004), and TOEFL-80 (Landauer and Dumais, 1997), using maxSim for multi-

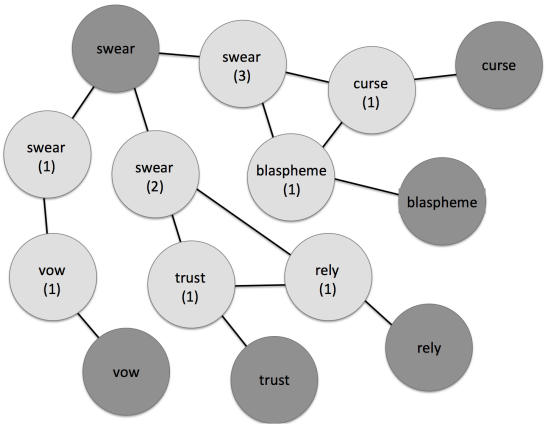

Figure 1: Illustration of WordNet-based sense graph.

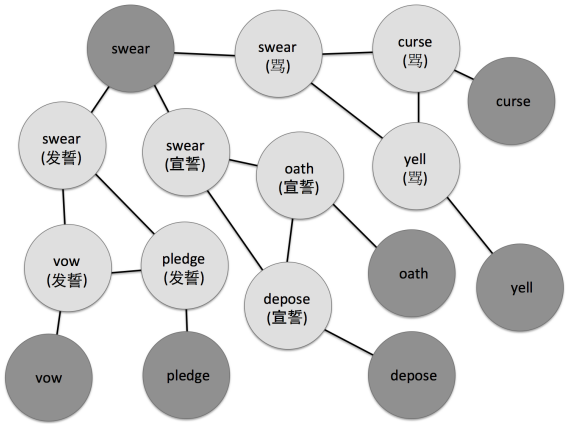

Figure 2: Illustration of parallel-text-based sense graph.

sense models (Jauhar et al., 2015, eq. 9) to select the most similar word.[1] For similarity rating, we again mirror Jauhar et al., testing with WS-353 (Finkelstein et al., 2001), RG-65 (Rubenstein and Goodenough, 1965), MC-30 (Miller and Charles, 1991), and the designated test subset (1000 items) of MEN-3k (Bruni et al., 2014), using avgSim (Jauhar et al., 2015, eq. 8) as the similarity rating, and evaluating model ratings against human similarity ratings via Spearman's rank correlation coefficient (ρ).[2]

Initial word representations. We use the word2vec (Mikolov et al., 2013) skip-gram architecture to train 80-dimensional word vectors (in keeping with Jauhar et al.), based on evidence that this model shows consistently strong performance on a wide array of tasks (Baroni et al., 2014; Levy et al., 2015). Training is on ukWaC (Ferraresi et al., 2008), a diverse 2B-word web corpus.[3]

Sense-graph construction from parallel text. To construct the sense graph per Section 2, we use $\sim$5.8M lines of segmented Chinese-English parallel text from the DARPA BOLT project and the Broadcast Conversation subset of the segmented Chinese-English parallel data in the OntoNotes corpus (Weischedel et al., 2013).[4] We perform word alignment with the Berkeley aligner (Liang et al., 2006). We filter out noisy alignments using the G-test statistic (Dunning, 1993), with a threshold selected during tuning on a development set.

We set α (see Equation 1) to 1.0. Each sense-sense edge $\langle e_i(c_j), e_{i'}(c_j)\rangle$ has individual weight $0 < \beta_r \leq 1$, computed by obtaining the G-test statistic for the alignment of e_i with c_j and for the alignment of $e_{i'}$ with c_j, running these values through a logistic function, and averaging. Parameters for these computations, as well as the G-test statistic threshold below which we filtered out noisy alignments, were selected during tuning on the development set.

Note that we have not currently incorporated special treatment for alignments of a single word to a multi-word phrase. This does create the possibility of noisy or uninformative sense annotations (e.g., sense annotations corresponding to parts of aligned Chinese phrases) when such alignments are not filtered out by the G-test thresholding.

Experimental conditions. We evaluate the following experimental conditions: Skip-gram (SG) uses the un-retrofitted word2vec vectors, Word-

[1] Because it is not clear how multi-word phrases should best be treated (and this is not a question being investigated here), we filter out any questions containing multi-word phrases for any of the relevant items (probe or possible response), and any questions for which any of the relevant items is completely out of vocabulary (no vectors available) for any of the evaluated models. This leaves 48 items in ESL, 87 items in RD, and 77 items in TOEFL.

[2] The designated development set of MEN-3k (2000 items) was used for tuning.

[3] To alleviate sparsity we lemmatized the ukWaC corpus. Runs without lemmatization produced weaker results.

[4] English was lemmatized post-alignment via lookup in the XTAG morphological database (XTAG Research Group, 2001).

Net (WN) retrofits using the WordNet-based sense graph, and Parallel Data (PD) retrofits using the sense graph built from parallel text. We also combine the two retrofitting approaches (PD-WN). For synonym selection, we compute maxSim over all sense pairs for WN and PD separately, and select the sense pair with the overall maximum cosine similarity across the two. For similarity rating, we explore two PD-WN combination approaches: for each word pair, we take the avgSim from each separate model, and then we (a) take the average of the values given by the two models (avg), or (b) take the maximum value between the two models (max).

	Synonym Selection SYMM (%)		
	ESL-48	RD-87	TOEFL-77
SG	58.3	58.6	71.4
WN	66.7	74.7	81.8
PD	68.8	62.1	80.5
PD-WN	**70.8**	**79.3**	**84.4**

Table 1: Synonym selection task results: accuracy

	Word similarity: avgSim SYMM (ρ)			
	WS-353	RG-65	MC-30	MEN-1k
SG	**.708**	.729	.722	.763
WN	.610	.725	.750	.739
PD	.636	**.777**	.715	.769
PD-WN (avg)	.666	**.777**	.742	**.773**
PD-WN (max)	.630	.731	**.758**	.756

Table 2: Similarity rating task results

4 Results

Table 1 shows that combining our new method with Jauhar et al.'s WN retrofitting performs best on synonym selection across all datasets, and both retrofitted models consistently outperform the no-retrofitting model (SG). Error analysis on RD-87, the only set on which WN substantially outperforms PD, suggests that PD's errors are driven by the large number of lower frequency items that characterize this dataset. Given that WordNet is a hand-curated lexicon while the parallel data mirrors actual usage, it is not surprising that the latter suffers when it comes to low frequency items.

Error analysis also indicates that PD performs particularly well on the synonym task precisely when one would expect: when the probe and the correct answer have an alignment to the same Chinese word form, so that the corresponding sense vectors are extremely close in vector space. Occasionally, PD yields "the wrong answer for the right reason", choosing an option for which there is indeed a correct alignment that matches an alignment of the probe word. For instance, though the probe *passage* is intended to have the answer *hallway*, PD chooses *ticket* because both *passage* and *ticket* have a sense defined by alignment to the Chinese word 机票 (*jipiao*), meaning "air ticket". Though this is a less frequent sense of *passage*, it is a reasonable one.

Results on the similarity rating task (presented in Table 2) are less clearly interpretable, top performance being divided between the PD model and the combined models—with the exception of WS-353. We note that WS-353 is a test set for which human raters were explicitly told to rate relatedness, rather than similarity, while the retrofitting process is intended to encourage similarity *per se*. If we exclude this set from consideration, we can observe that SG is outperformed by at least one sense-specific model in all cases.[5]

Note that as expected, the amount of training data has an impact on the quality of the alignments and of the sense graph. Retrofitting sense-specific embeddings using only 300k sentence pairs, which represent about 5% of the total training data, does not give clear benefit over word-form embeddings.

5 Conclusions and future work

Building on Jauhar et al. (2015), we have presented an alternative means of deriving information about senses and sense relations to build sense-specific vector space representations of words, making use of parallel text rather than a manually constructed ontology. We show that this is a viable alternative, producing representations that perform on par with those retrofitted to sense graphs based on WordNet.[6]

[5] We also explored using maxSim for similarity ratings, on the intuition that when human annotators give similarity judgments, they are likely to judge based on senses of the given words that are biased toward the words with which they are paired. However, top performance is similarly scattered when using maxSim for similarity scores and fails to improve over the SG baseline for two of the datasets.

[6] Sample sense-specific vectors and code for generating a sense graph from parallel data can be accessed at http://ling.umd.edu/~aetting/retropd.html.

Based on these results, it would be interesting to evaluate further refinements of the sense graph: alignment-based senses could be clustered, or further filtered to reduce the impact of alignment noise; new edges could be added using other multilingual resources. Finally, it will be important to evaluate the effectiveness of the retrofitted word embeddings on extrinsic tasks that require disambiguating word meaning in context.

Acknowledgments

The authors would like to thank Sujay Kumar Jauhar for sharing software and data and for helpful discussion. Thanks also to Manaal Faruqui and Peter Turney for help in acquiring evaluation datasets, to Amittai Axelrod for his assistance with data, and to the anonymous reviewers for valuable comments and suggestions. This work was supported in part by an NSF Graduate Research Fellowship under Grant No. DGE 1322106. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, volume 7, pages 61–72.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 255–262.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING*, pages 497–507.

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.

Mario Jarmasz and Stan Szpakowicz. 2004. Roget's thesaurus and semantic similarity. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2003:111.

Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL*, pages 683–693.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Els Lefever and Veronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Conference on Natural Language Learning (CoNLLL)*, pages 171–180.

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.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

George A Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 455–462.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING*, pages 151–160.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Peter Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for English. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.