Unsupervised Word Sense Induction using Distributional Statistics

Kartik Goyal
Carnegie Mellon Uniersity
kartikgo@cs.cmu.edu

Eduard Hovy
Carnegie Mellon University
hovy@cmu.edu

Abstract

Word sense induction is an unsupervised task to find and characterize different senses of polysemous words. This work investigates two unsupervised approaches that focus on using distributional word statistics to cluster the contextual information of the target words using two different algorithms involving latent dirichlet allocation and spectral clustering. Using a large corpus for achieving this task, we quantitatively analyze our clusters on the Semeval-2010 dataset and also perform a qualitative analysis of our induced senses. Our results indicate that our methods successfully characterized the senses of the target words and were also able to find unconventional senses for those words.

1 Introduction

Word Sense Induction (WSI) involves automatically determining the number of senses of a given word or a phrase and identifying the features which differentiate those senses. This task, although similar to the Word Sense Disambiguation (WSD) task, is fundamentally different because it does not involve any supervision or explicit human knowledge about senses of words. WSI has potential to be extremely useful in downstream applications because, apart from the savings on annotation costs, it also mitigates several theoretical conflicts associated with supervised WSD tasks, which generally involve deciding on the granularity of senses. Ideally, a WSI algorithm would be able to adapt to different tasks requiring different sense granularities. WSI algorithms can also be used to model the evolution of the senses of a word with time and hence can be much easier to maintain than existing fixed sense inventories like WordNet(Miller, 1995), Ontonotes(Hovy et al., 2006) etc. Automatic sense identification systems also have the potential to generalize well to large amounts of diverse data and hence be useful in various difficult domain independent tasks such as machine translation and information retrieval.

Several factors make the problem of word sense induction very challenging. Most importantly, it is not clear what should be the 'true' senses of a word. The semantic continuum makes it always possible to break a sense into finer grained subsenses. Thus, the problem is one of finding the optimal granularity for any given task. Even in a semi-supervised setting, it is unknown which sense inventories are most suited as starting points in a sense bootstrapping procedure.

Our unsupervised approach relies heavily on the distributional statistics of words which occur in the proximity of the target words. Hence, we first obtain the distributional statistics from a very large corpus to facilitate generalization and reliable estimation of different possible senses. Then we use these statistics in a novel manner to obtain a representation for the senses of the target word. In this paper, we discus the performance of induced senses on the Semeval 2010 WSD/WSI(Manandhar et al., 2010) task.

2 Related Work

Much of the work on word sense induction has been quite recent following the Semeval tasks on WSI in 2007(Agirre and Soroa, 2007) and 2010, but the task was recognized much earlier and various semi-supervised and unsupervised efforts were directed towards the problem.Yarowsky (1995) proposed a

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1302–1310, Dublin, Ireland, August 23-29 2014.

semi-supervised approach, which required humans to specify seed words for every ambiguous word and assumed one sense per discourse for an ambiguous word. The unsupervised approaches mainly focus on clustering the instances of the target words in a corpus, using first-order vectors, second-order vectors (Purandare and Pedersen, 2004)(Schütze, 1998) etc. Pantel and Lin (2002) used various syntactic and surface features for clustering the various occurences of a target word. Co-occurence graph-based approaches(Véronis, 2004) have also been used, which represent the words co-occuring with the target words as nodes and then identify the highly dense subgraphs or 'hubs' within this co-occurence graph. Brody and Lapata (2009) and Lau et al. (2012) proposed bayesian WSI systems which cluster the instances by applying Latent Dirichlet Allocation (LDA)(Blei et al., 2003), Hierarchical Dirichlet Processes (HDP)(Teh et al., 2006) etc. wherein each occurence of a target word is represented as a 'document' and its surrounding context as the 'observable content'. Choe and Charniak (2013) propose a 'naive bayes' model for WSI which assumes one sense per discourse and uses Expectation Maximization(EM) to estimate model parameters like the probability of generating an instance feature like a word in the context, given the sense of the target word in a particular instance. Reisinger and Mooney (2010) and Huang et al. (2012) have proposed sense dependent multiple prototypes for a word instead of the conventional one vector representation per word and have shown that this sense differentiation improves semantic similarity measurements between words.

3 Basic Motivation: Co-occurence graphs

Conventionally, each word is represented as a co-occurence vector which may contain frequency, point wise mutual information or some lower dimensional representation of context and this representation conflates all the senses of a word. These vectors can be viewed as a graph where words are nodes which have an edge between them if a word occurs in the distributional vector of another. Given a target ambiguous word w, we refer to those words as the 'first order' words(referred to by 'neighbors') which are directly connected to w. The 'second order' words are the words directly connected to the first order words and so on. This graph is cyclic and each node might have multiple senses conflated into it. In this work, we only consider the first and second order words, eg. a target word like 'bank' will have words like 'river' ,'money' etc in it's first order and the second order vectors will be the words from the context of the first order words like 'river':'flood','plains' etc, 'money': 'currency', 'economy' etc. Essentially, these second order words characterize the first order words and hence are very informative for clustering the first order words into different senses. Essentially, we use the second order words as features of the first order words and use them to cluster the first order words into different senses.It must be noted that the first order words themselves might have multiple senses and ideally, those words should also be disambiguated but in the current work we only focus on disambiguating the 'target' words.

4 Methodology

For clustering the neighbors of the target words, we implement and compare two methods which differ significantly in their technical details and employ distribtutional statistics of the neighbors differently, which we describe in the sections below. For obtaining the distributional statistics on a large scale, we used the 5-gram data of Google N-gram corpus(Michel et al., 2011) which effectively lets us use as 10 word window. No lemmatization or case normalization was performed because the large corpus size ameliorated the problem of sparseness. Only nouns, verbs, adjectives and adverbs were employed for the statistical estimation because our pilot studies suggested that these words were most informative.

4.1 Latent Dirichlet Allocation

LDA(Blei et al., 2003) is a well known bayesian generative topic model which models every 'document' as a mixture of latent 'topics' and all its 'words' as multinomial draws from those latent topics. In topic model parlance, a 'corpus' consists of various 'documents'. Each 'document' has a collection of tokens which is treated like a bag of words, where each word is drawn from a latent 'topic'. The topics are shared across documents thus giving each document a topic proportion based upon the topic assignment of the tokens in a document. The priors on topic proportions and the topic multimonial paramers are dirichlet parameters. An important characteristic of LDA is its clustering property which makes the model inclined to enforce sparseness with small dirichlet priors.

It is important to note that we employ LDA in a significantly different manner than the previous approaches which have used LDA or other related topic models for word sense induction. Other topic modelling based approaches for WSI represent each instance of the target word as a 'document' and the immediate context as the 'bag of words' for that 'document'. Unlike these approaches, we represented a target ambiguous word as the 'corpus' in the topic modelling parlance. Then we found out all the 'first order' words co-occuring with the target word within a 10 word window. Each 'first order' word/type is considered a 'document' in our LDA based approach. The latent 'topics' for each 'document' are the latent 'senses' and each first order type comprises of a 'sense distribution' which is indicative of its tendency to induce a particular sense in the target word. The 'second order' types are all the words occuring in a 10 word window of every 'first order' word. These types along with their frequency, form the 'bag of words' for the 'first order' type(LDA document). Hence, in our model, the latent senses are shared across all the first order neighbors of the target word and the second order tokens play the role of 'words' in our LDA based model. After getting the sense distributions for each first order type, we perform k-means over all the sense distribution vectors such that every first order neighbor gets assigned a cluster.

We posit that the distributional statistics of a large corpus helps in improving the coverage of second order words which are essential for reliable clustering of the first order words. However, the large number of occurences and a large vocabulary make it intractable to run LDA using the original frequency of the second order words. To overcome this computational hurdle, we posit that with a diverse representation of the second order words, LDA based parameter estimation relies more upon the relative distribution of the these words across all the first order words rather than their actual distributions. Hence, we decided to scale down the actual counts for each word so that we could run LDA with the finite resources available. An important parameter in this model is the number of latent topics/senses to use, which is specified to be the actual number of senses specified in the Ontonotes sense inventory. This is an idealized case in which the number of senses are known. The α hyperparameter is chosen to be small with respect to the average 'document lengths' we encounter. This has the effect of pushing most of the probabilistic weight to one topics instead of diluting it among many topics. We also decided to analyze the effect of part of speech tags of the second order words in clustering the first order words. The various configurations we experimented with were:

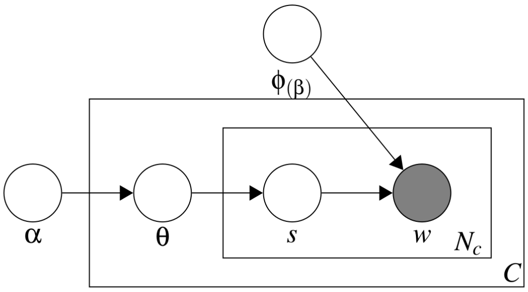

Figure 1: Figure1: s is the latent sense variable. θ is the sense distribution of a first order neighbor. w is a second order neighbor of a first order word. ϕ is the sense multinomial with a dirichlet prior β. α is a dirichlet prior on the sense proportion of a first order type.

- All: Considered nouns,verbs,adjectives and adverbs in second order bag of words.

- Nouns: Only considered second order words which were nouns to study the effect of Nouns on clustering.

- Verbs: Only considered second order words which were verbs.

- Nadj: Considered both nouns and adjectives to study the effect of Noun phrases over clustering.
- Vadv: Considered both verbs and adjectives for second order bag of words.

4.2 Spectral Clustering

Spectral Clustering(Ng et al., 2002) is a clustering technique which uses a pairwise similarity matrix, L, to find out clusters such that the seperation between the entities in two seperate clusters is maximum while implicitly taking into account the distances between groups of points instead of considering them individually. The aim is to find the eigenvectors of $D^{-1}L$ corresponding to smallest eigenvalues to minimize the similarity across two clusters. Here D is a diagonal matrix with degree of node i on entry D_{ii}. For k clusters, k eigenvectors ordered by their eigenvalues are found out. These k eigenvectors are used to form a $n \times k$ matrix where n is the number of datapoints. Each row of this matrix is considered a datapoint with a vector of length k, thus effectively reducing the dimension of the datapoints to k most prominent dimensions according to the similarity matrix decomposition. Finally, k-means is performed on the n vectors to assign a cluster to each datapoint.

We cluster the first order neighbors for each target word using spectral clustering. The crux of this algorithm lies in using appropriate pairwise distance matrices. For constructing the pairwise distance matrices of first order types, we used two vectorial representations of the first order words:

- **Senna embeddings**: The word embeddings trained by a neural network by (Weston et al., 2012)
- **Distributional vectors** comprised of the frequencies of the second order words.

Then we used these vectors to calculate mutual pairwise distance matrices(we experimented with Euclidean and Cosine distances), which were converted into similarity matrices by using Gaussian kernels. These matrices were used as input to the spectral clustering algorithm.

We chose to ignore very low frequency words for making word vectors. This cutoff was decided by analyzing the distributional frequency vs. rank curves of the words, which were heavy tailed. Again, we use the same number of clusters as the number of senses in Ontonotes sense inventory, so that we can study the correspondence between our clusters and the Ontonotes senses.

5 Quantitative Analysis

In this paper, we discus our systems' performances on the Semeval-2010 word sense induction/disambiguation dataset, which contains 100 target words: 50 nouns and 50 verbs. The test data is a part of OntoNotes (Hovy et al., 2006) and contains around 9000 instances of usage of the target words. For annotating a particular test instance, we first filtered the surrounding context to retain only salient Nouns, Verbs, Adverbs, and Adjectives. We report a mixture of senses for each instance, where the weight for each sense was proportional to the number of filtered surrounding words belonging to that sense/cluster. As mentioned earlier, we experimented with a variety of settings for spectral clustering and LDA based methods. The performance with different settings was generally similar and hence, we report our best results here. For a better insight into how our models in different settings performed, we also report the full tables for paired F-score. The performance trend of various systems is similar for other measures. We compare our results to three baselines:

- Most Frequent Sense (MFS) baseline: assigns all the test instances to the most frequent sense of the target word.
- Brown University's system results (Choe and Charniak, 2013).
- Lau (LDA) (Lau et al., 2012), who provide only the results for one of the three measures. In particular, we compare our system to their results obtained by a model that was based on LDA and used the gold standard number of senses as the number of topics to be used.

System	V-measure			Paired F-score			Supervised F-score			#cl
	all	nouns	verbs	all	nouns	verbs	all	nouns	verbs	
LDA	4.4	5.2	3.2	60.7	53.2	71.7	60.9	55.2	69.2	2.45
Spectral	4.5	4.6	4.2	61.5	54.5	71.6	60.7	55.1	68.8	1.87
MFS	0.0	0.0	0.0	63.5	57.0	72.7	58.7	53.2	66.6	1.00
Brown	18.0	23.7	9.9	52.9	52.5	53.5	65.4	62.6	69.5	3.42
Lau	-	-	-	-	-	-	64.0	60.0	69.0	-

Table 1: Performance on Paired F-score and supervised F-score. LDA and Spectral are the two methods proposed in this paper. Lau is the baseline in which LDA system of (Lau et al., 2012) is considered. It should be noted that in their paper, (Lau et al., 2012) did not report their performance on Paired F-score.

The Semeval-2010 task provides us with 3 evaluation metrics: V-measure, Paired F-score and Supervised F-score. It was noticed (Manandhar and Klapaftis, 2009) that V-measure tends to favour systems that produce a higher number of clusters than the gold standard and hence is not a reliable estimate of the performance of WSI systems. But, we report our results on V-measure too as it gives useful insight about the nature of data and the WSI algorithms.

It is important to note that all the measures treat Ontonotes sense annotations as the gold standard, which makes this task unfit for our evaluation purposes. As mentioned earlier, our argument is that several decisions related to the granularity of senses and definition of senses are a topic of dispute, and hence we believe that instead of relying upon a pre-annotated sense inventory, it should be more effective to induce senses automatically in an unsupervised manner using a large and unbiased corpus, and tune the granularity governing parameters for different downstream tasks which require sense disambiguation. But our performance on these annotations still provides us with valuable information about the agreement between Ontonotes senses and our systems' senses. In our experiments, we have not tried to tune the hyperparameters or perform agglomerative clustering to better fit our clusters to the gold standard clusters by using training/development set at all, because we wanted to analyze the performance of our algorithms in the most general setting.

5.1 V-Measure

The V-measure defines the quality of a cluster to be the harmonic mean of homogeneity and coverage. These can be viewed as precision and recall of the element-wise assignment to clusters, where homogeneity measures the 'pureness' of the clusters and coverage measures the 'cohesiveness'. It was noticed (Manandhar and Klapaftis, 2009) that V-measure tends to favour systems producing a higher number of clusters than the gold standard and hence is not a reliable estimate of the performance of WSI systems. In addition, the number of induced clusters in our systems is bounded at the top by the Gold Standard number of senses because of our choice of hyperparameters in both spectral clustering and LDA based approaches.

From the results, we realized that the number of senses induced in the test set by our system is quite low compared to the baselines and other systems that participated in Semeval-2010. This hurts our V-measure. Our systems perform better on nouns than verbs generally according to this measure. Also, LDA-based approaches with the number of topics equal to the number of gold-standard senses perform the best. For spectral clustering, euclidean distances seem to perform better.

5.2 Paired F-score

The paired F-score is the harmonic mean of precision and recall on the task of classifying whether the instances in a pair belong to the same cluster or not. This measure also penalizes the systems if the number of induced senses is not equal to the number of senses in the gold standard. It must be noted that in our approach, the induced number of senses on the test dataset is not equal to the original number of senses although we clustered with the number of clusters specified by Ontonotes, because our clusters are different from Ontonotes senses. MFS has a recall of 100% which makes it a very hard baseline to

P F-score(%)	all	nouns	verbs	#cl
CD20	60.5	53.1	71.3	2.12
CD15	57.9	50.8	68.2	2.26
CD10	58.5	50.7	69.7	2.27
ED20	61.5	54.5	71.6	1.87
ED15	60.6	53.1	71.5	2.12
ED10	60.0	52.3	71.3	2.45
CS15	59.6	52.9	69.4	2.25
CS10	60.1	51.9	72.0	2.07
ES15	59.8	52.9	71.3	2.15
ES10	60.8	53.5	71.4	2.21
MFS	63.5	57.0	72.7	1.00
Brown	52.9	52.5	53.5	3.42

Table 2: General trend for the various settings: Paired F-Score Evaluation: Spectral Clustering: ‘C’:cosine distance, ‘E’: Euclidean Distance, ‘D’: Second order Distributinal counts, ‘S’:Senna embeddings and the adjacent numbers are the number of nearest neighbors(in 1000s) considered for the distance matrix.

P F-score(%)	all	nouns	verbs	#cl
all	60.7	53.2	71.7	2.47
noun	59.6	52.1	70.7	2.32
verb	60.0	52.4	71.0	2.25
nadj	59.7	52.6	70.1	2.3
vadv	59.3	52.27	69.6	2.25
MFS	63.5	57.0	72.7	1.00
Brown	52.9	52.5	53.5	3.42

Table 3: General trend for the various settings: Paired F-Score Evaluation: LDA: ‘all’: All POS tags considered in the first order neighborhood, ‘noun’: Only nouns considere, ‘verbs’: Only verbs considered, ‘nadj’: nouns and adjectives considered, ‘vadv’:verbs and adverbs considered

beat. Semeval-2010 results show that none of the systems outperform the MFS baseline. Both of our systems perform better than other systems on this measure and are comparable to the performance of the MFS baseline.

5.3 Supervised F-score

For the supervised task, the test data is split into two parts: one for mapping the system senses to the gold standard senses, and the other for evaluation based upon the mapped senses. We report our performance on the 80% mapping and 20% evaluation split. The mapping is done automatically by the program provided by the organizers which is based upon representing the gold standard clusters as a mixture of the system senses.

Our different systems perform similarly on the supervised evaluation. We outperform the tough MFS baseline and perform competitively against other systems. We observe that other systems outperform us on the target nouns whereas our performance on verbs is similar to that of other systems. This can be attributed to the fact that our methods induce a small number of senses in general over the test set but according to the test data based upon Ontonotes, the senses of nouns have a much higher resolution than verbs.

5.4 Discussion on Quantitative Results

In general, we found our performance to be competetive with the other systems. Also, we perform significantly better than other Semeval-2010 systems on the paired F-score metric. In our experiments,

Sense	Cluster Words
1	Engineers,Presbyterian,Service,Jewish,Police,Ethnicity,Independent,Movements
2	membrane,complicated,surgical,hypothalamic,potassium,lymphatic,electron,tumor
3	Cynthia,Armstrong,Tracy,Marilyn,Stella,Abbot,Gustavus,Clark,Stewart,Monica
4	heels,noses,haze,hand,drooping,galloped,nakedness,pallid,anguish,palms
5	night,burdens,gut,assassins,witness,results,celestial,visual,deep,Hell
6	lifted,hastily,hovering,guiding,sinner,tendency,developing,sacrificed,condemned

Table 4: Example words in the clusters of ‘body.n’

we found that for spectral clustering, Euclidean distances tend to perform better than Cosine distances. Also, the distributional counts of the second order words tend to perform better than Senna vectors which is not surprising because the Senna vectors are trained with the philosophy of a language model, which results in words often being clustered according to their POS tags rather than their semantic closeness. Spectral methods, yield slightly better results on two metrics than LDA based clustering which suggests that similarity matrices give us a better idea about interactions between groups of words than simple occurence frequencies of the words. But a bigger advantage of spectral clustering techniques is the speed of computing SVD which is much better than that of slow inference algorithms of LDA based models.

For LDA based models, we also note that different settings focusing on different POS tags, performed very similarly and did not indicate any strong preference for any POS tag for the task of WSI using LDA. Finally, both our methods tend to induce a small number of senses in the test data, which suggests that the induced senses are relatively coarse-grained. Further splitting of coarse clusters using hierarchical clustering methods might be helpful if a task requires finer-grained senses.

6 Qualitative Analysis

In this section, we present some deductions drawn from the qualitative analysis of clusters generated by our methods which support our hypothesis. In particular, we discus the nature of clusters generated by the spectral clustering algorithm using the second order distributional vectors for obtaining the similarity matrix based on Euclidean distance.

A preliminary analysis of cluster sizes revealed that in almost all the cases, one of the clusters was very large(about 3 times larger than the second largest cluster) and this largest cluster seemed to conflate a lot of senses. Other clusters were generally similar sized and most of them represented a sense of the target word on their own. The results in general look very promising and many clusters can be easily interpreted as different senses of the target word.

In Table 4, we show the top few words for the word ‘body.n’. Some senses very clearly represent themselves : 1. Body as in organization, 2. Biological terms related to body, 4. Body in a more informal sense. Sense 5 seems like a mixture of two senses of body, one related to celestial bodies and other related to dead bodies/murder. Interestingly, sense 3 comprises proper nouns i.e. people whose bodies have been mentioned in the corpus. This is not a conventional sense listed in any of the sense inventories but based upon the requirements of a task, one might be interested in differentiating between general mentions of ‘bodies’ and mentions of ‘bodies’ which appear when mentioning famous people or celebrities. This sort of clustering can be incredibly useful in tasks like Machine Translation and Information Retreival which require us to model semantics of rare words such as important proper nouns.

7 Discussion and Future Work

We used a large corpus and its distributional statistics to perform word sense induction for a set of 100 target words. We proposed two algorithms which cluster the salient words surrounding the target word by using the distributions of surrounding words. Both LDA based algorithm and the spectral clustering algorithm yielded similar clusters. We believe that these clusters can be employed in downstream tasks and can be further broken into smaller fine grained clusters automatically if needed by the application.

We also evaluated our clusters arising from the distributional statistics, in the Semeval-2010 tasks without any tuning and showed that they perform competetively with other approaches.

We argue that treating existing sense inventories as gold standards for WSI tasks is not an appropriate measure for WSI systems because these inventories would not be able to measure two very important characteristics of WSI systems which make them more advantageous than supervised WSD systems:a) coverage and b) discovery of new senses.

Hence, the Semeval-2010 experiments are not an accurate reflection of the capabilities of WSI systems because they rely on the Ontonotes sense inventory for the Gold Standard judgements, which are admitted even by the OntoNotes builders to be only 85% reliable on average (Hovy et al., 2006). Our competetive performance on these tasks show that our methods can be compliant with standard word sense disambiguation tasks but more importantly, our qualitative analysis showed that our techniques can discover new unconventional senses too, which might not be present in the sense inventories but could be very useful in tasks requiring differentiations. Unfortunately, no metrics exist that can help us quantify the coverage of senses and their novelty. An ideal metric to evaluate the WSI systems in a better manner, would be their performance on extrinsic tasks like Machine Translation, Information Retreival, Machine Reading etc., which require differentiation of senses at different granular levels. WSI techniques have a potential of eliminating sense annotation costs hence enabling wider use of sense differentiation in a more generalized setting.

Our techniques resulted in coarse-grained senses. A major challenge in this task is to determine the appropriate number of senses to induce. To overcome this problem, non-parametric methods could be conceived to identify the ideal number of clusters automatically. In future, the WSI systems like ours can also be used to analyze the evolution of senses over a period of time or geographical variation of senses. As mentioned earlier, the co-occurence graph consists of many canonical representation of words which must be split according to their different senses. In our experiments, we considered a small number of target words and did not take into account the multiplicity of senses in the representation of 'first' and 'second' order neighbors. A more sophisticated iterative approach involving making several passes over a co-occurence graph and refining senses of different words in each pass can ameliorate the problem associated with a single canonical representation of neighboring words. Finally, designing extrinsic tasks to measure the efficacy of WSI systems will be extremely helpful in development of more robust and useful WSI systems.

Acknowledgments

This research was supported in part by DARPA grant FA8750-12-2-0342 funded under the DEFT program.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics.

Do Kook Choe and Eugene Charniak. 2013. Naive bayes word sense induction. In *EMNLP*, pages 1433–1437.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. 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.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Suresh Manandhar and Ioannis P Klapaftis. 2009. Semeval-2010 task 14: evaluation setting for word sense induction & disambiguation systems. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 117–122. Association for Computational Linguistics.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan. 2010. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

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

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.

Amruta Purandare and Ted Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48. Boston.

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. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Jean Véronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. 2012. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.