Distributed Representation-based Spoken Word Sense Induction

Justin Chiu, Yajie Miao, Alan W Black, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Jchiu1@andrew.cmu.edu, Yajiemiao@gmail.com, Awb@cs.cmu.edu, Alex.Rudnikcy@cs.cmu.edu

Abstract

Spoken Term Detection (STD) or Keyword Search (KWS) techniques can locate keyword instances but do not differentiate between meanings. Spoken Word Sense Induction (SWSI) differentiates target instances by clustering according to context, providing a more useful result. In this paper we present a fully unsupervised SWSI approach based on distributed representations of spoken utterances. We compare this approach to several others, including the state-of-the-art Hierarchical Dirichlet Process (HDP). To determine how ASR performance affects SWSI, we used three different levels of Word Error Rate (WER), 40%, 20% and 0%; 40% WER is representative of online video, 0% of text. We show that the distributed representation approach outperforms all other approaches, regardless of the WER. Although LDA-based approaches do well on clean data, they degrade significantly with WER. Paradoxically, lower WER does not guarantee better SWSI performance, due to the influence of common locutions.

Index Terms: Spoken Word Sense Induction, Spoken Language Understanding, Distributed Representations

1. Introduction

STD [1] focuses on finding instances of a text query in an audio corpus, and provides access to useful portions of the speech data. However, detecting the presence of a query may be insufficient if the query word happens to have multiple meanings. Presenting every instance of the query with different meaning is not efficient. Presenting the search result clustered by meaning could significantly increase the interpretability of the detected term.

Clustering target keyword according to the meaning requires Word Sense Induction (WSI) [2]. We explore Spoken Word Sense Induction (SWSI), which enables WSI on human speech instead of natural language text. Since speech data is noisier and (spontaneous) spoken language is less structured, we anticipate a greater challenge in SWSI, compared to a text-based WSI task.

In this paper, we describe a fully unsupervised SWSI approach that utilizes distributed representation [3] of spoken utterances. We compare our approach with several other approaches, including the state-of-the-art Hierarchical Dirichlet Process (HDP) which achieved the best result in SemEval-2013 WSI task [4]. We also test on three different levels of Word Error Rate (WER), as WER constitutes one of the major differences between SWSI and WSI. Related work is presented after our results and analysis section to provide boarder insight on the problem.

This paper makes three contributions:

- We present the Spoken Word Sense Induction (SWSI) task, together with a procedure that does not require human labeling for evaluation.
- We demonstrate that distributed representation-based approaches outperform other approaches regardless of the level of WER. LDA-based approaches do well on clean data. However, they significantly degrade as WER increases.
- We also show that the lower WER does not guarantee better performance on SWSI, possibly due to the reduced errors are mostly common locutions (phrases commonly used in spoken language), which does not contribute to the understanding of the content.

2. Approach

In this section, we will introduce our motivations and describe our techniques for constructing a distributed representation for spoken utterances.

2.1. The Skip-gram Model

Mikolov et al. [3] recently introduced the Skip-gram model. Skip-gram models and other Neural Network Language Models (NNLM) produce word representations for each word in the training data according to its surrounding words. Each word can be viewed as a point in a "Word Embedding" space, and if there are two words that are located closely in this space, it means those two words tend to show up in similar surrounding word contexts in training data. The advantage of using Skip-gram model instead of other NNLM is that the Skip-gram model requires much less computing resource yet it can still achieve good performance. (The comparisons between Skip-gram model and other NNLM are presented in the Related Works). We followed the standard training procedure of Skip-gram model in addition with Negative Sampling and Subsampling of Frequent Words. The parameter k for Negative Sampling is set to 5, and the parameter t for Subsampling of frequent word is set to 10^{-4}. For more detail of Skip-gram model training, please see [3].

The Skip-gram model will produce a single point in the "Word Embedding" space for each word in the training data. However, this is actually a limitation of the model, as each word is forced to be represented as a single point in the "Word Embedding" space. This is not an ideal situation, because if the *w* has different meanings, it is likely to occur with very different surrounding words. The computed single point for *w* is the average of all instances of *w*, which conflates the different meanings. If sense-labeled training data is available,

then it would be possible to train multiple distributed representations that differentiate the different meaning of the same word, yet such data would not be available in a typical SWSI situation.

2.2. Distributed Representation of Utterance

In order to overcome the limitation of existing Skip-gram models, we use a distributed representation for utterances to differentiate the meaning of multiple instances of the same word. Our intuition is that, if we can obtain the distributed representation for the entire utterance, which contains our target word and the surrounding words, we can then use that representation to differentiate the meanings of a specific word. Thus if the meaning of the utterance is different, we can expect that even the same word in an utterance is likely to have different senses. The SWSI task is usually considered to be a clustering task; clustering the utterance instances can be a good approximation of clustering the words by sense.

We obtain the distributed representation for an utterance as follows: We assume there is an extra "utterance token" associated with each utterance. This token will be trained with every other word in the sentence. So given a sequence of training word w_1, w_2, …, w_T in a specific utterance, the objective of the distributed representation of the utterance is to maximize the average log probability

$$\sum_{t=1}^{N} \log p(w_t \mid u) \tag{1}$$

where N is the size of the entire utterance and u is the "utterance token". This will map the utterance into the same space with other words in the training data, so the utterance can also be represented by the distributed representations used for the other words.

3. Experiments

3.1. Dataset

We use 60 hours of YouTube "How To" video for our experiments. The YouTube video corpus [5] we used includes human transcription, allowing us to compute the WER for ASR.

The ASR system we use to decode the speech is based on the Kaldi [6] toolkit. We have two different setups of acoustic model training to simulate different WER, which were 39.13% and 19.95% (nominally, 40% and 20%). The acoustic model of the 40% WER system is trained on the Wall Street Journal corpus consisting of approximately 80 hours of broadcast news speech. The 20% WER system's acoustic model is trained on 360 hours of video data that are in the same domain as the testing data. Speaker adaptive training (SAT) is conducted via feature-space MLLR (fMLLR) on LDA+MLLT features. DNN [7, 8, 9, 10] inputs include spliced fMLLR features. All decoding runs use a trigram language model that is trained from 480 hours of YouTube transcripts. The 40% WER system is meant to simulate a mismatch between training and testing data, common in real world use cases; it is about the same level as reported in [11]. The 20% WER system represent a more controlled environment (or more accurate ASR), as the mismatch between training data and testing data is much smaller. Together with the human transcription which is nominally 0% WER, we expect this can provide insight on how ASR performance affects SWSI performance. The number of word token and vocabulary size is reported in the following table:

Table 1. *Vocabulary size and number of tokens.*

WER (%)	40	20	0
Vocabulary Size	55266	52377	55162
Number of token	715849	745402	742260

In order to select the target queries for our SWSI task, we adopt the query selection process used in the SemEval-2013 WSI task. We selected those queries for which a sense inventory exists as a disambiguation page in the English Wikipedia[1]. As well, the queries we selected each have 3 senses among the WordNet 5000 most common senses [12] to ensure that the difficulties are comparable. Every query appears at least once in our 60 hours YouTube data.

3.2. Evaluation Metrics

A variety of evaluation metrics [13, 14, 15, 16] can be used for evaluating SWSI cluster quality. However, most of these will be affected by chance agreement caused by the number of clusters used. We therefore use the Adjusted Rand Index (ARI) [14] as our evaluation metric, as it removes the effect of the chance agreement; ARI was used in the SemEval-2013 WSI task. The standard ARI ranges from -1 to 1, however we follow the presentation format used in the SemEval-2013 WSI task and multiply the value by 100, to make it range from –100 to +100.

Defining the reference cluster for our queries is also a challenge, as asking human to label the actual word sense would require significant resources. Instead, we use a WordNet-based Word Sense Disambiguation (WSD) approach [17] to label the sense with the human transcript (0% WER) as our reference sense. If our query word is actually a recognition error (which means it does not occur in the human transcription), the reference sense for that instance is a specific sense of "Wrong Word" which only applies to recognition errors.

3.3. Experimental Setup

Our approach for using distributed representation of utterance for SWSI is straightforward. First, we train the distributed representation using the entire 60 hours of ASR transcription. For each of the utterance that contains the query word, we create a 100-dimension utterance vector. The utterance vector is trained using a standard toolkit[2]. We then perform repeated bisections clustering [18] on the utterance vector according to a pre-defined number of desired clusters using the CLUTO toolkit [19], and the MALLET toolkit [20] for the subsequent LDA-related processing. All the parameters are default values unless specified.

In order to estimate how our SWSI approach compares to the other existing approaches, we also conducted the same experiments using four different baseline systems:

[1]http://en.wikipedia.org/wiki/Category:Disambiguation_pages
[2]https://code.google.com/p/word2vec/.

Bag-of-Word (BOW) system: In BOW system, each utterance is represented by its BOW feature. We then perform repeated bisections clustering on the BOW feature. [21]

Latent Dirichlet Allocation feature (LDA-feature) system: Instead of using BOW as the feature for each utterance, it first builds a LDA model with 100 topics on the entire 60 hours of testing data. The repeated bisections clustering use the topic distribution of utterance as feature.

Latent Dirichlet Allocation (LDA) system: Described in [22], the LDA system trained the topic model only on the utterance that the query occurs. The number of topics is the desired cluster numbers, and each utterance is assigned to the topic that has the highest topical probability.

Hierarchical Dirichlet Processes (HDP) system: Also described in [22], the HDP system is trained and clustered in the similar way to the LDA system. However, it does not require any assignment for the topic (cluster) numbers, as the algorithm determines the number of topics automatically. HDP achieved the best performance in the SemEval-2013 WSI task.

We also evaluated our WordNet-based **WSD** system on the ASR transcription. This indicates how WSD system can perform given a widely-available knowledge source such as WordNet.

We conducted two different set of experiments. The first set of experiments show how different approaches perform with different assignment of senses (clusters) on 40% WER data, our expect real-world scenario. The second set of experiments compares how different approaches perform under different WER conditions. This shows how noise introduced by an ASR system affects the SWSI performance for each approach.

4. Results

4.1. Comparison between WSI approaches

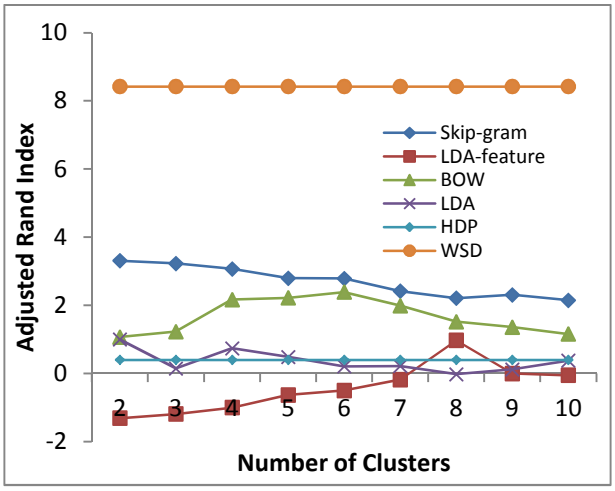

Figure 1: *ARI Comparison from different approaches with different numbers of clusters on 40% WER data.*

Figure 1 shows the ARI performance for our skip-gram based SWSI system as compared with the four baseline systems on 40% WER data. The WSD system is knowledge-based and indicates the performance achievable with a human-produced knowledge source such as WordNet. None of the other approaches rely on external knowledge. We vary the number of clusters to see how different approach interacts with the number of clusters. The only exception is the HDP system, as its algorithm will decide the most appropriate number of clusters using a data-driven method.

4.2. Comparison between WER

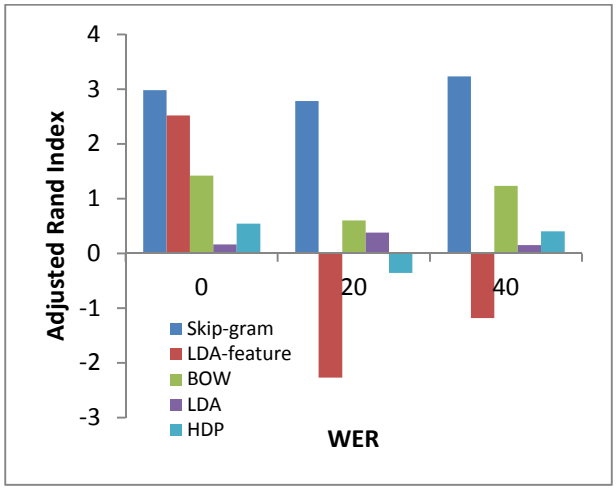

Figure 2: *ARI Comparison with number of cluster = 3 on different Word Error Rate.*

Figure 2 shows the comparison between the SWSI systems at different WERs. This result leads us to three conclusions. First, regardless of the varying WER, the Skip-gram based SWSI always achieves the best performance. Second, the LDA-feature system achieves decent performance in the 0% WER condition, but its performance is degraded significantly when noise (i.e. misrecognitions) is present. The noise due to ASR error disrupts the topical distribution, and hence degrades the quality of the LDA topical distribution feature. Third, in contrast to general expectation, reducing the WER does not directly transfer into a significantly better SWSI performance. We believe this is due to the presence of common locutions. Table 2 shows the percentage of the context words around the query that are high frequency (top 1%). Despite the significant difference on WER, the percentage of context consisting of frequently occurring words is similar. This implies that words benefiting from the lower WER may not be the ones that impact the meaning of the content. This also reflects human's conversational behavior, which is weighted towards high-frequency locutions.

Table 2. *Percentage of the context which is frequently occurring words.*

WER (%)	40	20	0
% of context is frequent word	76.9	78.8	78.1

5. Analysis

5.1. Exploring the Ideal Number of Senses

Deciding the correct number of senses/clusters is a perennial challenge in research. In this section, we provide our observations on how the number of reference senses interacts with the cluster numbers in the Skip-gram SWSI system.

Figure 3 shows the interaction between the number of assigned clusters and the number of reference senses for three different

levels of WER. The x axis shows the number of assigned clusters minus the number of reference clusters. The large decrease on the X = –1 is due to multiple instances of queries that have 2 meanings; assigning 1 sense to every word leads to an ARI of 0. According to the result, we observe that assigning 1 or 2 extra cluster compared to the reference sense inventory achieves the best performance. We conjecture that this is caused by the clustering algorithm benefitting by having an extra cluster to hold the "noisy" data. Without this extra cluster, the quality of the other clusters is reduced.

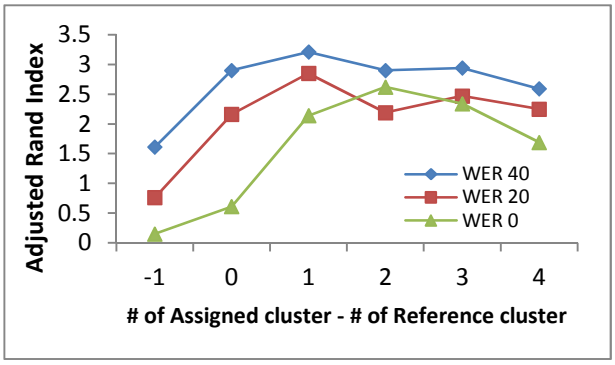

Figure 3: *ARI Comparison for interaction between the number of assigned and reference clusters.*

5.2. Related Experiments

Our Skip-gram based SWSI system achieves good performance on the described task, yet it still has limitations. The distributed representation requires a sufficient amount of training data to produce a stable vector space. We investigated reducing the amount of data used to train the distributed representation. When the video dataset is reduced to about 30 hours (which contains around 300,000 tokens) the SWSI performance is reduced to about the level of the BOW system. The performance continues to degrade with even less data are included. The BOW system, on the other hand, maintains roughly the same performance level despite reduction in the amount of data.

Distributed representation could be considered as a way to capture semantic information in the data. We also investigated its use as a way to identify possible recognition errors (that is, a given misrecognition may be occurring in an unexpected context). Accordingly, we conducted a preliminary experiment to test this possibility. We assume the cluster that has the highest variance would be the cluster that most likely pools recognition errors, as the source contexts would be very different. The experiment was inconclusive: high variance did not correlate with recognition error. We suspect that this was due to the fact that we trained the distributed representation using noisy data and that its variance is inherently high. We suspect using distributed representation based on a cleaner corpus (such as Wikipedia) might achieve better performance as the space would model the relationships in clean text.

We also investigated recognition error detection using the Word Burst phenomenon [23], a content word that occurs in isolation tends to be an instance of recognition error. We find that 85% of the recognition errors on query words in the 40% WER data match this assumption. We changed the cluster assignment for every instance of query word that matched the Word Burst assumption to a separate cluster that represents the "Wrong Word" sense. Performance does not improve, as for these data there are many correct instances that are singletons as well. Nevertheless we believe this can be a useful feature as it shows a very high recall rate (85%) for identifying possible recognition errors.

6. Related Work

Multiple authors address the WSI problem, from different perspectives. [22] investigates graphical model oriented approaches, including LDA and HDP which we use as baseline systems in this paper. [24] uses the concept of submodularity. The WSI task is treated as a submodular function maximization problem. [25] reported their WSI systems based on second order co-occurrence features which atttempts to capture the connection between words that are likely to co-occur with the same word. These investigations are reported on nature language text, and do not address the possible effect of noise (recognition errors or locutions) found in spoken data.

Other research [26, 27, 28] has investigated different neural network based distributed representations of words. [29] evaluated distributed representations on the word analogy task, and found that the Skip-gram models achieved the best performance by a significant margin. Regarding creating a distributed representation for multi-word instance [30], [31] reported a more sophisticated approach that combines the word vector in an order specified by a parse tree. However, due to its reliance on parsing, this approach only works on well-structured natural language sentences. Spoken utterances are harder to parse due to the presence of recognition errors and common locutions.

7. Conclusion

Our work makes several key contributions. We present the Spoken Word Sense Induction (SWSI) task, and describe an approach that does not require human labeling for evaluation. We also present a fully unsupervised SWSI approach based on the distributed representations for spoken utterances, which outperforms several existing approaches on different accuracies of ASR transcript. An interesting result is that, in contrast to expectation, improving WER does not guarantee an improvement in SWSI performance. We believe this is the main difference between SWSI and standard text-based WSI, as the words that benefit from the lower WER may not be the ones that impact the meaning of the content.

8. Acknowledgement

This work was funded in part by the Yahoo InMind project at Carnegie Mellon. We would like to thank Robert Frederking for his contributions.

9. References

[1] J.G. Fiscus, J. Ajot, J.S. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," *Proc. SIGIR*, Vol 7, pp. 51-57, 2007.

[2] R. Navigli, "Word Sense Disambiguation: a survey," *ACM Comupting Surveys*, 42(2):1-69, 2009.

[3] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.

[4] R. Navigli, and D. Vannella, "SemEval-2013 Task 11: Word Sense Induction & Disambiguation within an End-User Application," Proc. *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Vol 2, pp. 193-201, 2013.

[5] S.I. Yu, L. Jiang, A. Hauptmann, "Instructional Videos for Unsupervised Harvesting and Learning of Action Examples," Proc. *ACM International Conference on Multimedia*, pp. 825-828, 2014.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, "The Kaldi Speech Recognition Toolkit," *Proc. ASRU*, pp. 1-4, 2011.

[7] Miao, Y. and Metze, F., "Improving Low-Resource CD-DNN-HMM using Dropout and Multilingual DNN Training", *Proc. Interspeech*, pp. 2237-2241, 2013

[8] Miao, Y., Metze, F. and Rawat, S., "Deep Maxout Networks for Low-Resource Speech Recognition", *Proc. Automatic Speech Recognition and Understanding (ASRU)*,pp. 398-403, 2013

[9] Miao, Y., and Metze, F., "Distributed Learning of Multilingual DNN Feature Extractors using GPUs", *in Proc. Interspeech, 2015*. To appear

[10] Miao, Y., and Metze, F., "Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models", *in Proc. Interspeech, 2015*. To appear

[11] H. Liao, and E. McDermott, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," *Proc. ASRU*, pp. 368-373, 2013.

[12] P. Clark, C. Fellbaum, J.R. Hobbs, P. Harrison, W.R. Murray, and J. Thompson, "Augmenting WordNet for deep understanding of text," *Proc. Conference on Semantics in Text Processing*, pp. 45-57, 2010.

[13] W.M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, 66(336), pp. 846-850, 1971.

[14] L. Hubert, and P. Arabie, "Comparing Partitions," *Journal of Classification* 2(1), pp. 193-219, 1985.

[15] P. Jaccard, "Etude comparative de la distribution florale dans une portion des alpes et des jura," In *Bulletin de la Societ´ e Vaudoise des Sciences Naturelles*, Vol. 37, pp. 547-579, 1901.

[16] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, second edition, 1979.

[17] L. Tan, "Pywsd: Python Implementations of Word Sense Disambiguation (WSD) Technologies", Retrieved from https://github.com/alvations/pywsd

[18] Y. Zhao, and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 515-524, ACM, 2002

[19] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD workshop on text mining*, Vol. 400, No.1, pp. 525-526, 2000

[20] A.K. McCallum, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu, 2002.

[21] P. Pantel, and D. Lin, "Discovering Word Senses from Text", Proc. *8th International Conference on Knowledge Discovery and Data Mining*, pp. 613-619, Canada, 2002

[22] J. H. Lau, P. Cook, and T. Baldwin, "unimelb: Topic modelling-based word sense induction," Proc. *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, Vol 2, pp. 307-311, 2013.

[23] J. Chiu, A. Rudnicky, "Using Conversational Word Burst in Spoken Term Detection," Proc. *Interspeech*, pp 2247-2251, 2013

[24] S. Behera, R. Bairi, U. Gaikwad, and G. Ramakrishnan, "SATTY: Word Sense Induction Application in Web Search Clustering," Atlanta, Georgia, USA, 2013.

[25] T. Pedersen, "Duluth: Word Sense Induction Applied to Web Page Clustering," Atlanta, Georgia, USA, 2013.

[26] R. Collobert, and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," *Proceedings of the 25th international conference on Machine learning,* pp. 160-167, 2008.

[27] A. Mnih, and G.E. Hinton, "A scalable hierarchical distributed language model," *Advances in neural information processing systems*, pp. 1081-1088, 2009.

[28] J. Turian, L. Ratinov, Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384-394, ACL, 2010.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ICLR Workshop*, 2013

[30] Q.V. Le, and T. Mikolov, "Distributed representations of sentences and documents," *arXiv preprint arXiv:1405.4053*, 2014

[31] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," *Advances in Neural Information Processing Systems*, pp.926-934, 2013.