Interspeech 2018
2-6 September 2018, Hyderabad

INTENT DISCOVERY THROUGH UNSUPERVISED SEMANTIC TEXT CLUSTERING

Padmasundari[1], *Srinivas Bangalore*[2]

Interactions LLC

[1]padma@interactions.com, [2]sbangalore@interactions.com

Abstract

Conversational systems need to understand spoken language to be able to converse with a human in a meaningful coherent manner. This understanding (Spoken Language understanding - SLU) of the human language is operationalized through identifying intents and entities. While classification methods that rely on labeled data are often used for SLU, creating large supervised data sets is extremely tedious and time consuming. This paper presents a practical approach to automate the process of intent discovery on unlabeled data sets of human language text through clustering techniques. We explore a range of representations for the texts and various clustering methods to validate the clustering stability through quantitative metrics like Adjusted Random Index (ARI). A final alignment of the clusters to the semantic intent is determined through consensus labelling. Our experiments on public datasets demonstrate the effectiveness of our approach generating homogeneous clusters with 89% cluster accuracy, leading to better semantic intent alignments. Furthermore, we illustrate that the clustering offer an alternate and effective way to mine sentence variants that can aid the bootstrapping of SLU models.

Index Terms: spoken language recognition, conversational systems, intent discovery, semantic analysis and classification

1. Introduction

The task of understanding spoken language is central to the present day sophisticated human-computer conversational dialog systems. In order to achieve high accuracy and precision of understanding, the task of spoken language understanding (SLU) is typically reduced to identification of intent and extraction of named entity. We focus on the problem of identifying intents here and present our experimental study on the same.

Given a large collection of unlabeled documents, it is often a useful exercise to get the documents organized into smaller subsets of similar or related documents. Clustering is an important tool here. The vector space modelling of the data and the similarity metric determine whether the data in hand has a natural tendency to cluster or not. Using semantic representations in appropriate higher dimensional vector space and clustering the speech language text data, we attempt to relate the text intents to the characteristics of the resulting semantic clusters.

The process of clustering involves multiple steps, ideally starting with the screening and pre-processing of the data. Choosing the data representation and a suitable similarity measure is vital to determining the quality and usability of the clustering. The choice of representation is specific to the application problem domain but independent of the clustering algorithm being deployed. A "good" clustering of data requires effective feature selection, a proper choice of the algorithm and clustering strategy for the task at hand. Finally, cluster validation may be performed and the clustering results analyzed to draw meaningful conclusions regarding the subsets or the subgroups induced by the clustering.

The clustering algorithms have been studied extensively across a wide range of specific application domains and many algorithms have been proposed in the literature [1, 2]. The parametric clustering algorithms have been very popular due to their simplicity and scalability. However, the challenges of using clustering methods with unsupervised data are many. Clustering very large document collections by itself is a challenge. K-Means and most other parametric methods for clustering require prior knowledge of the "true" number of cluster partitions, which often comes from supervised data with assigned data labels. A major challenge is also due to the substantial variations in the results of the clustering exercise. In the absence of any real data with the correct ground truth classifications, further challenge is in evaluating the quality of clustering as well as in understanding and interpreting the data clusters.

The rest of the paper is organized as follows. Section 2 includes a brief review of related works. Section 3 introduces the required definitions and describes our practical approach to intent discovery based on a "stable" clustering solution. Section 4 details the datasets used in this work, the experiments performed to validate the approach and the analysis of results. Section 5 summarizes the findings with concluding remarks.

2. Related Study

A dialog system requires to understand the intentions of humans and extract the relevant information and actions from the intentions, to be able to converse with a human in a coherent manner. Various modes for communication like text, speech, graphics and other modes may be used on the input and output channels. Natural language processing aims at extracting the intention and relevant information from text. Document or text clustering and its applications in topic extraction, intent discovery and information retrieval have been quite frequently studied in the literature. This section includes a brief review of only the most relevant of them.

With respect to text clustering, the traditional expectation is that the hierarchical methods would perform much better than the rest. In contrast, the experimental studies in [3] find that the K-means or its variants actually do better than the other methods. [4, 5] explore and compare various text clustering algorithms. [4] evaluates that the partitional algorithms always lead to better solutions than the agglomerative algorithms. It proposes to combine features from partitional methods to reduce the initial errors in agglomerative methods and improve the quality of the clustering solution. Though this gave better solutions compared to the agglomerative methods, however, did not always outperform the partitional methods. Also the work finds that the partitional algorithms have relatively low computational requirements with higher clustering quality, thus

10.21437/Interspeech.2018-2436

making them ideal for use with large document collections.

Various measures to ascertain the validity and quality of a clustering solution have been developed and used in the literature. These may all be derived from the confusion matrix but are based on different ideas like counting of pairs, set overlaps summation, and mutual information. [6, 7] introduce re-sampling based approach to estimate the number of clusters and improve the accuracy of the clustering procedure. The *silhouette index* is used to determine the optimal number of clusters. [8] provides an overview and analysis of validity measures and further defines a set of properties as a guide to defining and detecting "good" measures. However, with regard to the problem of determining the appropriate number of clusters to be used with a clustering method on a given unsupervised dataset, cluster validation based on stability properties may come as an immediate choice; that is, one chooses the number of clusters which produces the most stable results.

[9] performs a stability study involving the *adjusted Rand index* and *adjusted mutual information index*. [10] provides a concise overview on clustering stability. [11] illustrates on supervised data that there exists a positive correlation between *impurity* and *sensitivity to random perturbations* of the input data set (supervised data) and then discusses an approach to evaluate clustering of unsupervised data.

Thus, researchers, in the past, have used measures such as F1-scores and purity to evaluate cluster validity. However, [12] argues that the measures based on Entropy, Purity and Mutual Information are not suitable for evaluating the K-means clustering and are "defective" validation measures for K-means clustering. [12] does an organized study of 16 external validation measures that are in use in such fields as data mining, information retrieval, machine learning, and statistics and provides a guide line for selecting suitable validation measures for K-means clustering. The work concludes that for evaluating quality of K-means clustering, one may only need to consider the *Rand statistic* based on the counts of agreements and disagreements of data object pairs in dierent partitions. External validations, measuring the similarity of two clustering solutions, take care of the point-level differences required to evaluate stability and performance of a clustering algorithm.

Document clustering and modeling for extracting information as topic, intent, etc.. often go well together with mutual benefits. Machine learning algorithms like clustering require the document or natural language text data input to be represented as a fixed-length feature vector. Our interests rest in the problem of clustering spoken language texts into groups by their meanings. The Bag-Of-Words (BOW) is a classical model used commonly with document clustering, however, generally insufficient to capture all the semantics. Vector space models (VSMs) represent the meanings of lexical items as vectors in a *semantic space*. The idea of semantic spaces has been researched for more than two decades. However, recent advances in neural networking techniques and computational approaches like tensors have led to creating better representations, capturing the meaning of a spoken language text. We highlight a few of the very recent attempts on the use of clustering on the semantic space for intent or topic modelling.

[13] proposes a *Multi-Grain Clustering Topic Model* (MGCTM) using a mixture component to discover groups in the dataset and a topic model component to fine-grain to further topics local to the clusters and globally shared across clusters. [14] starts with a semantic vector space model of words using Gaussian Mixture Models (GMMs) and does not use the notion of a document. The mixture components of the words capture the notion of latent topics here. [15] uses K-means to cluster the sentences based on vector representations derived as a weighted sum of the word embeddings. It uses intra-cluster perplexity and F1 scores to measure the effectiveness of the clustering. However the training corpus used is artificially composed for the purpose of experiments. Given the training labels of some known classes, being away from the traditional semi-supervised techniques, [16] studies the problem in an interactive context, incrementally proposing complementary clusters.

In the current work, we explore a range of semantic representations on spoken language text to best capture the meaning and intent through word semantics. While adopting a clustering approach on the unsupervised text data for intent discovery, the clustering solution is chosen based on the notion of stability with respect to the process of clustering itself (and not with respect to the data). We combine a few methods for generating multiple clustering and the final cluster membership of the documents is determined by consensus.

3. Approach

3.1. Semantic representations on spoken language text

Word embedding is, a numeric feature representation for words, learned from local co-occurrences of words in sentences in a language, generated by the neural networks. The word vectors have demonstrated nice semantic and syntactic behaviours. Semantically close words are also close in their vector representations. We have considered word2vec[17], gloVe[18] and ConceptNet-numberbatch[19] as the choices for semantic word representations.

The meanings and intents of sentences or human language text may be best captured through word semantics. doc2Vec is an extension of word embeddings. It is continuous distributed vector representation learned, from word embeddings, for pieces of texts (of variable length) [17] that can capture semantic similarity or relatedness between documents. Summing or averaging of the word vectors may also be used to represent sentences or documents, but the method is known to neglect a lot of information like the sequence of words. However, our initial experiments found that summing or averaging of word vectors often performed better than the doc2Vec. Hence based on the word embeddings, we also propose and derive the following vector representations for the spoken language text, to use in our experiments.

- w2vMean: Derive distributed representation of human language text as the average sum of the word representations. Text vector representation is of same dimension as the size of the word embedding.
- clustTfidf: Consider the corpus vocabulary and cluster the same using Kmeans. Replace the words in the sentences or utterances by their corresponding cluster label Ids. Similar to the classical model of Bag-Of-Words (BOW), derive the TF-IDF representation on the cluster label Ids for the sentence or human utterance. Vector dimension is equal to the number of clusters.
- clustModel: Cluster the corpus vocabulary using Kmeans. Represent the sentences or utterances as sequence of cluster labels in place of the words and build model on the word cluster labels. Use the vector representations of the cluster labels to compute the sentence vectors. The vector representation for text is taken as the average sum of the cluster label representations. Vector

dimension in this case is determined by the model (cluster label model) parameters.

- clustMean: Derive the vector representation for the text as the mean of the word cluster centroids. Text vector dimension is same as the size of the word embedding.

3.2. Clustering Algorithm

Taking queue from the related works quoted in section 2, we use the familiar K-means algorithm equipped with the Euclidean distance for the most part. Euclidean distance may not be the best metric to use in all scenarios, however it already provides us reasonable performance when used in conjunction with the proper choice of semantic representation.

3.3. Cluster Stability Measure

Adjusted Random Index (ARI) is an intuitive approach to comparing clustering solutions. It is based on counting pairs of objects that are classified in the same way in two given solutions. That is, pairs of data points that are in the same cluster (and in different clusters, respectively) under both clustering solutions. An ARI score of 1 indicates point-wise identical cluster structure and the changes to cluster groups may be perceived as perturbation. This reasons out why the argmax of ARI may be chosen as a criterion to measure stability and robustness of a clustering solution.

3.4. Automatic detection of number of clusters

Many researchers see the notion of stability as minimal sensitivity to perturbed versions of the data set. Running a clustering algorithm several times on slightly different data sets may be a way to achieve this. However, the stability or perturbation should be with respect to the process of clustering itself rather than with respect to the data; as our interests rest in evaluating the stability of a fixed clustering algorithm and being able to obtain a consistent clustering solution to base our classification. Hence we combine the following methods to perform multiple runs of the algorithms on the same dataset for generating a diverse set of clustering solutions.

1. Run clustering algorithm using different values (K) for the number of clusters
2. Perform multiple runs of a single algorithm with any given K
3. Use different clustering algorithms - like K-means and GMM

At the true value of K, the variations in the clustering solutions is expected to be at the minimal, thus making the cluster structure more robust and stable. In other words, the set of clustering solutions obtained for such choice of K for the number of clusters tends to be less diverse than the others. Hence compute the average of the ARI scores over all solution pairs of a specific algorithm for specific values of K and compare such average ARI scores over different values of K to determine the optimal value for K. ARI approach to finding K is supported by the arguments as in the stability study discussed in [9]. While the first of the above methods helps determine the "right" number of clusters, the second and third methods help ensure a good consensus on the final cluster memberships.

3.5. Intent Discovery and clustering

The consensus on the cluster memberships may be obtained based on the clustering of clusters (meta-clustering)[20]. Here, the centroid vectors of the clusters generated from the various clustering solutions are clustered. The cluster correspondence of the centroid vectors is then used to map the previously identified cluster labels of the data-points (cluster labels obtained from the different clustering solutions by methods 2 and 3 in section 3.4). The final consensus clusters for the ensemble is obtained, in terms of the cluster correspondences of the cluster centroids, by a majority voting. Experiments over popular benchmark datasets demonstrate the effectiveness of the approach, generating homogeneous clusters with scores comparable to the best available results for cluster classification accuracy. Profiling of consensus clusters, in terms of the frequently occurring words and word-clusters, leads to the semantic intents or labels characterizing the clusters.

Generally, while evaluating a clustering solution, the classes defined by the data labels are viewed as the correct clusters and the clustering algorithms are expected to detect the same. However a valid clustering solution may have many more clusters than there are labels in the data. This is because, as also argued in [11], similar labels may not always correspond to similar objects. There may be natural groups of mutually non-similar points that may share the same label; also a label may take multiple intents or tasks or objects under its umbrella. While points with distinct labels are expected to be assigned to different clusters, different related tasks classified with same intent label may not all fall in the same cluster. Thus our clustering solution may have multiple clusters corresponding to the same intent or semantics, bringing out the human language variants to the same or similar semantics.

4. Experiments and Results

The semantic clustering of words based on the distributed representation are well known and popular. We initially apply the stability approach to the corpus of words. This helps us to see if the procedure for estimating the number of clusters performs well and successfully uncovers the cluster structure of the data vocab. Based on the clustering of the words, the semantic representations for the spoken language text as defined in section 3 are computed. The text data is then clustered based on these representations. The clustering solution with number of clusters K corresponding to the maximum average ARI scores is the choice for the final intent groups by consensus. Intent group labels are compared with the available topic labeling for supervised data to validate the performance of the approach.

To take care not to consider too high values and too low values for K that may produce degenerate cases of clustering (degenerate cases like all data points getting to one cluster, each data point becoming a singleton cluster resulting in as many clusters as the number of data points), we also examine the cluster sizes that the algorithms produce. Given a data set on n points partitioned into k clusters, the expected cluster size of a balance clustering is n/k. We use this in determining the limits for the search range for the number of clusters K.

4.1. Datasets

Table 1 gives a summary of the datasets used in this experimental study.

4.2. Data Preprocessing

Data is pre processed to remove duplicates, stop-words, punctuation, special characters, to convert to lower case and perform stemming. The text was pre-processed in our experiments us-

Table 1: *Dataset details and the respective Cluster Homogeneity Scores. The best scores obtained for the public datasets(1-4) came through numberbatch representation and for the speech data (5-6) came with gloVe.*

Sl. No.	Dataset Name	Number of Docs	Sorted Vocab Size	Average Doc Size (in words)	Categories	Cluster Homogeneity	No. of clusters
1	[UCI] AAAI-13 Accepted Papers Abstract [21]	149	3642	154	12	89.9%	40
2	IMDB [22]	3381	22204	76	19	90.92%	85
3	Classic4 [23]	6902	27646	112	4	93.9%	63
4	Reuters21578 [24]	18526	137433	139	114	89.14%	314
5	Automobile Service Twitter Data	53705	41100	9	383	79.07%	420
6	Customer Care Dialog Data	573315	8583	92	156	83.34%	385

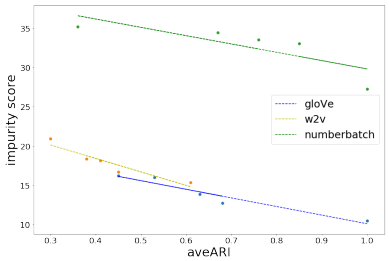

Figure 1: *Impurity vs ARI scores*

ing *gensim* package. We also group together any semantically related intents in the given labelled data.

4.3. Results

Our experiments on supervised data, with semantic labeling, show a positive correlation between the cluster homogeneity (or impurity = 1 - homogeneity) and stability of the clustering. Refer to figure 1. An higher average ARI implies low impurity (or high homogeneity) and a more stable and salient clustering. In a supervised scenario, we do observe that the clustering solution does show the optimal impurity (minimum) or homogeneity (maximum) for this choice of K, evidencing that the process of determining an appropriate number of clusters is closely related to the clustering process itself. Tables 2 and 3 provide the results on the UCI dataset, for aveARI and homogeneity values respectively; homogeneity or impurity scores use the given data labels, where as aveARI does not. Table 1 summarizes the cluster accuracy results on the various datasets. While conventional topic modeling schemes such as probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) need aggregation of short messages to avoid data sparsity in short documents [25, 26], our approach works on large amounts of raw short texts. We have also shown the utility of our experiment framework in intent classification of human utterances and learning sentence variants.

Table 2: *aveARI scores on UCI dataset*

	gloVe	w2v	number-batch
w2vMean	0.63	0.3	0.85
clustTfidf	0.68	**0.61**	**1**
clustModel	0.53	0.38	0.67
clustMean	0.45	0.41	0.36
doc2Vec	**1**	0.45	0.76
Best K	**113**	**99**	**78**

Table 3: *Cluster homogeneity scores on UCI dataset*

	gloVe	w2v	number-batch
w2vMean	86.15%	79.05%	66.94%
clustTfidf	87.25%	**84.62%**	**72.73%**
clustModel	83.97%	81.63%	65.56%
clustMean	83.79%	81.85%	64.82%
doc2Vec	**89.53%**	83.29%	66.46%
Best K	**113**	**99**	**78**

5. Conclusions

In this work, we presented a generic approach to model data intents for spoken language texts using clustering. The experiments and observations show how the average ARI over different values of K is positively correlated to clustering impurity and may effectively be used as a measure of quality for clustering in the case of unlabeled data. Also the stability based choice for the number of clusters clearly improved the cluster accuracy. With the right choice of semantic representation for the spoken language text data, clustering can serve as a tool to spot sentence variants for the same or similar intents. Our approach is scalable, simple and easy to use.

As future directions for research, we may focus on using better semantic representations or hybrid approach to improve efficiency. We may also consider further dividing the clusters using KMeans with DTW distances to arrive at semantically more consistent and meaningful smaller groups of documents, the characteristics of the finer groups corresponding to the semantic text intents. Providing suggested topics or intents to be used can help make the model more robust for a real time application. Designing unsupervised or semi-supervised solution to cluster and discover intents for dialogs may be another direction to look at. We also propose to use the clustering technique to analyze dialog structures and modelling dialog flow act. Adopting for a new language could be yet another direction for work.

6. References

[1] Lior Rokach and Oded Maimon. "Clustering Methods". *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*, chapter 15, pp. 321–352, Springer, 2010.

[2] Anil K.Jain and Richard C.Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[3] Michael Steinbach, George Karypis, Vipin Kumar. A Com-

parison of Document Clustering Techniques". In *Technical Report*, University of Minnesota, pp. 00–34, 2000.

[4] Zhao Y., Karypis G. and Fayyad U. "Hierarchical Clustering Algorithms for Document Datasets". In *Data Mining and Knowledge Discovery*, vol. 10, Issue 2, pp. 141-168, March 2005.

[5] Charu C.Aggarwal, ChengXiang Zhai. "A Survey of Text Clustering Algorithms". *Chapter 4 of MINING TEXT DATA*, pp. 77–128, Springer LLC 2012.

[6] Sandrine Dudoit and Jane Fridlyand. "A prediction-based resampling method for estimating the number of clusters in a dataset". In *Genome Biology*, 3:research0036.1, 2002.

[7] Sandrine Dudoit and Jane Fridlyand. "Bagging to improve the accuracy of a clustering procedure". In *Bioinformatics*, vol. 19, no. 9, pp. 1090-1099, 2003.

[8] Silke Wagner, Dorothea Wagner "Comparing Clusterings - An Overview". *Interner Bericht* series, Universitt Karlsruhe Fakultt fr Informatik 2007. https://books.google.com/books?id=V7x5mQEACAAJ.

[9] Nguyen Vinh and Julien Epps. "A novel approach for automatic number of clusters detection in microarray data based on consensus clustering". In *Proceedings of the Ninth IEEE International Conference on Bioinformatics and Bioengineering*, pp. 84-91. IEEE Computer Society, June 2009.

[10] Ulrike von Luxburg. "Clustering Stability: An Overview" *Foundations and Trends in Machine Learning*, vol. 2, no. 3, pp. 235–274, 2010

[11] Tuong Luu. "Approach to Evaluate Clustering using Classication Labelled Data". *Thesis presented to the University of Waterloo*, Waterloo, Ontario, Canada, 2010.

[12] Wu J. "Selecting External Validation Measures for K-means Clustering". In *Advances in K-means Clustering*, Springer Theses, Berlin, Heidelberg, 2012.

[13] Pengtao Xie and Eric P.Xing. "Integrating document clustering and topic modeling". In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 694–703, 2013.

[14] Vivek Kumar Rengarajan Sridhar. "Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words". In *Proc. of NAACL-HLT 2015*, pp. 192-200, Denver, Colorado, May 31–June 5, 2015.

[15] Chinea-Rios M., Sanchis-Trilles G. and Casacuberta F. "Sentence Clustering Using Continuous Vector Space Representation". In *Lecture Notes in Computer Science*, vol 9117, Springer, Cham. Paredes R., Cardoso J., Pardo X. (eds) Pattern Recognition and Image Analysis, IbPRIA 2015.

[16] Forman G., Nachlieli H. and Keshet R. "Clustering by Intent: A Semi-Supervised Method to Discover Relevant Clusters Incrementally". In *Lecture Notes in Computer Science*, vol. 9286, Springer, Cham. Bifet A. et al. (eds) Machine Learning and Knowledge Discovery in Databases, Springer, Cham, ECML PKDD 2015.

[17] Quoc Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In *Proceedings of the 31st International Conference on Machine Learning, Beijing,China*,JMLR:W&CP, vol. 32, 2014.

[18] Jeffrey Pennington, Richard Socher and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". http://nlp.stanford.edu/pubs/glove.pdf, 2015.

[19] Robert Speer, Joshua Chin and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". https://github.com/commonsense/conceptnet--numberbatch, AAAI 31, pp. 4444–4451, 2017.

[20] Rich Caruana, Mohamed Elhawary, Nam Nguyen and Casey Smith. "Meta Clustering" In *Proceedings of the International Conference on Data Mining*, 2006. http://www.cs.cornell.edu/~nhnguyen/metaclustering.pdf

[21] [UCI] AAAI-13 Accepted Papers Abstract dataset. https://archive.ics.uci.edu/ml/machine-learning-databases/00314

[22] IMDB dataset. https://github.com/hadley/data-movies

[23] Classic4 dataset. Retrieved November 29, 2009 from World Wide Web: ftp://ftp.cs.cornell.edu/pub/smart

[24] Reuters21578 dataset. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

[25] Rubayyi Alghamdi and Khalid Alfalqi. "A Survey of Topic Modeling in Text Mining". In *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, IJACSA 2015.

[26] Xiaojun Quan, Chunyu Kit, Yong Ge, Sinno Jialin Pan. "Short and Sparse Text Topic Modeling via Self-Aggregation". In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2270–2276, IJCAI 2015.