Clustering Prominent People and Organizations in Topic-Specific Text Corpora

ABDULKAREEM ALSUDAIS 1, 2 and HOVIG TCHALIAN 2
1. Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia. Email: A.Alsudais@psau.edu.sa
2. Claremont Graduate University, Claremont, California, USA Email: Hovig.Tchalian@cgu.edu

Abstract

Named entities in text documents are the names of people, organization, location or other types of objects in the documents that exist in the real world. A persisting research challenge is to use computational techniques to identify such entities in text documents. Once identified, several text mining tools and algorithms can be utilized to leverage these discovered named entities and improve NLP applications. In this paper, a method that clusters prominent names of people and organizations based on their semantic similarity in a text corpus is proposed. The method relies on common named entity recognition techniques and on recent word embeddings models. The semantic similarity scores generated using the word embeddings models for the named entities are used to cluster similar entities of the people and organizations types. Two human judges evaluated ten variations of the method after it was run on a corpus that consists of 4,821 articles on a specific topic. The performance of the method was measured using three quantitative measures. The results of these three metrics demonstrate that the method is effective in clustering semantically similar named entities.

1 Introduction

Researchers and scientists often turn to textual data to study social phenomena (Baker, Gabrielatos, and Mcenery 2013; Dimaggio, Nag, and Blei 2013; Jaworska and Krishnamurthy 2012). Mining those data allows researchers to generate insights about the grammatical features, linguistic content and social context of such phenomena (Li and Dash 2010; Yuan, Lau, and Xu 2016). Researchers use common text mining methods such as topic modeling and entity extraction to explore text corpora, identifying patterns and generating insights and also confirming and disconfirming hypotheses. Generating genuine insights through these methods requires the development of novel algorithms but, equally, their rigorous application in methods that are being continuously developed. Therefore, novel text mining algorithms that produce specific, consistent and relevant results are especially beneficial for many business intelligence systems and tools (Liu, Li, and Thomas 2017).

When examining a text corpus, one preliminary step in generating specific, relevant insights is to identify prominent entities, such as individuals, organizations and places mentioned in the corpus. Named Entity Recognition (NER) refers to the computational task of identifying real-world entities in text documents. Once identified and tagged, named entities may be utilized in information systems to enhance applications and tools as well as to generate useful information. The process of identifying and ranking the most prominent named entities in a corpus is often used in commercial text mining tools, which rely on popular NER algorithms. The need to generate relevant insights in text mining methods has substantially raised the bar for NER, necessitating additional investigations into the relatedness and similarity across tagged entities to locate groups of actors that
Influence and help structure a discourse. While earlier entity recognition methods helped answer simple questions about the frequency and occurrence of entities, newer methods help answer more complex questions about who the actors are (political, economic, technical), the consistent patterns of co-occurrence they exhibit, and their shared interests in relation to a topic of interest. While generating new analytical possibilities, however, many novel NER features have suffered from questionable rigor and replicability. The novel text mining method known as word embeddings promises to remedy this situation. Word embeddings refers to models that create dense vector representations for words or phrases in a text corpus by utilizing their immediate syntactic context, defined by a window of proximate terms (Alsuadais, Tchalian, and Hilton 2016). These models have recently gained popularity, due in large part to the higher analytical bar required of text mining methods. Their greater analytical power is in turn due to advances in computing power that have made the creation of vectors for large corpora more feasible and less time- and compute-intensive. Successful models for generating vector representations for words and phrases such as Skip-gram and CBOW (Mikolov et al. 2013), GloVe (Pnnington, Socher, and Manning 2014), and others (Levy and Goldberg 2014; Turian, Ratinov, and Bengio 2010) have proven successful in performing various language-related analytics tasks. These tasks include solving analogies and translating text across natural languages with the use of dictionaries. One of the most beneficial features of these new word embedding methods captures semantic similarities across words and phrases in a corpus. This similarity feature may be used, for example, to capture and detect the semantic changes in the meaning of particular words over a pre-defined period of time (Hamilton, Leskovec, and Jurafsky 2016; Kulkarni et al. 2015). This application demonstrates the potential of using the new feature to solve new and challenging questions in the text-mining domain. The objective of this paper is to resolve one of these questions: Is it possible to create an NLP pipeline that reliably and consistently clusters named entities (people and organizations) in a corpus by utilizing their semantic similarity as generated by a semantic word embeddings model?

In this paper, semantic similarities generated by Skip-gram and CBOW are employed to cluster the most prominent named entities in a corpus. Clustering named entities is a recognized approach for leveraging tagged entities (Hasegawa, Sekine, and Grishman 2004). When clustering named entities, similar named entities are grouped within the same cluster based on their similarity values, which are detected using a purpose-built, reliable similarity function. Thus, an underlying metric for capturing the similarities between the entities must be utilized. The method offers a rigorous and reliable alternative to simple clustering methods, offering information systems researchers a replicable method for generating results and insights in various text mining and NER applications.

When training a Skip-gram or CBOW model on a topic-specific corpus, one may also discover relationships between words and entities in the collection that are otherwise latent and difficult to detect. According to Thomas Mikolov, the chief scientist behind the two models, the learned word vectors are only effective if they are used to accomplish other tasks. For example, Mikolov suggests using word vectors and K-means clustering to create classes for the words in the collection. While that is useful and currently achievable using the word2vec software, additional modifications and customizations are required to implement problem-specific solutions that build on that idea. Generating additional insights in a consistent, replicable way from the relationship among tagged named entities therefore offers information systems researchers numerous benefits, including enhancing Question Answering (QA) systems and automating the population process in ontologies (Marrero et al., 2013). This paper develops a novel full-scale, reliable and replicable method for semantically clustering similar named entities, primarily of the “people” and “organization” type, based on the similarity scores as generated by a Skip-gram or CBOW model. It does so by: (1) using NER to locate named entities in the text, such as the names of real-world organizations; (2) using semantic similarity vectors as captured by a word embedding model; (3) deploying a reliable algorithm to cluster the entities according to the resulting vectors. The preliminary empirical findings and quantitative evaluation of the method demonstrate that the it is successful in clustering and grouping named entities that have similar roles in the text or are members of the same abstract classes.
2 Methodology

In this section, the method proposed to identify and cluster named entities in a text corpus is described. The method relies on three major components: named entity recognition, word embeddings, and clustering. Figure 1 demonstrates the three major steps in the method and the relevant tasks that take place in each step. In the first step, named entities in the corpus are identified and then ranked according to the frequency in which they were used in the articles of the corpus. In the second step, a word embeddings model is applied on the collection to generate semantic similarity scores across the named entities in the corpus. In the final step, a clustering operation is performed on a symmetrical matrix that consists of top named entities to identify clusters of similar individuals and organizations in the corpus. This section describes the three steps, the rationale behind each task proposed for each step, and the underlying assumptions that were made.

2.1 Named Entity Recognition

Named entities refers to the names of people and organizations that exist in the real world. Named Entity Recognition (NER) is the task of processing a body of text and classifying segments in the text that are named entities. Researchers have been attempting to solve this problem for decades (Nadeau and Sekine 2006). Several highly accurate and reliable algorithms for capturing named entities currently exist (Finkel, Grenager, and Manning 2005; McCallum and Li 2003). Recent work in the area has focused on increasing the accuracy of named entity taggers by leveraging advances in neural networks (Chiu and Nichols 2016; Lample et al. 2016). However, these new models are only slightly more effective than their traditional counterparts. Several named entity extraction methods target unique types of corpora such as tweets (Habib and Van Keulen 2015; Ritter, Clark, and Etzioni 2011) and biomedical data (Habibi et al. 2017; Tang et al. 2015). Additionally, certain methods are designed for specific languages such as Chinese (Lei et al. 2014; Peng and Dredze 2016) and Arabic (Althobaiti, Kruschwitz, and Poesio 2015; Oudah and Shaalan 2016).

There are multiple categories for named entities including the names of people, organizations, locations, and time units. Since the primary motivation of the proposed method is to semantically cluster similar people and organizations, named entities of other types such as “location” and “time” have been discarded. However, other named entity types should be considered in future extensions of the method.

In this paper, Stanford NER (Finkel et al., 2005) is used in the first step in the proposed method. Stanford NER, which is based on an older mechanism of detecting named entities, is a highly accurate tagger that performs well on text extracted from newspaper articles. According to Finkel et al. (2005), Stanford NER achieved an F1 score (a composite measure employed to determine the accuracy of a classifier) of 86% when trained on the CoNLL-2003 named entity dataset (Tjong et al., 2003). The NER process followed in this paper is initiated by tokenizing each article in the collection individually, in order to locate the set of sentences in the article. Subsequently, the Stanford NER program is run on each extracted sentence.
The next task in this step is entity linking. Entity linking is the process of matching different segments of text that belong to the same real-world entity. For instance, “Barack Obama” and “Barack H. Obama” refer to the same person, and an entity linking procedure should be able to recognize this and match the two variations of the name accordingly. To increase the performance of the proposed method, a “soft” entity linking process is performed on tagged names of people at the article-level. An attempt to match entities is performed by taking advantage of the fact that in English, full names of individuals are commonly written in their entirety the first time they are mentioned in an article, whereas subsequently, they are referred to only by their last name. The task relies on removing middle initials and the word “Mr.” that often precedes people’s last names from tagged named entities in these later mentions. Therefore, the proposed task of entity linking relies on inspecting whether a “last name” in the article has been mentioned earlier as a portion of a full name. Future implementations of this approach will extend the methodology to recognize other titles prefixing a person’s name such as “Dr.,” “Mrs.,” “Miss,” etc. For organizational names, no attempts were made to match entities. For example, “SEC” and “Securities and Exchange Commission” were captured as separate entities. This is a limitation of the current method that will also be explored in future extensions.

The final task in the named entities recognition process involves replacing named entities in each article with their full names as captured in the “entity linking” task with their corresponding entity type. For example, each occurrence of the names “Barack Obama” and “Barack H. Obama” would be resolved to “Barack_Obama_PER” in a given article. The underscores replace empty spaces so that the word embeddings model can recognize the name as a single entity and then process it accordingly. The tag “_PER” is used to indicate that the named entity is of the “person” type while the tag “_ORG” is used to indicate that the named entity is of the “organization” type. The types are added to the names so that they may be later retrieved using a word embeddings model.

2.2 Word Embeddings

Word embeddings refers to models that create dense vector representations for words or phrases in a text corpus by utilizing their immediate syntactic context, defined by a window with proximate terms. Word embeddings models (Baroni et al., 2014), which rely on attempts to “predict” word vectors, are better at capturing semantic similarities between words as compared to traditional “counting” methods that use either Positive Pointwise Mutual Information (PPMI) or Local Mutual Information (LMI) to weigh the features in the vectors. Researchers have demonstrated how this similarity feature available in word embeddings models may be employed to accomplish several challenging language-related tasks.

Once all the named entities in the corpus are tagged, the next step is to run a word embeddings model on the modified corpus. The objective of this step is to detect the semantic similarity across named entities in the corpus. In this study, two popular word embeddings models, CBOW and Skip-gram, are tested and applied on a text corpus. Both models are used to investigate whether the performance of the proposed method is affected by the model that is employed to capture the similarity between the named entities. The implementation of CBOW and Skip-gram in the Python package gensim (Rehurek and Sojka 2010) is used to run the models. Running the two models generates word vectors for named entities in the collection. These vectors and similarity scores are used in the following step to cluster the named entities.

2.3 Clustering Prominent Named Entities

Clustering is a traditional data mining technique that is used to group similar items based on an underlying similarity metric. Clustering may be applied to different types of data including text. In the text mining domain, clustering may be used to group text segments such as words, terms, sentences, topics, or documents. The results of clustered text data may be utilized to enhance text mining tasks such as corpus summarization and document classification (Aggarwal and Zhai 2012). In this paper, clustering is applied on a matrix that consists of the top named entities in a collection and their similarity values. The purpose of this clustering procedure is to discover semantic roles, labels, and categories.

Several clustering algorithms currently exist, and for this paper, K-means is used as the clustering algorithm. In K-means, objects are portioned into different groups where each cluster contains at least one item. No items may be placed into more than one cluster. The number of clusters ‘K’ must be specified prior to initiating the algorithm. Choosing
the most appropriate K is an ongoing research problem. For the proposed method, K is calculated by dividing the number of top terms to be clustered by the number 10, a common practice when using K-means clustering. Implementing a different method of calculating K or using a clustering algorithm other than K-means such as the silhouette method or X-means clustering might produce better results. These methods were not employed in this study and present opportunities for future research and refinement.

The final task in the method is to identify and cluster the N top named entities in the corpus. To accomplish this goal, a symmetrical N x N matrix is first constructed. In this matrix, the columns and rows consist of all the named entities in the top N named entities list. For each entity in the rows section, a vector is created. The vector consists of the similarity values as generated by the word embeddings model between the entity and the corresponding entities in the columns section. The concluding step is to then cluster the entities in the rows section using K-means clustering.

The top terms lists are generated by counting the number of articles for each named entity in the corpus where the named entity appeared and then selecting the top N named entities. The number of articles was used as a unit of measure rather than the raw frequency to provide additional rigor and to avoid assigning a large weight to named entities that appeared many times in the corpus but in only a small number of articles. Using a larger corpus is recommended, since a larger set of terms is more likely to accurately and precisely identify a subset of top terms. (See further details in section 2.4.) To maintain the automation of the method, qualitatively refining and filtering terms in the top terms lists was avoided. Thus, terms that are erroneously tagged as named entities and appear in the top terms list are kept, even though they contribute negatively to the results. The rationale behind this approach is to refrain from using experts’ opinions to modify the lists. The approach allows greater replicability of the method by relying exclusively on quantitative results, without subjective modifications by experts, while sacrificing very little in accuracy and precision.

In this paper, five different top named entities lists are tested. The lists differ in two core variables: (1) the number of terms in the list (either 100 or 200) and (2) the type of named entities in the list. Two of the lists contains entities of the “people” type, two of the lists contains entities of the “organizations” type, and one list contains entities of both the “people” and “organization” types.

2.4 Data

The method proposed in this paper was developed for use on a large text corpus. A large corpus is required because word embeddings models perform best on large corpora and often fail when applied on small ones. In this study, a large text corpus on “corporate governance” was used (Alsudais and Tchalian 2016). Table 1 includes the sources used to build the dataset.

| Newspaper          | Number of Articles |
|--------------------|--------------------|
| The Wall Street Journal | 2,381              |
| The New York Times      | 1,090              |
| The Washington Post       | 701                |
| The Los Angeles Times       | 649                |
| Total                    | 4,821              |

Table 1. Summary of data sources

3 Experiment

In this section, an experiment to test and evaluate the proposed method is explained. The method was tested on the “corporate governance” corpus described in the previous section. Five top terms lists were investigated. The lists were: (1) top 100 tagged entities of the “person” type (T100_P); (2) top 200 tagged entities of the “person” type (T200_P); (3) top 100 tagged entities of the “organization” type (T100_O); (4) top 200 tagged entities of the “organization” type (T200_O); and (5) a combined list of the top 100 tagged entities of both the “person” and “organization” types (T200_PO). Both CBOW and Skip-gram models were used to generate vectors for the named entities. The models’ implementations in the Python package gensim (Rehurek and Sojka 2010) were used. For both models, the number of dimensions was set to 500 and the window size was set to 10. For each one of the five lists, two N x N matrices were constructed with the semantic similarity between the variables in the row and column as the values of the cells. Therefore, ten matrices were constructed. The values in the first matrix were derived from the results generated using CBOW whereas the values in the second were derived from the results generated using Skip-gram. The purpose of testing these variations was to investigate whether the changes in the models or the type of input would generate improved results.
3.1 Evaluation Metrics

For quantitative evaluation, three metrics were used. These metrics relied on two domain experts who manually reviewed the results. These experts were asked to complete two tasks: (1) indicate whether a logical and semantic role or class might be inferred and used to label at least the majority of terms for each cluster identified by the method; and, if that was the case, (2) indicate for each named entity in the cluster whether the entity belongs to the category identified for the cluster. Accordingly, three evaluation metrics were used. In text mining, these metrics are typically used to evaluate topic modeling and clustering algorithms. The first metric is a coherence measure that is used to evaluate the average coherence of the clusters. The second metric is a standard precision measure that is employed to quantify how “precise” the method is in assigning terms to clusters. The third metric is a coherent clusters metric that is used to quantify the number of clusters that comprise a meaningful class or category and may be tagged with a class label.

3.1.1 Coherence Measure

The coherence measure is a metric that is commonly used when evaluating topic models. Coherence measures have been used in other studies (Qiang et al., 2016; Xie & Xing, 2013; Xie et al., 2015). There are many automated methods for producing coherence scores. However, it has been argued that coherence values based on human judgment are still superior (Röder et al., 2015). Additionally, many of these automated metrics rely on co-occurrence values between two terms at the sentence or article levels. This makes the metrics unsuitable for use in the proposed method since two terms might be accurately placed in the same cluster according to the proposed method even if their co-occurrences value was low. The terms would be placed in the same cluster since they are semantically similar because they share similar neighboring words. An example of this is a cluster labeled “corrupt CEOs” where accurately placed terms in the cluster would be names of corrupt CEOs who may have similar neighboring words such as “CEO,” “crooked,” and “corrupt” yet still have low values as generated by traditional co-occurrence-based metrics.

Human judges are therefore used to evaluate the coherence or clusters for purposes of developing and testing the method proposed in this paper. The coherence of a cluster is calculated based on the ratio of the number of relevant terms in the cluster to the total number of terms in the cluster. A judge is asked to infer a label for an examined cluster. If a label can be found, the judge manually tags irrelevant terms in the cluster that do not fit and therefore do not belong to the cluster. The overall coherence for a method is calculated by averaging the individual coherence values for the clusters in the set. The formula for calculating the coherence value of a cluster is:

\[
\frac{\text{Number of relevant named entities in cluster}}{\text{Number of named entities in the cluster}}
\]

Similar to the evaluation process in Qiang et al. (2016), all terms within a cluster are considered irrelevant if the judge is unable to infer a semantic label or class for the cluster.

3.1.2 Precision Measure

For the precision measure, the values are calculated by considering the number of True Positive (TP) and False Positive (FP) named entities in the results previously tagged by the judges. The standard formula for calculating the precision value is:

\[
\frac{TP}{TP + FP}
\]

A TP named entity is defined as a named entity that has been accurately placed in a cluster and confirmed by a judge. Conversely, a FP named entity is an entity that the judge has tagged as an entity that has been inaccurately placed in the cluster. Similar to the process followed for the coherence measure, all entities in a cluster are tagged as false positive (Type I error) if the judge is unable to infer a semantic label or class for the cluster.

3.1.3 Coherent Clusters Measure

The third measure used in this paper is a coherent clusters measure. The purpose of this measure is to quantify the number of clusters that are labeled as valid by the judges. The formula for calculating the coherent clusters measure is the following:

\[
\frac{\text{Number of clusters tagged as "accurate"}}{\text{Total Number of Clusters (K)}}
\]

The judges were instructed to tag a cluster as accurate if at least the majority of the entities in the cluster could be grouped under the same semantic class or label. An example of a label for a cluster would be “corrupt CEOs” or “companies that faced
scandals in the past.” Since the previous two measures were heavily influenced by false positive terms, this metric could be used to provide additional insights on the performance of the method and to measure the method’s ability to produce strong clusters even when some of the clusters contain several noisy terms. Furthermore, this measure might be employed to hint at how the results of the method might be interpreted by a domain expert who might manually eliminate problematic terms that negatively affect otherwise strong clusters.

4 Results

In this section, results of running the method on the “corporate governance” corpus using the previously described top named entities lists as well as both CBOW and Skip-gram are described. The judges were asked to analyze the generated clusters using ten variations of the method. The judges were given the results of the ten variations without any indicators of the underlying model or list used. Selected clusters from the first nine methods are displayed in Appendix A. Named entities that judges deemed unfit for a discovered cluster are designated in italics. The first line in each table describes the type of top named entities list and word embeddings model used. For instance, T100_O and Skip-gram refer to the list of the top 100 named entities of the “organizations” type and to the Skip-gram model. The second line in each table includes the labels created by the judge for the cluster. The third row in each table contains the terms within the clusters.

While improvements can be made, these results demonstrate the method’s ability to detect semantically similar clusters of people and organizations in a text corpus. The method was quantitatively evaluated based on the three quantitative measures: coherence measure, precision measure, and coherent clusters measure.

The results for the three measures indicate that the method was successful in detecting semantically similar people and organizations. Under all three measures, one of the judges evaluated the method slightly more favorably than the other. One possible explanation for this is that one of the judges tagged some of the broad clusters as incoherent. With respect to the performance of CBOW and Skip-gram, averaging the results for all ten variations indicated that the results were more accurate when using Skip-gram. These findings are identical to those of previous studies that compared the performance of these two methods for various language-related tasks (Chen et al., 2015).

The quantitative results suggest that the method performed best when the lists included only people or only organizations. The best results were achieved when the top 100 named entities tagged with the “organization” type was processed using Skip-gram and then clustered using a K size of 10. The coherence and precision measures were 80.8% and 80.4%, respectively, and the judges found that eight and one half out of the ten clusters were valid. The results of the coherence measure indicate that the average values for all variations were higher than 50%. Table 2 includes the detailed results of the coherence measures for the ten variations of the method. The results show that the method performs slightly higher when Skip-gram was used and that the most coherent clusters were identified when Skip-gram was employed to cluster the entities in the top 100 organizations list.

The results of the precision measure were similar to those of the coherence measure, with Skip-gram averaging a few percentage points higher than CBOW. Table 3 shows a complete breakdown of the precision values for the ten variations. The best overall precision value was attained when Skip-gram and the T200_O list were used. The lowest value was 58.7% suggesting that the method provided useful information even when performing at its worst. Several inaccurately tagged named entities that were meaningless and did not represent a real-world person or organization, such as “John” and “Messes,” negatively affected the performance of the precision measure as they were flagged as unfit by the judges.

The third and final measure was the coherent clusters measure, which was based on counting the number of clusters that the judges deemed as being logical and accurately representing a semantic category. The results of this measure were the most promising. For many of the variations, the judges found that 80% of the groups captured by the method were valid and coherent. This suggests that the method may be produce more accurate results and demonstrate greater precision if a domain expert examines the results and removes unfit terms from each cluster. Users of the method can use these tradeoffs to help balance automation with error rates for individual cases.
| List: | T100_P | T200_P | T100_O | T200_O | T100_PO | Average |
|------|--------|--------|--------|--------|---------|---------|
| K size: | 10 | 20 | 10 | 20 | 20 | |
| Judge 1 | CBO | 44.2% | 54.0% | 55.3% | 36.4% | 50.42% |
| | SkipGram | 54.2% | 82.0% | 71.8% | 29.4% | 57.54% |
| Judge 2 | CBO | 72.5% | 73.4% | 73.9% | 74.0% | 71.72% |
| | SkipGram | 81.7% | 79.6% | 76.6% | 77.6% | 75.40% |
| Average | CBO | 58.35% | 62.2% | 55.3% | 36.4% | 50.42% |
| | SkipGram | 66.7% | 82.0% | 71.8% | 29.4% | 66.47% |
| Both | 63.15% | 61.5% | 69.40% | 54.35% | 63.77% |

Table 2. Results of the coherence measure

| List: | T100_P | T200_P | T100_O | T200_O | T100_PO | Average |
|------|--------|--------|--------|--------|---------|---------|
| K size: | 10 | 20 | 10 | 20 | 20 | |
| Judge 1 | CBO | 48.0% | 52.0% | 62.0% | 36.0% | 51.50% |
| | SkipGram | 57.0% | 82.8% | 80.5% | 31.5% | 63.06% |
| Judge 2 | CBO | 86.5% | 92.7% | 71.5% | 83.8% | 80.90% |
| | SkipGram | 90.8% | 78.0% | 86.0% | 86.0% | 81.56% |
| Average | CBO | 67.25% | 72.35% | 66.75% | 59.90% | 66.20% |
| | SkipGram | 73.90% | 80.40% | 83.25% | 58.75% | 72.31% |
| Both | 70.58% | 76.38% | 75.00% | 59.33% | 69.26% |

Table 3. Results of the precision measure

| List: | T100_P | T200_P | T100_O | T200_O | T100_PO | Average |
|------|--------|--------|--------|--------|---------|---------|
| K size: | 10 | 20 | 10 | 20 | 20 | |
| Judge 1 | CBO | 50% | 70% | 70% | 45% | 62% |
| | SkipGram | 70% | 90% | 80% | 40% | 68% |
| Judge 2 | CBO | 90% | 90% | 80% | 90% | 85% |
| | SkipGram | 90% | 80% | 90% | 90% | 85% |
| Average | CBO | 70% | 80% | 75% | 68% | 74% |
| | SkipGram | 80% | 85% | 85% | 65% | 77% |
| Both | 75% | 82% | 80% | 66% | 75% |

Table 4. Results of coherent clusters measure

5 Discussion and Conclusion

In this paper, a new method that utilizes existing and accepted techniques to cluster named entities of the “person” and “organization” types in a topic-specific corpus based on the semantic similarities between the entities as predicted by the CBO and Skip-gram models is introduced. The method was tested on a corpus that consists of news articles containing the term “corporate governance” published in four of the leading newspapers in the United States between 1978 and 2004. The results demonstrate the effectiveness of the method when evaluated using quantitative metrics.

Observations with respect to the results of the process performed by the two annotators suggest that the method effectively captured semantic clusters of named entities. For example, when using Skip-gram and the list of top 100 organizations, investment banks was one of the identified clusters (Appendix A). The investment banks cluster included banks such as “Morgan Stanley” and “Merrill Lynch”. Another cluster found while using Skip-gram and the list of top 100 organizations was a narrower cluster that included companies dealing with financial crises in the early 2000s and two of their auditors. This kind of information might be valuable for non-domain experts as it captures information that is specific to a topic-specific corpus. An additional observation was that changing the list of top named entities affected the performance of the clustering procedure. Using lists of only people or only organizations generated more coherent results. Furthermore, results indicated that using Skip-gram generated more coherent clusters when compared to CBO. The result of the quantitative measures agreed with these observations.

As shown in Appendix A, it was observed that some of the clusters were narrow and defined
whereas others were broader and more abstract. Narrow clusters are ones that are unique to the dataset and the analyzed corpus. Example of such clusters are “corporate governance experts” and “companies in crisis in the 1980s.” On the other hand, broad clusters are those that represent semantic categories that are more universal such as “universities” and “financial services firms.” Table 5 includes a sample of narrow and broad clusters.

| Narrow Clusters                          | Broad Clusters                      |
|------------------------------------------|-------------------------------------|
| Corporate governance thought leaders     | Universities                        |
| Companies in crisis in 1980s             | Company CEOs and founders            |
| Companies experiencing financial scandals early 2000s | Large, publicly-traded companies |
| Corporate governance experts             | Financial services firms             |
| Company executives involved in financial or investment scandal | Company executives |
| Corporate governance associations        | American business executives         |
| Corporate governance experts             | American government executives       |
| SEC leaders (chairs or commissioners)    |                                     |

Table 5. Examples of narrow and broad clusters

While the results are promising, there remain several limitations in the method and the paper. First, the method was tested and evaluated on a single corpus. Thus, testing the method on a different dataset might not produce results with higher error rates than those attained in this corpus. Additionally, latent issues in the method might be revealed. Second, due to the historical context of the corpus used in this paper, it was difficult for the two judges to fairly evaluate some of the extracted clusters. Some of the clusters represented narrow and specific categories that might be challenging to interpret and label, even for domain experts, suggesting that some of the reported results may have fewer errors than those reported in the paper. Finally, while entity linking techniques were used for entities of the person type, similar techniques were not employed for entities of the organization type. Thus, many replicates existed in the top organizations lists.

There are also several areas where the method employed in this paper may be enhanced in future work. First, the techniques used to generate the top terms lists may be improved. Currently, the top named entities lists are generated by counting the number of articles that named entities appear within, and then ranking the named entities accordingly. It is possible that the use of more complex alternatives that leverage additional information pertaining to the entities could significantly improve the results. Second, the current method assumes a fully automated process with no expert intervention. The coherence of the clusters will improve if a domain expert is providing expert feedback at various points. This expert feedback may occur after generating the top terms lists by having the expert flag terms that are inaccurately tagged as named entities. Third, since the method was performed on a text corpus with a long duration, running the method on different time periods in the collections might produce different or more informative results. For example, one of the identified clusters was labeled “companies facing crises in the early 2000s.” This category would clearly not have been identified if the method was run on a period consisting of articles published from 1995 to 1999. Running the analysis on a different text corpus with a narrower or less historically bounded context may provide fewer time markers and therefore produce different results. Finally, the application of the method on time periods in the collection might provide additional insights with respect to the evolution of semantic

References

Aggarwal, Charu C, and Chengxiang Zhai. 2012. “Chapter 4 A Survey of Text Clustering Algorithms.” In Mining Text Data, Springer, 77–128.

Alsudais, Abdulkareem, and Hovig Tchalian. 2016. “Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus.” In Proceedings of the 22nd Americas Conference on Information Systems (AMCIS), San Diego, California, August 11-13, 2016,.

Alsudais, Abdulkareem, Hovig Tchalian, and Brian Hilton. 2016. “Labeled Topics for News Corpora Using Word Embeddings and Keyword Identification.” In IJCAI Workshop on NLP Meets Journalism,.

Althobaiti, Maha, Udo Kruschwitz, and Massimo Poesio. 2015. “Combining Minimally-Supervised Methods for Arabic Named Entity Recognition.” Transactions of the Association for Computational Linguistics.
Baker, Paul, Costas Gabrielatos, and Tony Mcenery. 2013. “Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word ‘Muslim’ in the British Press 1998–2009.” Applied Linguistics 34(3): 255–78.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. “Don’t Count, Predict! A Cystematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors.” In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA: Association for Computational Linguistics, 238–47.

Chen, Xinxiong et al. 2015. “Joint Learning of Character and Word Embeddings.” In IJCAI’15 Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 1236–42.

Chiu, Jason P C, and Eric Nichols. 2016. “Named Entity Recognition with Bidirectional LSTM-CNNs.” Transactions of the Association for Computational Linguistics. 4: 357–70.

Dimaggio, Paul, Manish Nag, and David Blei. 2013. “Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding.” Poetics 41(6): 570–606. http://dx.doi.org/10.1016/j.poetic.2013.08.004.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. “Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling.” In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), , 363–70.

Hasegawa, Takaaki, Satoshi Sekine, and Ralph Grishman. 2004. “Discovering Relations among Named Entities from Large Corpora.” In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL ’04), Association for Computational Linguistics, Stroudsburg, PA, USA.

Jaworska, Sylvia, and Ramesh Krishnamurthy. 2012. “On the F Word: A Corpus-Based Analysis of the Media Representation of Feminism in British and German Press Discourse, 1990-2009.” Discourse & Society 23(4): 401–31.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. “Statistically Significant Detection of Linguistic Change.” In Proceedings of the 24th International Conference on World Wide Web (WWW ’15), Florence, Italy, 625–35.

Lei, Jianbo et al. 2014. “A Comprehensive Study of Named Entity Recognition in Chinese Clinical Text.” J Am Med Inform Assoc 21(5): 808–14.

Levy, Omer, and Yoav Goldberg. 2014. “Dependency-Based Word Embeddings.” In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, Maryland, USA: Association for Computational Linguistics, 302–8.

Li, Nan, and Desheng Dash. 2010. “Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast.” Decision Support Systems 48(2): 354–68. http://dx.doi.org/10.1016/j.dss.2009.09.003.

Liu, Dapeng, Yang Li, and Manoj A Thomas. 2017. “A Roadmap for Natural Language Processing Research in Information Systems.” In Proceedings of the 50th Hawaii International Conference on System Sciences, , 1112–21.
Recognition: Fallacies, Challenges and Opportunities.” *Computer Standards & Interfaces* 35(5): 482–89.

Mccallum, Andrew, and Wei Li. 2003. “Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons.” In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL '03)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 188–91.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *Advances in Neural Information Processing Systems* 26: 3111–19.

Nadeau, David, and Satoshi Sekine. 2006. “A Survey of Named Entity Recognition and Classification.” (1991): 1–20.

Oudah, Mai, and Khaled Shaalan. 2016. “NERA2.0: Improving Coverage and Performance of Rule-Based Named Entity Recognition for Arabic.” *Natural Language Engineering* 23(3): 441–72.

Peng, Nanyun, and Mark Dredze. 2016. “Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning.”

Pinnington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 1532–43.

Qiang, Jipeng, Ping Chen, and Tong Wang. 2016. “Topic Modeling over Short Texts by Incorporating Word.”

Rehurek, Radim, and Petr Sojka. 2010. “Software Framework for Topic Modelling with Large Corpora.” In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Ritter, Alan, Sam Clark, and Oren Etzioni. 2011. “Named Entity Recognition in Tweets: An Experimental Study.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1524–34.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures.” In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

Tang, Zhuo et al. 2015. “CRFs Based Parallel Biomedical Named Entity Recognition Algorithm Employing MapReduce Framework Recognition.” *Cluster Computing* 18(2): 493–505. http://dx.doi.org/10.1007/s10586-015-0426-z.

Tjong, Erik F, Kim Sang, and Fien De Meulder. 2003. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.”

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. “Word Representations: A Simple and General Method for Semi-Supervised Learning.” In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–94.

Xie, Pengtao, and Eric P Xing. 2013. “Integrating Document Clustering and Topic Modeling.” In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.

Xie, Pengtao, Diyi Yang, and Eric P Xing. 2015. “Incorporating Word Correlation Knowledge into Topic Modeling.” In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Yuan, Hui, Raymond Y K Lau, and Wei Xu. 2016. “The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach.” *Decision Support Systems* 91: 67–76. http://dx.doi.org/10.1016/j.dss.2016.08.001.
**Appendix A.** Sample of clusters found, their labels, and the entities in the clusters. Entities in red colors are ones that were inaccurately placed in a cluster

| Count | Method            | Label                                                                 | Entities                                                                   |
|-------|-------------------|----------------------------------------------------------------------|---------------------------------------------------------------------------|
| 1     | T100_O & skip-gram | Companies experiencing financial scandals in early 2000s and their auditors | 1. Enron Corp  
2. Enron  
3. WorldCom  
4. WorldCom Inc.  
5. Tyco International Ltd.  
6. Pricewaterhouse Coopers  
7. MCI  
8. Tyco International  
9. Tyco |
| 2     | T100_O & skip-gram | Investment Banks                                                       | 1. Merrill Lynch  
2. Goldman Sachs  
3. Deutsche Bank  
4. Morgan Stanley  
5. Credit Suisse First Boston  
6. Lehman Brothers |
| 3     | T100_P & skip-gram | Company leaders involved in financial or investment scandal           | 1. Arthur Andersen  
2. Dennis Kozlowski  
3. Bernard Ebbers  
4. Kenneth Lay  
5. Morgan Stanley  
6. Fannie Mae  
7. Martha Stewart  
8. Mark Swartz  
9. **Freddie Mac** |
| 4     | T100_P & skip-gram | Corporate Governance thought leaders                                   | 1. Ira Millstein  
2. John Coffee  
3. Weil  
4. Graef Crystal  
5. Joseph Grundfest  
6. **Spencer Stuart**  
7. Jay Lorsch |
| 5     | T100_O & CBOW      | Regulatory bodies                                                      | 1. Congress  
2. Senate  
3. Federal Reserve  
4. House  
5. Delaware Chancery Court  
6. Senate Banking Committee  
7. European Union  
8. Supreme Court |
| 6     | T100_O & CBOW      | Universities                                                           | 1. Columbia University  
2. Harvard Business School  
3. **Gotshal & Manges**  
4. Stanford University  
5. Harvard University  
6. Harvard |
| 7     | T100_P & CBOW      | Activist Investors/officials                                            | 1. Richard Koppes  
2. Ira Millstein  
3. Ralph Whitworth  
4. Sean Harrigan  
5. Dale Hanson  
6. Phil Angelides  
7. Alan Hevesi |
| Page | Technique & Technique | Cluster | Members |
|------|------------------------|---------|---------|
| 8    | T100_P & CBOW          | Corporate Governance Experts | 1. Nell Minow  
2. Patrick McGurn  
3. Sarah Teslik  
4. Ann Yerger  
5. Carol Bowie  
6. Paul Hodgson  
7. Ken Bertsch |
| 9    | T100_PO & skip-gram    | Companies in Crisis in 1980s | 1. Vivendi  
2. Walt Disney Co  
3. Compaq Computer Corp  
4. IBM  
5. GE  
6. Chrysler Corp  
7. GM  
8. Hewlett-Packard  
9. Comcast Corp |
| 10   | T100_PO & skip-gram    | SEC Officials | 1. William Donaldson  
2. Harvey Pitt  
3. Arthur Levitt  
4. John Reed  
5. Carl McCall  
6. Richard Breeden  
7. Henry Paulson  
8. Franklin Raines  
9. William Webster |
| 11   | T100_PO & CBOW         | Failed Cluster | 1. European Union  
2. International Monetary Fund  
3. Labor Department  
4. World Bank  
5. Supreme Court  
6. Philippines |
| 12   | T100_PO & CBOW         | Failed Cluster | 1. Capitol Hill  
2. House Financial Services Committee  
3. Michael Oxley  
4. John McCain  
5. Senate Banking Committee  
6. Paul Sarbanes |
| 13   | T200_O & skip-gram     | Accounting Firms | 1. PricewaterhouseCoopers  
2. Ernst & Young  
3. Xerox Corp  
4. Arthur Andersen LLP  
5. Ernst & Young LLP  
6. Deloitte & Touche  
7. KPMG |
| 14   | T200_O & skip-gram     | Executive search and compensation firms | 1. KornFerry International  
2. Pearl Meyer & Partners  
3. Cleveland |
| 15   | T200_P & skip-gram     | American Business Executives | 1. Berkshire Hathaway  
2. Jean-Marie Messier  
3. Rupert Murdoch  
4. Conrad Black  
5. Bill Gates  
6. Fox  
7. Time Warner  
8. Barry Diller  
9. Ted Turner |
| Page | Model | Cluster | Items |
|------|-------|---------|-------|
| 16   | T200_P & skip-gram | Failed Cluster | 1. Bush  
2. Vladimir Putin  
3. George Bush  
4. Mikhail Khodorkovsky  
5. Boris Yeltsin  
6. God  
7. Saddam Hussein |
| 17   | T200_O & CBOW | Failed Cluster | 1. Standard & Poor  
2. WSJ  
3. Philippines  
4. U.S. Treasury  
5. FED  
6. Social Security  
7. Standard & Poor |
| 18   | T200_O & CBOW | Federal regulatory agencies and divisions | 1. Wall Street Journal  
2. Senate Banking Committee  
3. American Stock Exchange  
4. Delaware Chancery Court  
5. House Financial Services Committee  
6. Labor Department  
7. Public Company Accounting Oversight Board  
8. Treasury Department  
9. Office of Federal Housing Enterprise Oversight  
10. Federal Communications Commission |