Cluster Labeling by Word Embeddings and WordNet's Hypernymy

Hanieh Poostchi
University of Technology Sydney
Capital Markets CRC
hpoostchi@cmcrc.com

Massimo Piccardi
University of Technology Sydney
massimo.piccardi@uts.edu.au

Abstract

Cluster labeling is the assignment of representative labels to clusters of documents or words. Once assigned, the labels can play an important role in applications such as navigation, search and document classification. However, finding appropriately descriptive labels is still a challenging task. In this paper, we propose various approaches for assigning labels to word clusters by leveraging word embeddings and the synonymy and hypernymy relations in the WordNet lexical ontology. Experiments carried out using the WebAP document dataset have shown that one of the approaches stand out in the comparison and is capable of selecting labels that are reasonably aligned with those chosen by a pool of four human annotators.

1 Introduction and Related Work

Document collections are often organized into clusters of either documents or words to facilitate applications such as navigation, search and classification. The organization can prove more useful if its clusters are characterized by *sets of representative labels*. The task of assigning a set of labels to each individual cluster in a document organization is known as cluster labeling (Wang et al., 2014) and it can provide a useful description of the collection in addition to fundamental support for navigation and search.

In Manning et al. (2008), cluster labeling approaches have been subdivided into i) differential cluster labeling and ii) cluster-internal labeling. The former selects cluster labels by comparing the distribution of terms in one cluster with those of the other clusters while the latter selects labels that are solely based on each cluster individually. Cluster-internal labeling approaches include computing the clusters' centroids and using them as labels, or using lists of terms with highest frequencies in the clusters. However, all these approaches can only select cluster labels from the terms and phrases that explicitly appear in the documents, possibly failing to provide an appropriate level of abstraction or description (Lau et al., 2011). As an example, a word cluster containing words *dog* and *wolf* should not be labeled with either word, but as *canids*. For this reason, in this paper we explore several approaches for labeling word clusters obtained from a document collection by leveraging the synonymy and hypernymy relations in the WordNet taxonomy (Miller, 1995), together with word embeddings (Mikolov et al., 2013; Pennington et al., 2014).

A hypernymy relation represents an asymmetric relation between a class and each of its instances. A hypernym (e.g., *vertebrate*) has a broader context than its hyponyms (*bird, fishes, reptiles* etc). Conversely, the contextual properties of the hyponyms are usually a subset of those of their hypernym(s). Hypernymy has been used extensively in natural language processing, including in recent works such as Yu et al. (2015) and HyperVec (Nguyen et al., 2017) that have proposed learning word embeddings that reflect the hypernymy relation. Based on this, we have decided to make use of available hypernym-hyponym data to propose an approach for labeling clusters of keywords by a representative selection of their hypernyms.

In the proposed approach, we first extract a set of keywords from the original document collection. We then apply a step of hierarchical clustering on the keywords to partition them into a hierarchy of clusters. To this aim, we represent each keyword as a real-valued vector using pre-trained word embeddings (Pennington et al., 2014) and repeatedly apply a standard clustering algorithm.

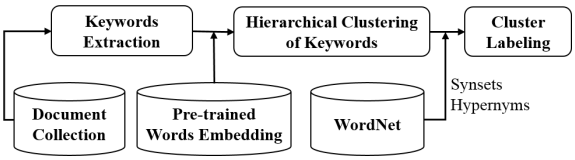

Figure 1: The proposed cluster labeling pipeline.

For labeling the clusters, we first look up all the synonyms of the keywords and, in turn, their hypernyms in the WordNet hierarchy. We then encode the hypernyms as word embeddings and use various approaches to select them based on their distance from the clusters' centers. The experimental results over a benchmark document collection have shown that such a distance-based selection is reasonably aligned with the hypernyms selected by four, independent human annotators. As a side result, we show that the employed word embeddings spontaneously contain the hypernymy relation, offering a plausible justification for the effectiveness of the proposed method.

2 The Proposed Pipeline

The proposed pipeline of processing steps is shown in Figure 1. First, keywords are extracted from each document in turn and accumulated in an overall set of unique keywords. After mapping such keywords to pre-trained word embeddings, hierarchical clustering is applied in a top-down manner. The leaves of the constructed tree are considered as the clusters to be labeled. Finally, each cluster is labeled automatically by leveraging a combination of WordNet's hypernyms and synsets and word embeddings. The following subsections present each step in greater detail.

2.1 Keyword Extraction

For the keyword extraction, we have used the rapid automatic keyword extraction (RAKE) of Rose et al. (2010). This method extracts keywords (i.e., single words or very short word sequences) from a given document collection and its main steps can be summarized as:

1. Split a document into sentences using a predefined set of sentence delimiters.
2. Split sentences into sequences of contiguous words at phrase delimiters to build the candidate set.
3. Collect the set of unique words (W) that appear in the candidate set.
4. Compute the word co-occurrence matrix $X_{|W|\times|W|}$ for W.
5. Calculate word score $score(w) = deg(w)/freq(w)$, where $deg(w) = \sum_{i\in\{1,\ldots,|W|\}} X[w,i]$ and $freq(w) = \sum_{i\in\{1,\ldots,|W|\}} (X[w,i] \neq 0)$.
6. Score each candidate keyword as the sum of its member word scores.
7. Select the top T scoring candidates as keywords for the document.

Alternatively, RAKE can use other combinations of $deg(w)$ and $freq(w)$ as the word scoring function. The keywords extracted from all the documents are accumulated into a set, C, ensuring uniqueness.

2.2 Hierarchical Clustering of Keywords

A top-down approach is used to hierarchically cluster the keywords in C. First, each component word of each keyword is mapped onto a numerical vector using pre-trained GloVe50d[1] word embeddings (Pennington et al., 2014); missing words are mapped to zero vectors. Then, each keyword k is represented with the average vector $\overrightarrow{k}$ of its component words. Then, we start from set C as the root of the tree and follow a branch-and-bound approach, where each tree node is clustered into c clusters using the k-means algorithm (Hartigan and Wong, 1979). A node is marked as a leaf if it contains less than n keywords or it belongs to level d, the tree's depth limit. The leaf nodes are the clusters to be named with a set of verbal terms.

2.3 Cluster Labeling

As discussed in Section 1, we aim to label each cluster with descriptive terms. The labels should be more general than the cluster's members to abstract the nature of the cluster. To this end, we leverage the hypernym-hyponym correspondences in the lexical ontology. First, for each cluster, we create a large set, L, of candidate labels by including the hypernyms[2] of the component words, expanded by their synonyms, of all the keywords. The synonyms are retrieved from the WordNet's sets of synonyms, called *synsets*. Then, we apply the four following approaches to select l labels from set L:

[1] http://nlp.stanford.edu/data/wordvecs/glove.6B.zip

[2] Nouns only (not verbs).

- *FreqKey*: Choose the l most frequent hypernyms of the l most frequent keywords.
- *CentKey*: Choose the l most central hypernyms of the l most central keywords.
- *FreqHyp*: Choose the l most frequent hypernyms.
- *CentHyp*: Choose the l most central hypernyms.

Approaches *FreqKey* and *FreqHyp* are based on frequencies in the collection. For performance evaluation, we sort their selected labels in descending frequency order. In *CentKey* and *CentHyp*, the centrality is computed with respect to the cluster's center in the embedding space as the average vector of all its keywords $\overrightarrow{K} = \frac{1}{|K|}\sum_{k\in K}\overrightarrow{k}$. The distance between hypernym h and the cluster's center is $d(\overrightarrow{h}, \overrightarrow{K}) = ||\overrightarrow{h} - \overrightarrow{K}||$, where $\overrightarrow{h}$ is the average vector of the hypernym's component words. The labels selected by these two approaches are sorted in ascending distance order.

3 Experiments and Results

For the experiments, we have used the WebAP dataset[3] (Keikha et al., 2014) as the document collection. This dataset contains $6,399$ documents of diverse nature with a total of $1,959,777$ sentences. For the RAKE software[4], the hyper-parameters are the minimum number of characters of each keyword, the maximum number of words of each keyword, and the minimum number of times each keyword appears in the text, and they have been left to their default values of 5, 3, and 4, respectively. Likewise, parameter T has been set to its default value of one third of the words in the co-occurrence matrix. For the hierarchical clustering, we have used $c = 8$, $n = 100$ and $d = 4$ based on our own subjective assessment.

3.1 Human Annotation and Evaluation

For the evaluation, eight clusters (one from each sub-tree) were chosen to be labeled manually by four, independent human annotators. For this purpose, for each cluster, we provided the list of its keywords, K, and the candidate labels, L, to the annotators, and asked them to select the best $l = 10$ terms from L to describe the cluster. Initially, we had considered asking the annotators to also select representative labels from K, but a preliminary analysis showed that they were unsuitable to describe the cluster as a whole (Table 1 shows an example). Although the annotators were asked to provide their selection as a ranked list, we did not make use of their ranking order in the evaluation.

[3] https://ciir.cs.umass.edu/downloads/WebAP/

[4] https://github.com/aneesha/RAKE

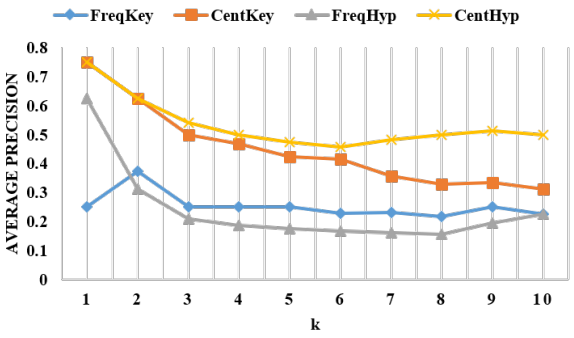

Figure 2: Precision at k ($P@k$) for $k = 1, \ldots, 10$ averaged over the eight chosen clusters for the compared approaches.

To evaluate the prediction accuracy, for each cluster we have considered the union of the lists provided by the human annotators as the ground truth (since $|L|$ was typically in the order of $150-200$, the intersection of the lists was often empty or minimal). As performance figure, we have decided to report the well-known precision at k ($P@k$) for values of k between one and ten. We have not used the recall since the ground truth had size 40 in most cases while the prediction's size was kept to $l = 10$ in all cases, resulting in a highest possible recall of 0.25. Figure 2 compares the average $P@k$ for $k = 1, \ldots, 10$ for the four proposed approaches. The two approaches based on minimum distance to the cluster center (*CentKey* and *CentHyp*) have outperformed the other two approaches based on frequencies (*FreqKey* and *FreqHyp*) for all values of k. This shows that the word embedding space is in good correspondence with the human judgement. Moreover, approach *CentHyp* has outperformed all other approaches for all values of k, showing that the hypernyms' centrality in the cluster is the key property for their effective selection.

3.2 Visualization of Keywords and Hypernyms

Hypernyms are more general terms than the corresponding keywords, thus we expect them to be in larger mutual distance in the word embedding

Keywords	website www, clearinghouse, nih website, bulletin, websites, hotline, kbr publications, pfm file, syst publication, gov web site, dhhs publication, beta site, lexis nexis document, private http, national register bulletin, daily routines, data custodian, information, serc newsletter, certified mail, informational guide, dot complaint database, coverage edit followup, local update, mass mailing, ahrq web site, homepage, journal messenger, npl site, pdf private, htm centers, org website, web site address, telephone directory, service records, page layout program, service invocation, newsletter, card reader, advisory workgroup, library boards, full text online, usg publication, webpage, bulletin boards, fbis online, teleconference info, journal url, insert libraries, headquarters files, volunteer website http, bibliographic records, vch publishers, ptd web site, tsbp newsletter, electronic bulletin boards, email addresses, ecommerce, traveler, api service, intranet, website http, newsletter nps files, mail advertisement transmitted, subscribe, nna program, npci website, bulletin board, fais information, archiving, page attachment, nondriver id, mail etiquette, ip address, national directory, web page, pdq editorial boards, aml sites, dhs site, ptd website, directory ers web site, forums, digest, beta site management, directories, ccir papers, ieee press, fips publication, org web site, clearinghouse database, monterey database, hotlines, dslip description info, danish desk files, sos web site, bna program, newsletters, inspections portal page, letterhead, app ropri, image file directory, website, electronic mail notes, web site http, customized template page, mail addresses, health http, internet questionnaire assistance, electronic bulletin board, eos directly addresses, templates directory, beta site testers, informational, dataplot auxiliary directory, coverage edit, quarterly newsletter, distributed, reader, records service, web pages.
Annotator 1	**electronic_communication**, computer_network, **web_page**, **web_site**, mail, **text_file**, **computer_file**, protocol, software, electronic_equipment
Annotator 2	computer_network, telecommunication, computer, mail, **web_page**, information, news, press, code, software
Annotator 3	news, **informing**, medium, **web_page**, **computer_file**, written_record, document, press, article, essay
Annotator 4	communication, **electronic_communication**, **informing**, press, medium, document, electronic_equipment, computer_network, transmission, record
CentHyp	**electronic_communication**, information_measure, **text_file**, **web_page**, **informing**, print_media, **web_site**, **computer_file**, commercial_enterprise, reference_book

Table 1: An example cluster. The hypernyms selected by *CentHyp* and by at least one annotator are shown in boldface.

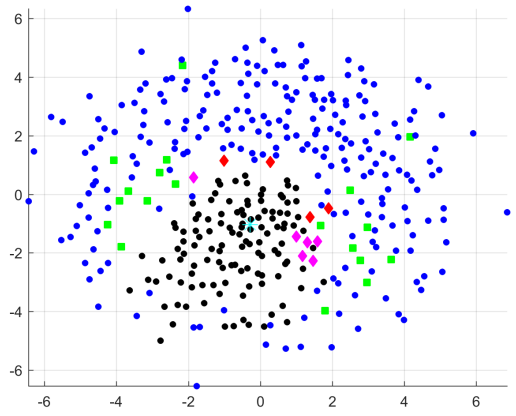

Figure 3: Two-dimensional visualization of an example cluster (this figure should be viewed in color). The black and blue dots are the cluster's keywords and the keywords' hypernyms, respectively. The green dots are the hypernyms selected by the human annotators, the red dots are the hypernyms selected by *CentHyp*, and their intersection is recolored in magenta. The cluster's center is the turquoise star.

space. To explore their distribution, we have used two-dimensional multidimensional scaling (MDS) visualizations (Borg and Groenen, 2005) of selected clusters. For each cluster, the keywords set K, the hypernyms set L, and the cluster's center have all been aggregated as a single set before applying MDS. An examples is shown in Figure 3. As can be seen, the hypernyms (blue dots) nicely distribute as a circular crown, external and concentric to the keywords (black dots), showing that the hypernymy relation corresponds empirically to a radial expansion away from the cluster's center. This likely stems from the embedding space's requirement to simultaneously enforce meaningful distances between the different keywords, the keywords and the corresponding hypernyms, and between the hypernyms themselves. The hypernyms selected by the annotators (green and magenta dots) are among the closest to the cluster's center, and thus those selected by *CentHyp* (red and magenta dots) have the best correspondence (magenta dots alone) among the explored approaches.

3.3 A Detailed Example

As a detailed example, Table 1 lists all the keywords of a sample cluster and the hypernyms selected by the four human annotators and *CentHyp*. Some of the hypernyms selected by more than one annotator (e.g., "electronic communication", "web page" and "computer file") have also been successfully identified by *CentHyp*. On the other hand, *CentHyp* has selected at least two terms ("commercial enterprise" and "reference book') that are unrelated to the cluster. Qualitatively, we deem the automated annotation as noticeably inferior to the human annotations, yet usable wherever manual annotation is infeasible or impractical.

4 Conclusion

This paper has explored various approaches for labeling keyword clusters based on the hypernyms from the WordNet lexical ontology. The proposed approaches map both the keywords and their hypernyms to a word embedding space and leverage the notion of centrality in the cluster. Experiments carried out using the WebAP dataset have shown that one of the approaches (*CentHyp*) has outperformed all the others in terms of precision at k for all values of k, and it has provided labels which are reasonably aligned with those of a pool of annotators. We plan to test the usefulness of the labels for tasks of search expansion in the near future.

Acknowledgments

This research has been funded by the Capital Markets Cooperative Research Centre in Australia and supported by Semantic Sciences Pty Ltd.

References

I. Borg and P. J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer New York.

J A. Hartigan and M. A. Wong. 1979. A K-Means Clustering Algorithm. *JSTOR: Applied Statistics* 28(1):100–108.

M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson. 2014. Retrieving Passages and Finding Answers. In *Proceedings of the 2014 Australasian Document Computing Symposium (ADCS)*. pages 81–84.

J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*. volume 1, pages 1536–1545.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, volume 2, pages 3111–3119.

G. A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

K. A. Nguyen, M. Köeper, S. Schulte im Walde, and N. T. Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of the 2017 Empirical Methods in Natural Language Processing (EMNLP)*. pages 233–243.

J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing (EMNLP)*. volume 14, pages 1532–1543.

S. Rose, D. Engel, N. Cramer, and W. Cowley. 2010. Automatic Keyword Extraction from Individual Documents. In *Text Mining. Applications and Theory*, Wiley-Blackwell, chapter 1, pages 1–20.

J. Wang, C. Kang, Y. Chang, and J. Han. 2014. A Hierarchical Dirichlet Model for Taxonomy Expansion for Search Engines. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*. pages 961–970.

Z. Yu, H. Wang, X. Lin, and M. Wang. 2015. Learning Term Embeddings for Hypernymy Identification. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. pages 1390–1397.