Clustering Moderate-size Collections of Short Texts

Lili Kotlerman · Ido Dagan · Oren Kurland

(work in progress, will be submitted to IRJ)

In real-life industrial settings there is often a need to perform clustering of short-text collections of moderate size, originating from a specific target domain. Typical examples are clustering for exploration of user interactions in a company call center or for analysis of statements about a definite company or product in social media.

Such data is challenging for traditional clustering methods. Due to the low number of texts (typically a couple of hundreds), there is not enough data to obtain meaningful information about semantic/topical relatedness of terms, as it is done by co-clustering and topic-modelling approaches. From the other side, short texts do not allow to effectively measure similarity between the texts directly, based on their term overlap, as done by classical clustering techniques, such as Complete Link, K-means etc.

To overcome this problem external knowledge is usually involved:

- One way of utilizing external knowledge is applicable for classical clustering algorithms, which are the common practice in many application settings (e.g. [1], [2], [3]). According to this approach the texts are augmented with semantically related terms and then clustering algorithm is applied over the obtained longer texts (see [4] for an overview). Semantically related terms typically used are WordNet synonyms and, sometimes, hypernyms (e.g. [5], [6], [7]). Various distributional similarity methods can also be used to obtain information on semantic relatedness of terms from large corpora, as well as a number of additional resources.
- Another way is to train a topic model, such as LDA [8], over a large external corpus with further application of this model for clustering of the original text collection.

Our research was motivated by joint work with industrial partners. Thus, we collected real-life industrial data from different domains and source channels, and annotated the data in collaboration with domain experts. As a result we present 4 new datasets: (1) excerpts from e-mail feedbacks to a railway company, (2) tweets with criticism towards a bank, (3) tweets mentioning iPhone4, (4) excerpts from speech transcripts of customer interactions with a call center of an airline.

We analyse the datasets with clustering in mind and, based on this analysis, suggest a number of techniques to enhance clustering in our setting. We suggest utilizing external knowledge in an alternative way: given a text collection, automatically cluster the terms in the collection using external resources of semantic relatedness between terms, and further represent the texts as vectors of term clusters (bag-of-clusters) rather than bag-of-words vectors. This representation resembles that of co-clustering and topic-modelling methods, which detect groups of semantically related terms to enhance text clustering. Yet, being explicit, it allows control over feature weighting, provides a natural way of incorporating external lexical resources of any type and enables application of classical clustering methods, such as Complete Link, over a more compact and neat representation. We suggest a feature weighting scheme geared for our type of data and suggest retaining only most prominent features in each vector.

In addition, we analyse the state-of-the-art clustering algorithms and hypothesize that K-medoids approach ([9]) is most suitable for our type of data. We show that K-medoids algorithm over bag-of-clusters vector representation with our suggested weighting scheme considerably and significantly outperforms the following baselines:

- Complete Link and K-medoids over original texts and over texts augmented with semantically related terms (see below)
- Co-clustering of [10] over original texts and over augmented texts
- LDA model trained over original texts and over augmented texts, and applied to original texts and to augmented texts
- LDA model trained over UKWaC corpus and applied to original texts and to augmented texts

To augment the original texts and to obtain term clusters, for each term t_i in bag-of-words vectors of the original texts we extracted its WordNet synonyms, derivationally related terms, hypo- and hypernyms, holonyms and meronyms, as well as top-20 terms similar to t_i according to a distributional similarity resource derived from the UKWaC corpus .

Finally, we intend to publicly release our clustering tool, with access to all baselines and external knowledge resources, as well as the datasets created for this research.

References:

[1] Nomoto, T., Matsumoto, Y.: A new approach to unsupervised text summarization. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 26–34. ACM (2001).

[2] Naughton, M., Kushmerick, N., Carthy, J.: Clustering sentences for discovering events in news articles. In: Advances in Information Retrieval, pp. 535–538. Springer (2006)

[3] Ye, H., Young, S.: A clustering approach to semantic decoding. In: Ninth International Conference on Spoken Language Processing (2006)

[4] Hu, Jian, et al.:Enhancing text clustering by leveraging Wikipedia semantics. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. (2008).

[5] Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pp. 541–544. IEEE (2003)

[6] Shehata, Shady. "A wordnet-based semantic model for enhancing text clustering. In: Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on. IEEE (2009).

[7] Jing, Liping, Michael K. Ng, and Joshua Z. Huang.: Knowledge-based vector space model for text clustering. In: Knowledge and information systems 25.1. (2010).

[8] Blei, D., Ng, Y., Jordan, M.: Latent Dirichlet Allocation. In: Journal of machine learning research. (2003).

[9] Kaufmann, L., Rousseeuw, P.J.: Clustering by means of medioids. In: Statistical data analysis based on the L1 norm. (1987).

[10] Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 89–98. ACM (2003)