Sentence Clustering via Projection over Term Clusters

Lili Kotlerman, Ido Dagan
Bar-Ilan University
Israel
Lili.Kotlerman@biu.ac.il
dagan@cs.biu.ac.il

Maya Gorodetsky, Ezra Daya
NICE Systems Ltd.
Israel
Maya.Gorodetsky@nice.com
Ezra.Daya@nice.com

Abstract

This paper presents a novel sentence clustering scheme based on projecting sentences over term clusters. The scheme incorporates external knowledge to overcome lexical variability and small corpus size, and outperforms common sentence clustering methods on two real-life industrial datasets.

1 Introduction

Clustering is a popular technique for unsupervised text analysis, often used in industrial settings to explore the content of large amounts of sentences. Yet, as may be seen from the results of our research, widespread clustering techniques, which cluster sentences directly, result in rather moderate performance when applied to short sentences, which are common in informal media.

In this paper we present and evaluate a novel sentence clustering scheme based on projecting sentences over term clusters. Section 2 briefly overviews common sentence clustering approaches. Our suggested clustering scheme is presented in Section 3. Section 4 describes an implementation of the scheme for a particular industrial task, followed by evaluation results in Section 5. Section 6 lists directions for future research.

2 Background

Sentence clustering aims at grouping sentences with similar meanings into clusters. Commonly, vector similarity measures, such as cosine, are used to define the level of similarity over bag-of-words encoding of the sentences. Then, standard clustering algorithms can be applied to group sentences into clusters (see Steinbach et al. (2000) for an overview).

The most common practice is representing the sentences as vectors in term space and applying the K-means clustering algorithm (Shen et al. (2011); Pasquier (2010); Wang et al. (2009); Nomoto and Matsumoto (2001); Boros et al. (2001)). An alternative approach involves partitioning a sentence connectivity graph by means of a graph clustering algorithm (Erkan and Radev (2004); Zha (2002)).

The main challenge for any sentence clustering approach is language variability, where the same meaning can be phrased in various ways. The shorter the sentences are, the less effective becomes exact matching of their terms. Compare the following newspaper sentence "*The bank is phasing out the EZ Checking package, with no monthly fee charged for balances over $1,500, and is instead offering customers its Basic Banking account, which carries a fee*" with two tweets regarding the same event: "*Whats wrong.. charging $$ for checking a/c*" and *"Now they want a monthly fee!*". Though each of the tweets can be found similar to the long sentence by exact term matching, they do not share any single term. Yet, knowing that the words *fee* and *charge* are semantically related would allow discovering the similarity between the two tweets.

External resources can be utilized to provide such kind of knowledge, by which sentence representation can be enriched. Traditionally, WordNet (Fellbaum, 1998) has been used for this purpose (Shehata (2009); Chen et al. (2003); Hotho et al. (2003); Hatzivassiloglou et al. (2001)). Yet, other resources

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 38–43,
Montréal, Canada, June 7-8, 2012.

of semantically-related terms can be beneficial, such as WordNet::Similarity (Pedersen et al., 2004), statistical resources like that of Lin (1998) or DIRECT (Kotlerman et al., 2010), thesauri, Wikipedia (Hu et al., 2009), ontologies (Suchanek et al., 2007) etc.

3 Sentence Clustering via Term Clusters

This section presents a generic sentence clustering scheme, which involves two consecutive steps: (1) generating relevant term clusters based on lexical semantic relatedness and (2) projecting the sentence set over these term clusters. Below we describe each of the two steps.

3.1 Step 1: Obtaining Term Clusters

In order to obtain term clusters, a term connectivity graph is constructed for the given sentence set and is clustered as follows:

1. Create initially an undirected graph with sentence-set terms as nodes and use lexical resources to extract semantically-related terms for each node.
2. Augment the graph nodes with the extracted terms and connect semantically-related nodes with edges. Then, partition the graph into term clusters through a graph clustering algorithm.

Extracting and filtering related terms. In Section 2 we listed a number of lexical resources providing pairs of semantically-related terms. Within the suggested scheme, any combination of resources may be utilized.

Often resources contain terms, which are semantically-related only in certain contexts. E.g., the words *visa* and *passport* are semantically-related when talking about tourism, but cannot be considered related in the banking domain, where *visa* usually occurs in its *credit card* sense. In order to discard irrelevant terms, filtering procedures can be employed. E.g., a simple filtering applicable in most cases of sentence clustering in a specific domain would discard candidate related terms, which do not occur sufficiently frequently in a target-domain corpus. In the example above, this procedure would allow avoiding the insertion of *passport* as related to *visa*, when considering the banking domain.

Clustering the graph nodes. Once the term graph is constructed, a graph clustering algorithm is applied resulting in a partition of the graph nodes (terms) into clusters. The choice of a particular algorithm is a parameter of the scheme. Many clustering algorithms consider the graph's edge weights. To address this trait, different edge weights can be assigned, reflecting the level of confidence that the two terms are indeed validly related and the reliability of the resource, which suggested the corresponding edge (e.g. WordNet synonyms are commonly considered more reliable than statistical thesauri).

3.2 Step 2: Projecting Sentences to Term Clusters

To obtain sentence clusters, the given sentence set has to be projected in some manner over the term clusters obtained in Step 1. Our projection procedure resembles unsupervised text categorization (Gliozzo et al., 2005), with categories represented by term clusters that are not predefined but rather emerge from the analyzed data:

1. Represent term clusters and sentences as vectors in term space and calculate the similarity of each sentence with each of the term clusters.
2. Assign each sentence to the best-scoring term cluster. (We focus on hard clustering, but the procedure can be adapted for soft clustering).

Various metrics for feature weighting and vector comparison may be chosen. The top terms of term-cluster vectors can be regarded as labels for the corresponding sentence clusters.

Thus each sentence cluster corresponds to a single coherent cluster of related terms. This is contrasted with common clustering methods, where if sentence A shares a term with B, and B shares another term with C, then A and C might appear in the same cluster even if they have no related terms in common. This behavior turns out harmful for short sentences, where each incidental term is influential. Our scheme ensures that each cluster contains only sentences related to the underlying term cluster, resulting in more coherent clusters.

4 Application: Clustering Customer Interactions

In industry there's a prominent need to obtain business insights from customer interactions in a contact center or social media. Though the number of key

sentences to analyze is often relatively small, such as a couple hundred, manually analyzing just a handful of clusters is much preferable. This section describes our implementation of the scheme described in Section 3 for the task of clustering customer interactions, as well as the data used for evaluation. Results and analysis are presented in Section 5.

4.1 Data

We apply our clustering approach over two real-life datasets. The first one consists of 155 sentences containing reasons of account cancelation, retrieved from automatic transcripts of contact center interactions of an Internet Service Provider (ISP). The second one contains 194 sentences crawled from Twitter, expressing reasons for customer dissatisfaction with a certain banking company. The sentences in both datasets were gathered automatically by a rule-based extraction algorithm. Each dataset is accompanied by a small corpus of call transcripts or tweets from the corresponding domain.[1]

The goal of clustering these sentences is to identify the prominent reasons of cancelation and dissatisfaction. To obtain the gold-standard (GS) annotation, sentences were manually grouped to clusters according to the reasons stated in them.

Table 1 presents examples of sentences from the ISP dataset. The sentences are short, with only one or two words expressing the actual reason stated in them. We see that exact term matching is not sufficient to group the related sentences. Moreover, traditional clustering algorithms are likely to mix related and unrelated sentences, due to matching non-essential terms (e.g. *husband* or *summer*). We note that such short and noisy sentences are common in informal media, which became a most important channel of information in industry.

4.2 Implementation of the Clustering Scheme

Our proposed sentence clustering scheme presented in Section 3 includes a number of choices. Below we describe the choices we made in our current implementation.

Input sentences were tokenized, lemmatized and cleaned from stopwords in order to extract content-word terms. Candidate semantically-related terms were extracted for each of the terms, using WordNet synonyms and derivations, as well as DIRECT[2], a directional statistical resource learnt from a news corpus. Candidate terms that did not appear in the accompanying domain corpus were filtered out as described in Section 3.1.

he hasn't been using it all summer long
it's been sitting idle for about it almost a year
I'm getting married my husband has a computer
yeah I bought a new laptop this summer so
when I said faces my husband got laid off from work
well I'm them going through financial difficulties

Table 1: Example sentences expressing 3 reasons for cancelation: the customer (1) does not use the service, (2) acquired a computer, (3) cannot afford the service.

Edges in the term graph were weighted with the number of resources supporting the corresponding edge. To cluster the graph we used the Chinese Whispers clustering tool[3] (Biemann, 2006), whose algorithm does not require to pre-set the desired number of clusters and is reported to outperform other algorithms for several NLP tasks.

To generate the projection, sentences were represented as vectors of terms weighted by their frequency in each sentence. Terms of the term-cluster vectors were weighted by the number of sentences in which they occur. Similarity scores were calculated using the cosine measure. Clusters were labeled with the top terms appearing both in the underlying term cluster and in the cluster's sentences.

5 Results and Analysis

In this section we present the results of evaluating our projection approach, compared to the common K-means clustering method[4] applied to:

(A) Standard bag-of-words representation of sentences;

[1] The bank dataset with the output of the tested methods will be made publicly available.

[2] Available for download at `www.cs.biu.ac.il/~nlp/downloads/DIRECT.html`. For each term we extract from the resource the top-5 related terms.

[3] Available at `http://wortschatz.informatik.uni-leipzig.de/~cbiemann/software/CW.html`

[4] We use the Weka (Hall et al., 2009) implementation. Due to space limitations and for more meaningful comparison we report here one value of K, which is equal to the number of clusters returned by projection (60 for the ISP and 65 for the bank dataset). For K = 20, 40 and 70 the performance was similar.

(B) Bag-of-words representation, where sentence's words are augmented with semantically-related terms (following the common scheme of prior work, see Section 2). We use the same set of related terms as is used by our method.

(C) Representation of sentences in term-cluster space, using the term clusters generated by our method as vector features. A feature is activated in a sentence vector if it contains a term from the corresponding term cluster.

Table 2 shows the results in terms of Purity, Recall (R), Precision (P) and F1 (see "Evaluation of clustering", Manning et al. (2008)). Projection significantly[5] outperforms all baselines for both datasets.

Dataset	Algorithm	Purity	R	P	F1
ISP	**Projection**	**.74**	**.40**	**.68**	**.50**
	K-means A	.65	.18	.22	.20
	K-means B	.65	.13	.24	.17
	K-means C	.65	.18	.26	.22
Bank	**Projection**	**.79**	**.26**	**.53**	**.35**
	K-means A	.61	.14	.14	.14
	K-means B	.64	.13	.19	.16
	K-means C	.67	.17	.21	.19

Table 2: Evaluation results.

For completeness we experimented with applying Chinese Whispers clustering to sentence connectivity graphs, but the results were inferior to K-means.

Table 3 presents sample sentences from clusters produced by projection and K-means for illustration. Our initial analysis showed that our approach indeed produces more homogenous clusters than the baseline methods, as conjectured in Section 3.2. We consider it advantageous, since it's easier for a human to merge clusters than to reveal sub-clusters. E.g., a GS cluster of 20 sentences referring to fees and charges is covered by three projection clusters labeled *fee*, *charge* and *interest rate*, with 9, 8 and 2 sentences correspondingly. On the other hand, K-means C method places 11 out of the 20 sentences in a messy cluster of 57 sentences (see Table 3), scattering the remaining 9 sentences over 7 other clusters.

In our current implementation *fee*, *charge* and *interest rate* were not detected by the lexical resources we used as semantically similar and thus were not grouped in one term cluster. However, adding more resources may introduce additional noise. Such dependency on coverage and accuracy of resources is apparently a limitation of our approach. Yet, as our experiments indicate, using only two generic resources already yielded valuable results.

a. Projection

credit card, card, mastercard, visa (38 sentences)
XXX has the worst credit cards ever
XXX MasterCard is the worst credit card I've ever had
ntuc do not accept XXX visa now I have to redraw $150...
XXX card declined again , $40 dinner in SF...

b. K-means C

fee, charge (57 sentences)
XXX playing games wit my interest
arguing w incompetent pol at XXX damansara perdana
XXX's upper management are a bunch of rude pricks
XXX are ninjas at catching fraudulent charges.

Table 3: Excerpt from resulting clusterings for the bank dataset. Bank name is substituted with XXX. Cluster labels are given in italics. Two most frequent terms are assigned as cluster labels for K-means C.

6 Conclusions and Future Work

We presented a novel sentence clustering scheme and evaluated its implementation, showing significantly superior performance over common sentence clustering techniques. We plan to further explore the suggested scheme by utilizing additional lexical resources and clustering algorithms. We also plan to compare our approach with co-clustering methods used in document clustering (Xu et al. (2003), Dhillon (2001), Slonim and Tishby (2000)).

Acknowledgments

This work was partially supported by the MAGNETON grant no. 43834 of the Israel Ministry of Industry, Trade and Labor, the Israel Ministry of Science and Technology, the Israel Science Foundation grant 1112/08, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886 and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

[5] p=0.001 according to McNemar test (Dietterich, 1998).

References

Chris Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City, USA.

Endre Boros, Paul B. Kantor, and David J. Neu. 2001. A clustering based approach to creating multi-document summaries.

Hsin-Hsi Chen, June-Jei Kuo, and Tsei-Chun Su. 2003. Clustering and visualization in a multi-lingual multi-document summarization system. In *Proceedings of the 25th European conference on IR research*, ECIR'03, pages 266–280, Berlin, Heidelberg. Springer-Verlag.

Inderjit S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 269–274, New York, NY, USA. ACM.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

C. Fellbaum. 1998. *WordNet – An Electronic Lexical Database*. MIT Press.

Alfio Massimiliano Gliozzo, Carlo Strapparava, and Ido Dagan. 2005. Investigating unsupervised learning for text categorization bootstrapping. In *HLT/EMNLP*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min yen Kan, and Kathleen R. McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *In Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49.

A. Hotho, S. Staab, and G. Stumme. 2003. Wordnet improves text document clustering. In Ying Ding, Keith van Rijsbergen, Iadh Ounis, and Joemon Jose, editors, *Proceedings of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR 2003), August 1, 2003, Toronto Canada*. Published Online at `http://de.scientificcommons.org/608322`.

Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. 2009. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 389–396, New York, NY, USA. ACM.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *JNLE*, 16:359–389.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, Juli.

Tadashi Nomoto and Yuji Matsumoto. 2001. A new approach to unsupervised text summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 26–34, New York, NY, USA. ACM.

Claude Pasquier. 2010. Task 5: Single document keyphrase extraction using sentence clustering and latent dirichlet allocation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 154–157, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shady Shehata. 2009. A wordnet-based semantic model for enhancing text clustering. *Data Mining Workshops, International Conference on*, 0:477–482.

Chao Shen, Tao Li, and Chris H. Q. Ding. 2011. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases. In *AAAI*.

Noam Slonim and Naftali Tishby. 2000. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 208–215, New York, NY, USA. ACM.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. *KDD Workshop on Text Mining*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A large ontology from wikipedia and wordnet.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 297–300, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA. ACM.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR*, pages 113–120.