

A Compact Representation for Cross-Domain Short Text Clustering

Alba Núñez-Reyes^{1,2}, Esaú Villatoro-Tello^{2(✉)}, Gabriela Ramírez-de-la-Rosa²,
and Christian Sánchez-Sánchez²

¹ Maestría en Diseño, Información y Comunicación (MADIC),
División de Ciencias de la Comunicación y Diseño,
Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa,
Mexico City, Mexico
`ar.nunezreyes@gmail.com`

² Language and Reasoning Research Group, Information Technologies Department,
Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa,
Mexico City, Mexico
`{evillatoro,gramirez,csanchez}@correo.cua.uam.mx`

Abstract. Nowadays, Twitter depicts a rich source of on-line reviews, ratings, recommendations, and other forms of opinion expressions. This scenario has created the compelling demand to develop innovative mechanisms to store, search, organize and analyze all this data automatically. Unfortunately, it is seldom available to have enough labeled data in Twitter, because of the cost of the process or due to the impossibility to obtain them, given the rapid growing and change of this kind of media. To avoid such limitations, *unsupervised categorization strategies* are employed. In this paper we face the problem of cross-domain short text clustering through a compact representation that allows us to avoid the problems that arise with the high dimensionality and sparseness of vocabulary. Our experiments, conducted on a cross-domain scenario using very short texts, indicate that the proposed representation allows to generate high quality groups, according to the value of Silhouette coefficient obtained.

Keywords: Short text clustering · Unsupervised categorization · Cross-domain clustering · Compact text representation · Silhouette coefficient

1 Introduction

Nowadays, microblogging media has become a universal tool for sharing and obtaining information from millions of people. For instance, Twitter represents a rich source of trends, reviews, ratings, recommendations, and many other forms of online opinion expressions. This scenario has created a compelling demand for innovative mechanisms to automatically store, search, organize and analyze all this data. An example of such mechanisms are *topic detection* methods, which implies the automatic discovery of thematic-related groups into free-text documents, namely *tweets*.

Traditional approaches for topic detection, such as supervised classification strategies, assume that training and test documents are drawn from the same distribution. However, in many cases this scenario is unreal, especially in datasets extracted from Twitter. Thus, the process of using a statistical model trained in one (source) domain, for categorizing information contained in a different (target) domain, requires bridging the gap between the two domains to facilitate the knowledge transfer¹. Accordingly, *cross-domain clustering* represents an unsupervised process where instances from the target domain are categorized using information obtained from the formed clusters within the source domain [2–5].

In this paper we describe an analysis on the pertinence of using a very compact representation for texts in a cross-domain clustering scenario. Thus, we employed a representation based on the transition point (TP) technique, which dictates that terms surrounding the TP value are closely related to the conceptual content of a document; the TP is located at the middle point between the terms with most and less frequency. Therefore, we hypothesise that the terms whose frequencies are closer to the TP can be used as features to represent a document to transfer the knowledge from one domain to another. Although this technique has been extensively validated for solving Natural Language Processing (NLP) tasks such as text clustering [6] and text classification [7]; a common characteristic among these research works is that experiments were performed using formal documents, *i.e.*, news reports, books, articles, abstracts, etc. Currently, to the best of our knowledge, there is no formal study regarding the evaluation on the pertinence of a representation based on TP for a scenario with the following characteristics: (i) documents are very short texts, specifically Twitter posts, (ii) documents belong to a very dynamic and constantly growing environment, and, (iii) a proposal where a representation that fits to the distribution of unknown domains is used, *i.e.*, cross-domain topic detection. According to these observations, our research looks for answers to the following questions: *how useful is a compact document representation based on the TP for topics detection in Twitter?* and, *to what extent can be improved the clustering quality of tweets, by means of the acquired knowledge from other domains?*

The rest of the paper is organized as follows. Section 2 describes recent approaches in domain adaptation for text clustering in short texts. Section 3 explains our experimental methodology, and Sect. 4 presents the obtained results. Finally, Sect. 5 depicts our conclusions and some ideas for future work.

2 Related Work

Several approaches have been proposed for the automatic topic detection problem. The most popular proposal considers the application of *supervised strategies* to infer a classification function from a training hand-labelled set (*i.e.*, *source domain*), and then, the learnt function is used for predicting the category of new unlabeled documents (*i.e.*, *target domain*) [8]. As expected, the more the number of labeled documents, the better the results in the target distribution,

¹ A research problem known as *domain adaptation* [1].

and consequently the categorization performance. Unfortunately, as we previously mentioned, for many real life tasks, such as Twitter data analysis, having enough labeled data is uncommon, and often, too expensive or some times even impossible to obtain given the rapid growing and change of such media. As a consequence of this limitations, *unsupervised strategies* are employed.

In the work described in [9], authors proposed a method that allows the identification of news topics in Twitter. In general the system works as follows: (i) tweets are obtained by means of querying specific keywords, then, (ii) a supervised classifier distinguishes between news-tweets and junk-tweets, (iii) news-tweets are clustered into topics using a leader-follower algorithm, which requires a set of initial handpicked tweets to lead the algorithm; and finally, (iv) a geo-tag is assigned to tweets for visualizing the results to the user. A key difference between this work and ours is that we do not require training documents, neither a set of manually picked tweets for building the groups.

In Atefeh et al. [10], authors provide an extensive review of several methods for topic detection in Twitter media. Generally speaking, the topic detection methods are categorized based on the following characteristics: (i) whether the topics are known a priori or not, (ii) if the employed approach is supervised, unsupervised or hybrid, and, (iii) if a retrospective analysis is needed or if the recently acquired data need to be analysed to discover new topics. Most of the works described in [10] employ a traditional Bag-of-Terms representation, in which the terms vary from the full vocabulary to specific named entities and even to some other probabilistic representations (*e.g.*, language models). Regarding similarity metrics, the most common are: Euclidean distance, Pearson’s correlation coefficient, cosine similarity and others like Hellinger distance and clustering index. It is worth mentioning that most of these works assume an in-domain scenario, *i.e.*, data were drawn from the same distribution. Contrastingly, our experiments are conducted on a cross-domain configuration, and our proposal of a compact representation depicts a balance between the high dimensionality and the sparseness of vocabulary.

Concerning to the cross-domain clustering, there has been a lesser amount of research, and even less for cross-domain text clustering [3,5]. In [3] the authors propose learning from a subspace of common knowledge while at the same time clustering both the source domain and target domain, such that the data distribution of different domains are similar when projected in the subspace. In [5] authors propose an iterative clustering method that allows transfer the knowledge of distributions of clusters from the source domain to the target domain, in addition, its approach uses a density based method for clustering the target domain. In these works, the employed datasets for experimentation are formal documents (*e.g.*, news reports) and the number of groups in both domains have to be known a priori. In our work, we use more challenging data, namely tweets, and, as in real life, the number of existing clusters in both domains are unknown.

In summary, most of the previous work apply either a supervised or an unsupervised categorization strategy that assumes a similar distribution of the data, hence most of the reported experiments are performed under an in-domain

scenario using large-formal documents. On the contrary, our proposal faces the problem of cross-domain clustering by means of using a compact representation that efficiently deals with the vocabulary high dimensionality and sparseness of short text documents.

3 Experimental Methodology

In this section we describe our proposed methodology for tackling the problem of cross-domain short text clustering. Firstly, we explain the pre-processing operations applied to our data, then we briefly report how the compact representation was computed; and finally, we describe the employed clustering method as well as the evaluation metric considered for reporting our experimental results.

3.1 Pre-processing Stage

As pre-processing steps we applied the following operations to each tweet contained in the employed dataset: (1) tweets are transformed to lowercase; (2) user mentions (*i.e.*, @USER), hashtags (*i.e.*, #HASHTAG), and outgoing links are deleted from the tweet; (3) all punctuation mark as well as emoticons are deleted; (4) a stemming process is applied using the Porter algorithm [11]; and, (5) all stopwords are removed.

3.2 Document Representation

As mentioned before, we propose to use a compact representation for tackling the problem of cross-domain short texts clustering. Accordingly, we employed the Transition Point (TP) technique as a key aspect in our proposed solution. It is known that the TP depicts a frequency value that divides the vocabulary of a document into those terms with high and low frequency values respectively. This technique is based on the Zipf law of word occurrences [12, 13]. From these studies it is possible to hypothesise that those terms whose frequency value is closer to the tp factor, are strongly related to the main conceptual content of a document, and therefore are good elements for its representation. A typical formula for computing the tp value is shown in expression 1.

$$tp_T = \frac{\sqrt{8 * I_1 + 1} - 1}{2} \quad (1)$$

where I_1 represents the number of hapax² contained in the collection vocabulary. Once the tp is defined, a document is then represented by means of a BoW-like³ technique considering only those words near to the tp frequency value. In our experiments, we employ a fixed percentage of terms near to the tp frequency value, particularly 40%.

² A word that occurs just once within a text.

³ Bag-of-Words representation.

As described in [6], the *tp* technique has been employed as a term selection strategy in different NLP related tasks. However, as far as we know, it has never been applied for representing Twitter posts, for the task of cross-domain clustering.

3.3 Clustering Method

As the main clustering method we employed the well known *k*-means algorithm. The *k*-means algorithm is an iterative approach that executes two basic steps: first, assigns documents to existing centroids and second, reduces the divergence by the re-estimation of the centroids based on the current assignment of documents.

Although the *k*-means method has the disadvantage of requiring a manually-set parameter by the user, *i.e.*, the number of groups (*k*), we decided to use this algorithm since it has demonstrated to be an effective method for many clustering related tasks [14]. The used similarity metric was the Euclidean distance⁴.

3.4 Evaluation

Given that the number of existing topics within the data are unknown, supervised evaluation metrics such as *accuracy*, *F-score*, *precision* or *recall* can not be applied. Nevertheless, it is necessary to measure the quality of formed groups, aiming at determining the clustering tendencies of the data, *i.e.*, distinguishing whether (or not) a random structure actually exists in the data.

Accordingly, as evaluation measure we employ the popular metric of silhouette coefficient [15], which combines the concepts of cohesion and separation for each point of a clusters. As it is known, the cluster’s cohesion measures how closely the objects in a cluster are related among them. Whilst cluster’s separation measures how well a cluster is distinguished from other clusters. The silhouette coefficient value varies from -1 to 1 . A negative silhouette value depicts an undesirable result, meaning a bad clustering; on the contrary, positive values define a better quality in the clustering result, which is a preferable performance.

4 Empirical Evaluation and Results

In order to answer our stated research questions, we employed a dataset of tweets in English and Spanish gathered for the Online Reputation Laboratory (RepLap⁵). For the experiments we used the subset belonging to the topic detection task in its 2013 edition [16]. This dataset contains tweets from four different domains: Automotive (Au), Banking (Ba), Music (Mu) and University (Un). In Table 1 some statistics of this dataset are shown.

From Table 1 can be observed some interesting facts regarding the nature of the dataset. For instance, the comparison of the data across domains shows that

⁴ We employed the *k*-means as implemented in <http://scikit-learn.org/>.

⁵ <http://nlp.uned.es/replab2013/>.

Table 1. Statistics for the dataset used for the experimental evaluation. *Au*, *Ba*, *Mu* and *Un* refer to Automotive, Banking, Music and University, respectively.

| | Spanish | | | | English | | | |
|-----------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| | <i>Au</i> | <i>Ba</i> | <i>Mu</i> | <i>Un</i> | <i>Au</i> | <i>Ba</i> | <i>Mu</i> | <i>Un</i> |
| Avg. words per tweet | 9.9 | 10.1 | 9.9 | 10.0 | 10.6 | 11.3 | 10.0 | 10.9 |
| Avg. vocabulary per tweet | 9.5 | 9.8 | 9.5 | 9.7 | 10.0 | 10.8 | 9.5 | 10.4 |
| Avg. tweet length (chars.) | 63.8 | 69.5 | 62.2 | 68.7 | 60.6 | 66.4 | 56.6 | 65.9 |
| Avg. words length per tweet | 6.4 | 6.9 | 6.3 | 6.9 | 5.7 | 5.9 | 5.6 | 6.1 |
| Total number of tweets | 5735 | 6552 | 8288 | 1056 | 33453 | 14378 | 33016 | 17863 |

tweets from the Banking domain have more words than the rest of the domains (in both languages); however the size of the vocabulary per tweet, in average, is similar across the four domains. Another interesting statistic is that tweets in Spanish have less number of words than tweets in English, but the number of English tweets is four times the sample in Spanish. Additionally, notice that the smaller domain (according to the number of tweets) in Spanish is University; while the smaller domain in English is Banking followed by University.

For the empirical evaluation two main set of experiments were designed. The first set evaluates the pertinence of the proposed representation, *i.e.* the TP approach, in a in-domain clustering scenario. The second set of experiments evaluates if the representation learned from a source domain, is able to produce quality clusters in a target domain, *i.e.*, cross-domain clustering. Following subsections describe the setup configuration as well as the obtained results.

4.1 In-Domain Clustering Experiments

The main goal of this set of experiments is to find out if the compact representation, based on the TP, is useful for topic detection in Twitter. As mentioned in Sect. 3.3, as clustering algorithm we employed the k -means method, using six different values of $k = \{5, 10, 15, 20, 25, 30\}$.

We established the maximum number of groups $k = 30$ because we considered this number small enough for a posterior analysis of the data by an online reputation manager; otherwise, the analysis could become very tedious. It is important to mention that RepLab organizers provided information regarding the number of *named entities* contained in each domain, for instance, some of the entities contained in the Automotive domain are: Nissan, Honda, BMW, etc. The quantity of entities in each domain is: 10 for Music, 11 for Banking, 19 for University and 21 for Automotive. Although this information was known a priori, it was not used in our experiments, but provides a referential parameter regarding the number of clusters that may be useful in a real case scenario.

As we are proposing a compact representation, we selected as a baseline a representation based on the *bag-of-words* (BOW) technique. As known, this type of representation considers the entire vocabulary for representing each document.

Results for *in-domain* experiments, both for Spanish and English languages, can be seen in Tables 2 and 3 respectively. Obtained results are reported in terms of the silhouette coefficient (SC). Notice that we compare the quality of the formed clusters for the two types of representation, *BOW* and *TP*, for each domain, as well as for the 6 different values of k . According to the SC values, the proposed compact representation outperforms the baseline representation. Interestingly, we can observe a difference between SC values for Spanish and English languages, mainly for University and Banking domains, where the quality of formed groups is better in Spanish than in English. This difference may have been caused by the size of the datasets between the two languages. In consequence, the number of elements contained into the clusters formed for the English data is bigger and thus the clusters may contain tweets with a higher diversity of topics.

4.2 Cross-Domain Clustering Experiments

Previous experiments demonstrate that the employed compact representation improves the performance of the clustering in very short texts. Taking this into

Table 2. Silhouette coefficient values for in-domain clustering for tweets in Spanish language. In the third row, the vector’s dimension for each type of representation.

| k | Automotive | | Banking | | Music | | University | |
|-----|------------|--------------|------------|--------------|------------|--------------|------------|--------------|
| | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> |
| | 13944 | 86 | 14284 | 79 | 14089 | 114 | 4419 | 22 |
| 5 | 0.008 | 0.109 | 0.031 | 0.197 | 0.029 | 0.143 | 0.028 | 0.251 |
| 10 | 0.012 | 0.129 | 0.035 | 0.205 | 0.037 | 0.143 | 0.026 | 0.371 |
| 15 | −0.007 | 0.137 | 0.018 | 0.200 | 0.04 | 0.178 | 0.013 | 0.407 |
| 20 | −0.016 | 0.142 | 0.013 | 0.217 | 0.038 | 0.179 | 0.020 | 0.452 |
| 25 | 0.007 | 0.152 | 0.021 | 0.224 | 0.016 | 0.179 | 0.020 | 0.483 |
| 30 | −0.219 | 0.163 | −0.004 | 0.233 | 0.024 | 0.178 | −0.002 | 0.503 |

Table 3. Silhouette coefficient values for in-domain clustering for tweets in English language. In the third row, the vector’s dimension for each type of representation.

| k | Automotive | | Banking | | Music | | University | |
|-----|--------------|-----------|------------|--------------|------------|--------------|------------|--------------|
| | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> | <i>BOW</i> | <i>TP</i> |
| | 30358 | 527 | 17988 | 310 | 22356 | 502 | 22593 | 327 |
| 5 | 0.120 | 0.061 | 0.01 | 0.085 | 0.023 | 0.126 | 0.010 | 0.083 |
| 10 | 0.130 | 0.088 | 0.021 | 0.089 | 0.028 | 0.120 | 0.021 | 0.095 |
| 15 | 0.102 | 0.094 | 0.003 | 0.079 | 0.030 | 0.060 | 0.005 | 0.067 |
| 20 | 0.017 | 0.064 | 0.019 | 0.091 | 0.030 | 0.072 | 0.019 | 0.078 |
| 25 | −0.116 | 0.002 | 0.016 | 0.090 | 0.040 | 0.118 | 0.016 | 0.059 |
| 30 | 0.113 | 0.009 | 0.003 | 0.090 | 0.024 | 0.088 | −0.003 | 0.056 |

account, for the next set of experiments we used the *TP* strategy for representing documents. The goal of the *cross-domain* clustering scenario is to determinate the robustness of the acquired knowledge from a source domain into a target domain. Results for the cross-domain experiments are shown in Tables 4 and 5 for Spanish and English languages respectively.

Table 4. Silhouette coefficient values for cross-domain clustering for tweets in Spanish language. The target domain is given in the first row and the source domain is indicated by *Au*, *Ba*, *Mu* and *Un* to refer to Automotive, Banking, Music and University, respectively.

| k | Automotive | | | Banking | | | Music | | | University | | |
|-----|------------|-----------|-------------|-----------|-----------|-------------|-----------|-----------|-------------|------------|-------------|-----------|
| | <i>Ba</i> | <i>Mu</i> | <i>Un</i> | <i>Au</i> | <i>Mu</i> | <i>Un</i> | <i>Au</i> | <i>Ba</i> | <i>Un</i> | <i>Au</i> | <i>Ba</i> | <i>Mu</i> |
| 5 | 0.27 | 0.27 | 0.65 | 0.24 | 0.26 | 0.67 | 0.25 | 0.29 | 0.66 | 0.24 | 0.28 | 0.29 |
| 10 | 0.30 | 0.29 | 0.81 | 0.27 | 0.31 | 0.80 | 0.27 | 0.33 | 0.82 | 0.27 | 0.32 | 0.30 |
| 15 | 0.32 | 0.32 | 0.87 | 0.31 | 0.33 | 0.86 | 0.30 | 0.36 | 0.87 | 0.30 | 0.36 | 0.32 |
| 20 | 0.35 | 0.34 | 0.89 | 0.32 | 0.38 | 0.88 | 0.32 | 0.38 | 0.89 | 0.33 | 0.37 | 0.36 |
| 25 | 0.38 | 0.35 | 0.90 | 0.35 | 0.40 | 0.90 | 0.33 | 0.42 | 0.91 | 0.37 | 0.42 | 0.39 |
| 30 | 0.40 | 0.37 | 0.92 | 0.37 | 0.43 | 0.91 | 0.36 | 0.44 | 0.93 | 0.38 | 0.43 | 0.42 |

Table 5. Silhouette coefficient values for cross-domain clustering for tweets in English language. The target domain is given in the first row and the source domain is indicated by *Au*, *Ba*, *Mu* and *Un* to refer to Automotive, Banking, Music and University, respectively.

| k | Automotive | | | Banking | | | Music | | | University | | |
|-----|-------------|-------------|-----------|-----------|-------------|-----------|-----------|-------------|-------------|------------|-------------|-----------|
| | <i>Ba</i> | <i>Mu</i> | <i>Un</i> | <i>Au</i> | <i>Mu</i> | <i>Un</i> | <i>Au</i> | <i>Ba</i> | <i>Un</i> | <i>Au</i> | <i>Ba</i> | <i>Mu</i> |
| 5 | 0.13 | 0.14 | 0.12 | 0.10 | 0.14 | 0.10 | 0.10 | 0.13 | 0.13 | 0.11 | 0.15 | 0.13 |
| 10 | 0.11 | 0.12 | 0.13 | 0.10 | 0.15 | 0.10 | 0.11 | 0.11 | 0.12 | 0.11 | 0.12 | 0.11 |
| 15 | 0.11 | 0.13 | 0.10 | 0.09 | 0.15 | 0.11 | 0.09 | 0.11 | 0.13 | 0.10 | 0.13 | 0.09 |
| 20 | 0.13 | 0.13 | 0.10 | 0.11 | 0.13 | 0.11 | 0.12 | 0.13 | 0.12 | 0.10 | 0.13 | 0.13 |
| 25 | 0.13 | 0.12 | 0.11 | 0.10 | 0.14 | 0.12 | 0.10 | 0.13 | 0.12 | 0.11 | 0.14 | 0.12 |
| 30 | 0.14 | 0.13 | 0.11 | 0.10 | 0.13 | 0.12 | 0.09 | 0.14 | 0.14 | 0.10 | 0.14 | 0.13 |

Several observations can be made from these results; firstly, similarly to the in-domain experiments, the quality of the clusters are better for Spanish than for English. The advantage of representing a domain in terms of another one is significantly better for the Spanish language. For example, for the *Music* (*Mu*) domain, the best in-domain performance was of $SC = 0.179$ with $k = 20$ (see Table 2), whilst for the cross-domain scenario the obtained result is $SC = 0.930$

Table 6. Silhouette coefficient values for cross-domain clustering of two source domains for Spanish language. The target domain is given in the first row and the source domain is indicated by *Au*, *Ba*, *Mu* and *Un* to refer to Automotive, Banking, Music and University, respectively. The - symbol indicates that the target domain does not contain any terms from the source domain.

| k | Automotive | | | Banking | | | Music | | | University | | |
|-----|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | <i>BaMu</i> | <i>BaUn</i> | <i>MuUn</i> | <i>AuMu</i> | <i>AuUn</i> | <i>MuUn</i> | <i>AuBa</i> | <i>AuUn</i> | <i>BaUn</i> | <i>AuBa</i> | <i>AuMu</i> | <i>BaMu</i> |
| 5 | 0.24 | 0.25 | 0.26 | 0.26 | 0.21 | 0.25 | - | 0.22 | 0.30 | - | 0.22 | 0.23 |
| 10 | 0.26 | 0.28 | 0.29 | 0.26 | 0.24 | 0.30 | - | 0.26 | 0.30 | - | 0.24 | 0.30 |
| 15 | 0.28 | 0.31 | 0.31 | 0.25 | 0.28 | 0.34 | - | 0.28 | 0.35 | - | 0.27 | 0.26 |
| 20 | 0.28 | 0.34 | 0.33 | 0.28 | 0.29 | 0.37 | - | 0.29 | 0.36 | - | 0.28 | 0.31 |
| 25 | 0.31 | 0.37 | 0.34 | 0.29 | 0.31 | 0.39 | - | 0.31 | 0.40 | - | 0.31 | 0.34 |
| 30 | 0.31 | 0.38 | 0.36 | 0.32 | 0.32 | 0.41 | - | 0.33 | 0.41 | - | 0.31 | 0.36 |

with $k = 30$ (see Table 4). The latter means that formed groups have good cohesion and good separability when knowledge from the University domain is considered to categorize tweets about Music.

In general, the knowledge extracted from the University domain turns out to be the best source domain for the Spanish language, however it was not the same for the English tweets. In Table 5 can be noticed that the improvement of using knowledge from other domains is smaller. For instance, SC values for the in-domain clusters formed for the Banking domain is $SC = 0.091$ with $k = 20$ (see Table 3). However, when using Music as source domain, the quality of the formed clusters in the Banking data is improved by 6%, *i.e.*, $SC = 0.15$ with $k = 15$ (see Table 5). In addition, notice that for the Spanish experiments, the better results are more consistent with $k = 30$; contrastingly, for the English data there is not a clear tendency for the k value.

4.3 Are the More Domains the Better?

Our performed experiments in a *cross-domain* scenario indicate that the *TP* representation depicts an effective approach for detecting topics across different target domains. Hence, we design a third set of experiments aiming to explore to what extent the quality of the clustering can increase, when more than one domain are used as source data.

To perform this set of experiments, the compact vocabulary (based on the TP) was unified from two and from three source domains, then the combined vocabulary was used to represent one target domain. The obtained results for these experiments in Spanish tweets are shown in Tables 6 and 7. In general, it is not possible to further improve the quality of the formed groups (see Table 4), however, the obtained performance still is better than when the in-domain clustering is performed (see Table 2). We think that the produced detriment is due to the inclusion of noisy terms when more source domains are considered.

Table 7. Silhouette coefficient values for cross-domain clustering of three source domains for Spanish language. The target domain is given in the first row and the source domain is indicated by *Au*, *Ba*, *Mu* and *Un* to refer to Automotive, Banking, Music and University, respectively.

| k | Automotive | Banking | Music | University |
|-----|---------------|---------------|---------------|---------------|
| | <i>BaMuUn</i> | <i>AuMuUn</i> | <i>AuBaUn</i> | <i>AuBaMu</i> |
| 5 | 0.23 | 0.22 | 0.26 | 0.25 |
| 10 | 0.24 | 0.26 | 0.30 | 0.24 |
| 15 | 0.26 | 0.27 | 0.31 | 0.21 |
| 20 | 0.27 | 0.29 | 0.33 | 0.27 |
| 25 | 0.29 | 0.31 | 0.35 | 0.29 |
| 30 | 0.30 | 0.32 | 0.38 | 0.25 |

5 Conclusions

In this paper we tackled the problem of cross-domain short text clustering, particularly topic detection in Twitter posts. We carried out a study on the pertinence of using a compact document representation, specifically a term selection method based on the Transition Point technique, which establishes that terms surrounding the transition point are good terms for capturing the conceptual content of a document.

Two research questions were stated at the beginning of this paper. First, we were interested in validating the usefulness of the *TP* representation for topic detection in Twitter. The performed experiments showed that the proposed compact representation allows to produce high quality groups, particularly for tweets in Spanish language. Second, we were interested in evaluating if the proposed representation was suitable to identify topics under a cross-domain scenario, *i.e.*, if the *TP* facilitates the knowledge transfer between the source and target domains. Our experiments showed that the proposed methodology produces high quality groups under a *cross-domain* scenario, specially for tweets in Spanish. Finally, an additional experiment, showed that the combination of the knowledge extracted from two or three domains, is not useful for improving the clustering results in the target domain.

As future work, we want to explore the sensitivity of the proposed compact representation to the number of selected terms by the *TP* technique. Furthermore, we want to incorporate contextual information, namely, word n-grams. Our intuition is that if some contextual information is added, specially for English tweets, the quality of the formed clusters could be improved. Additionally, we intent to determine the pertinence of the proposed representation for solving non-thematic text classification tasks, such as author profiling problems (*e.g.*, age, gender, and personality recognition), where not enough/reliable labeled data are available.

Acknowledgments. This work was partially funded by CONACyT: through project grant 258588, the Thematic Networks program (Language Technologies Thematic Network projects 260178, 271622), and scholarship number 587804. We also thank to UAM Cuajimalpa and SNI-CONACyT for their support.

References

1. Li, Q.: Literature survey: domain adaptation algorithms for natural language processing. Department of Computer Science, The Graduate Center, The City University of New York, pp. 8–10 (2012)
2. Dai, W., Yang, Q., Xue, G.-R., Yu, Y.: Self-taught clustering. In: Proceedings of the 25th International Conference on Machine Learning, ICML 2008, pp. 200–207. ACM, New York (2008)
3. Gu, Q., Zhou, J.: Learning the shared subspace for multi-task clustering and transductive transfer classification. In: 2009 Ninth IEEE International Conference on Data Mining, pp. 159–168, December 2009
4. Bhattacharya, I., Godbole, S., Joshi, S., Verma, A.: Cross-guided clustering: transfer of relevant supervision across tasks. *ACM Trans. Knowl. Discov. Data* **6**, 9:1–9:28 (2012)
5. Samanta, S., Selvan, A.T., Das, S.: Cross-domain clustering performed by transfer of knowledge across domains. In: 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1–4, December 2013
6. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Gelbukh, A. (ed.) *CICLing 2006*. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg (2006). doi:[10.1007/11671299_55](https://doi.org/10.1007/11671299_55)
7. Moyotl-Hernández, E., Jiménez-Salazar, H.: Enhancement of DTP feature selection method for text categorization. In: Gelbukh, A. (ed.) *CICLing 2005*. LNCS, vol. 3406, pp. 719–722. Springer, Heidelberg (2005). doi:[10.1007/978-3-540-30586-6_80](https://doi.org/10.1007/978-3-540-30586-6_80)
8. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* **34**, 1–47 (2002)
9. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2009, pp. 42–51. ACM, New York (2009)
10. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. *Comput. Intell.* **31**, 132–164 (2015)
11. Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)
12. Zipf, G.: *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge (1949)
13. Booth, A.D.: A law of occurrences for words of low frequency. *Inf. Control* **10**(4), 386–393 (1967)
14. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Aggarwal, C., Zhai, C. (eds.) *Mining Text Data*, pp. 77–128. Springer US, Boston (2012)
15. Rousseeuw, P.J.: Silhouettes: graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
16. Amigó, E., et al.: Overview of replab 2013: evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) *CLEF 2013*. LNCS, vol. 8138, pp. 333–352. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-40802-1_31](https://doi.org/10.1007/978-3-642-40802-1_31)

Advances in Computational Intelligence
15th Mexican International Conference on Artificial
Intelligence, MICA 2016, Cancún, Mexico, October
23–28, 2016, Proceedings, Part I
Sidorov, G.; Herrera-Alcántara, O. (Eds.)
2017, XXVII, 552 p. 184 illus., Softcover
ISBN: 978-3-319-62433-4