A compact representation for cross-domain short text clustering

Presentation · October 2016

DOI: 10.13140/RG.2.2.33206.70726

CITATIONS
0

READS
91

4 authors, including:

Esaú Villatoro-Tello
Metropolitan Autonomous University
67 PUBLICATIONS **278** CITATIONS
SEE PROFILE

Gabriela Ramirez-de-la-Rosa
Metropolitan Autonomous University
39 PUBLICATIONS **89** CITATIONS
SEE PROFILE

A compact representation for cross-domain short text clustering

Alba Núñez-Reyes, Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa and Christian Sánchez-Sánchez

Language and Reasoning (LyR) Research Group, Information Technologies Dept.,UAM, México

MICAI, Cancun, Mexico. October 27th 2016

Outline

1 Introduction

2 Proposed method

3 Experimental setup

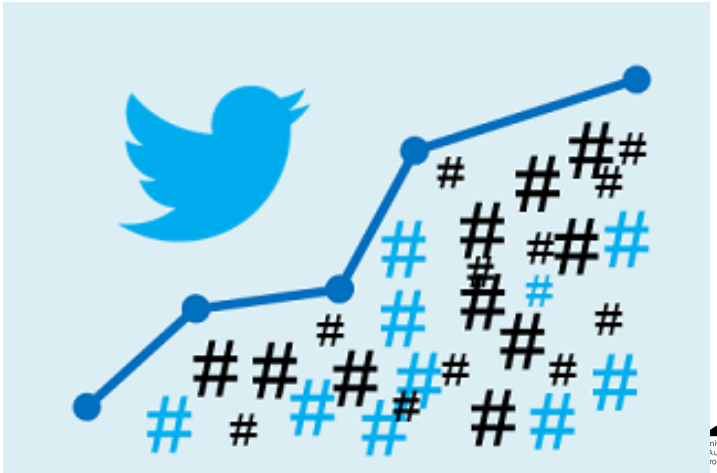
4 Results

5 Conclusions and Future work

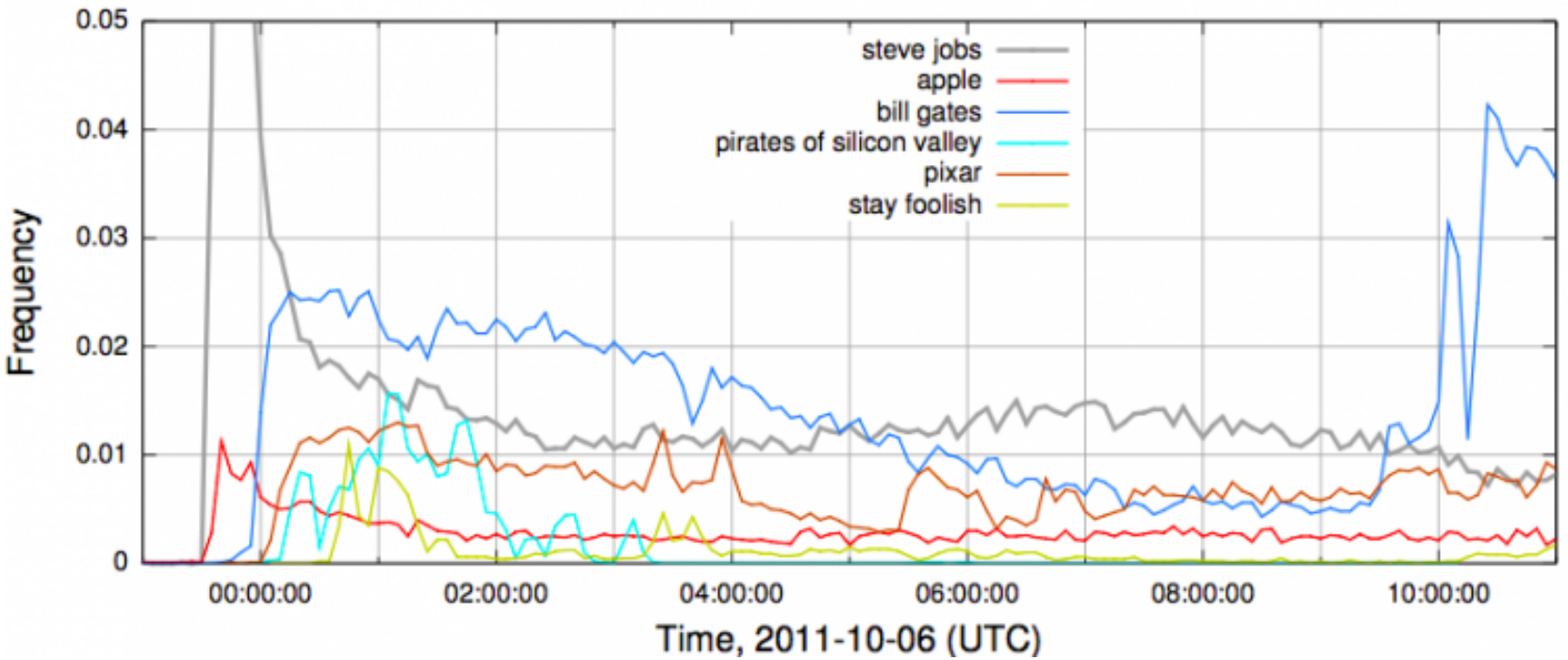
Introduction

Twitter in our everyday life :

- Twitter represents a rich source of on-line opinion expressions.
- Innovative mechanisms to automatically store, search, organize and analyze all this data.
- Traditional approaches for topic detection of same distributions.



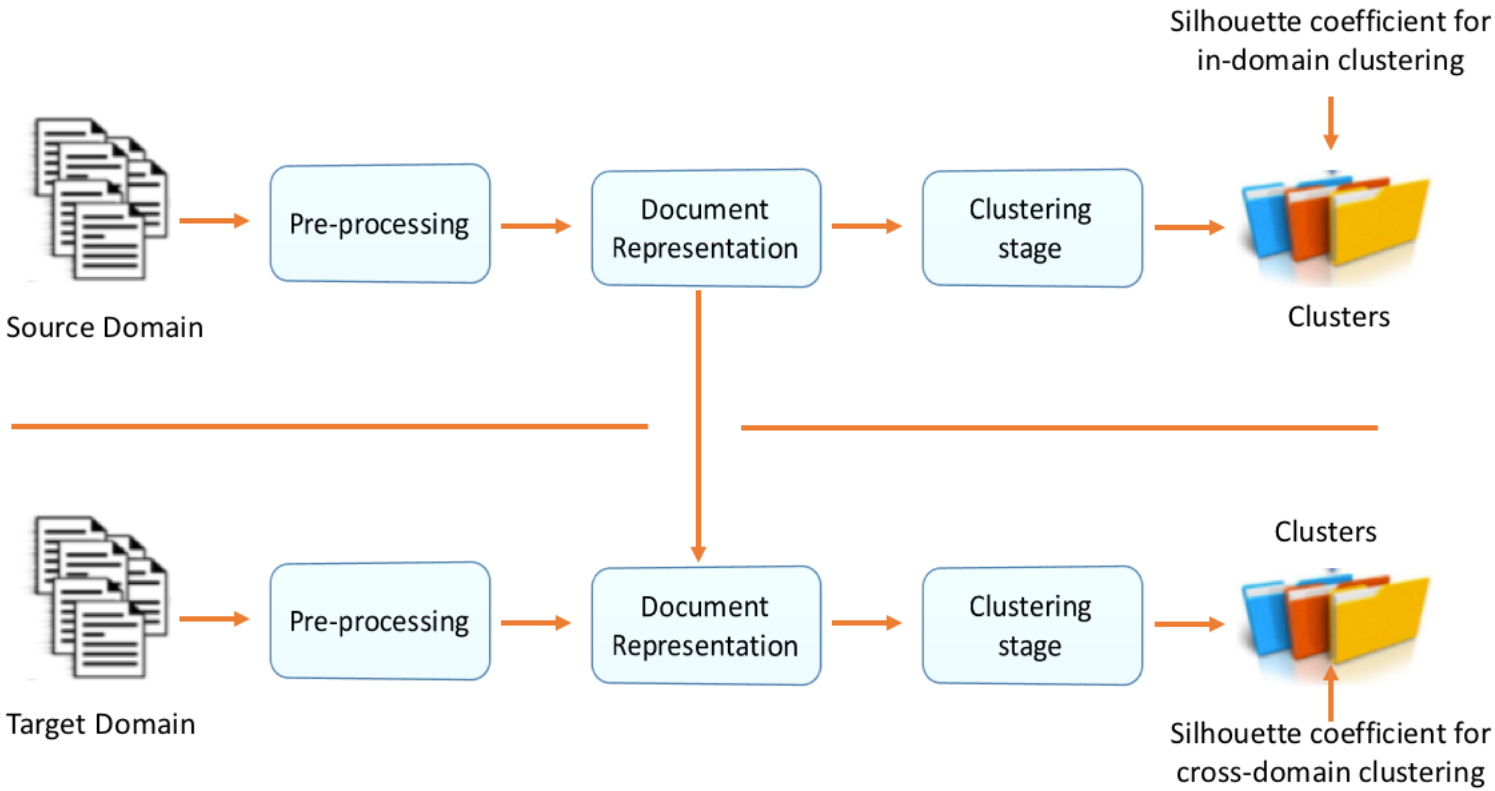
Supervised classification strategies, assume that training and test documents are drawn from the same distribution. However, in many cases this scenario is unreal, especially in data sets extracted from Twitter. Thus, the process of using a statistical model trained in one (source) domain, for categorizing information contained in a different (target) domain, requires bridging the gab between the two domains to facilitate the knowledge transfer.

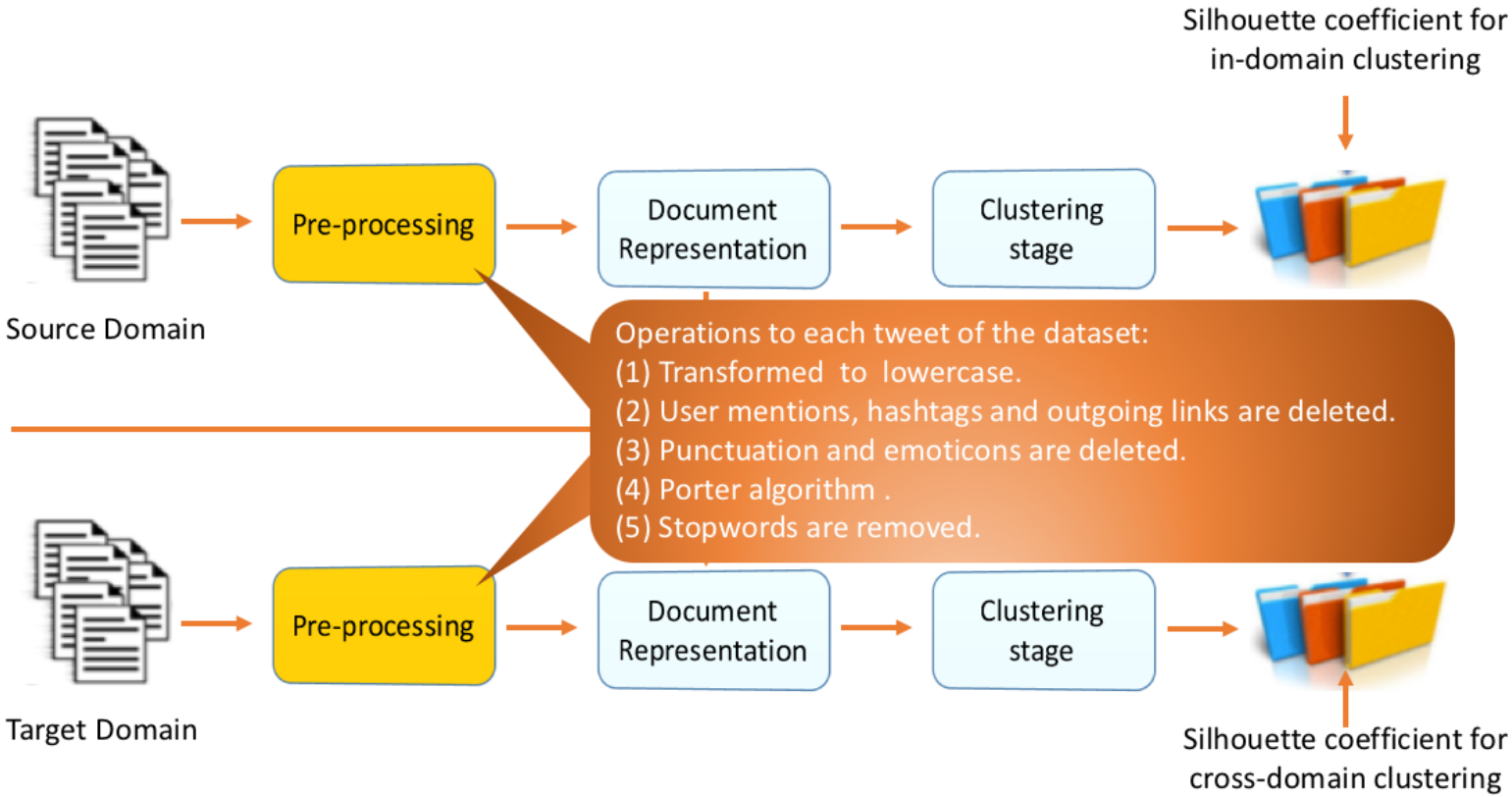


Research questions

- We hypothesise that a compact representation can be used as feature to transfer the knowledge from one domain to another.
- Thus, our main research questions are :
  1. **How useful is a compact document representation for topics detection in Twitter ?**
  2. **What extent can be improved the clustering quality of tweets, by means of the acquired knowledge from other domains ?**

General architecture



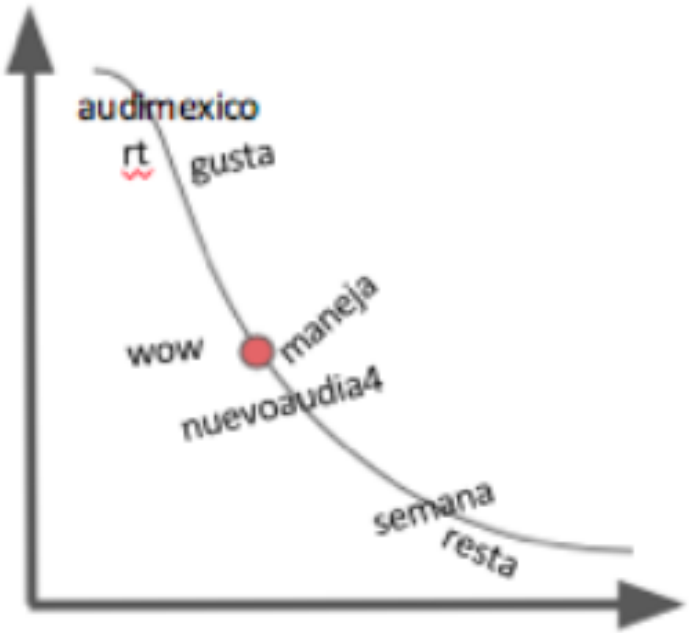

Silhouette coefficient for in-domain clustering
Source Domain
Pre-processing
Document Representation
Clustering stage
Operations to each tweet of the dataset:
(1) Transformed to lowercase.
(2) User mentions, hashtags and outgoing links are deleted.
(3) Punctuation and emoticons are deleted.
(4) Porter algorithm .
(5) Stopwords are removed.
Target Domain
Pre-processing
Document Representation
Clustering stage
Silhouette coefficient for cross-domain clustering

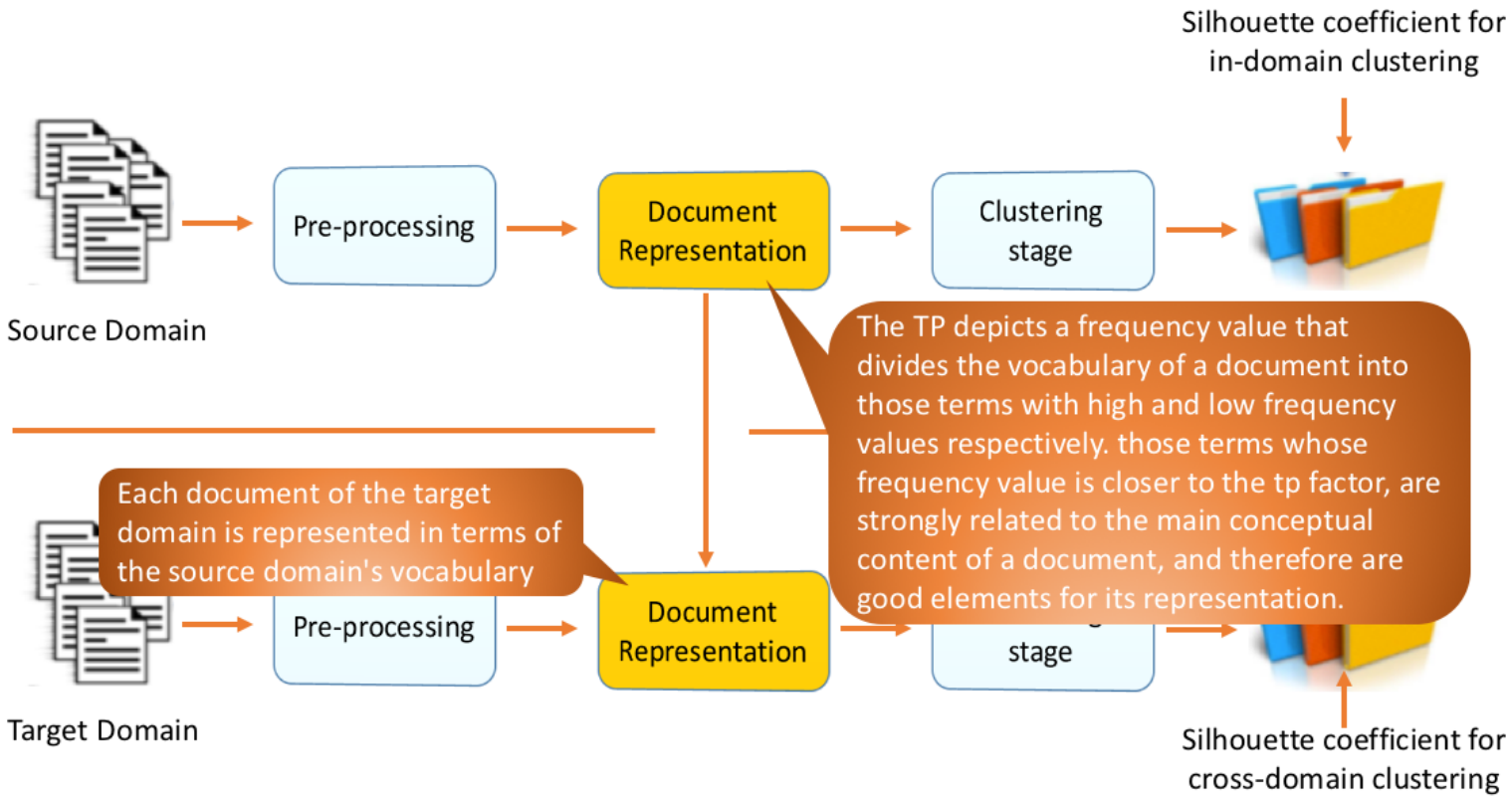
Baseline

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.

Transition Point (TP) Technique

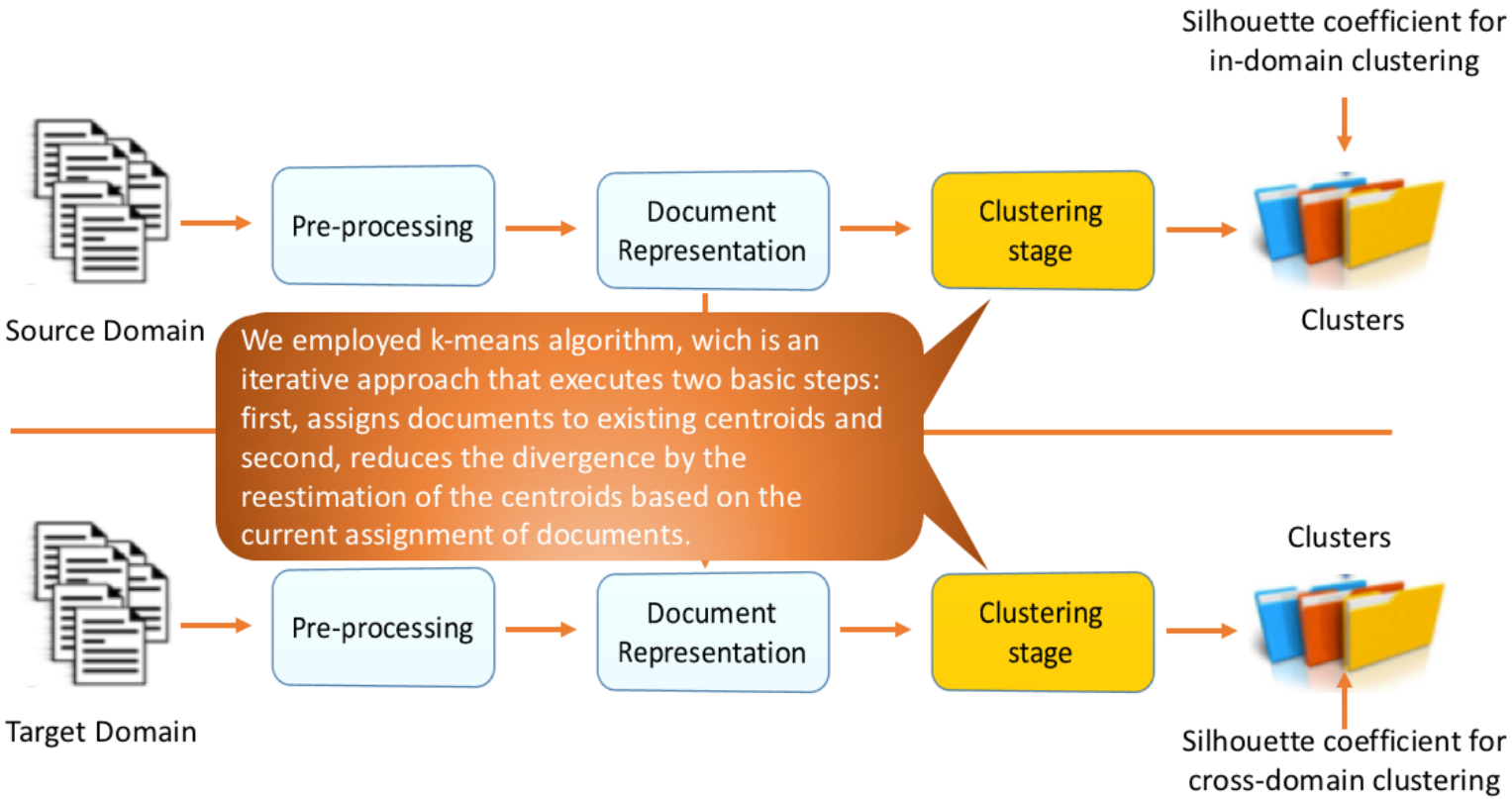
Which dictates that terms surrounding the TP value are closely related to the conceptual content of a document ; the TP is located at the middle point between the terms with most and less frequency.

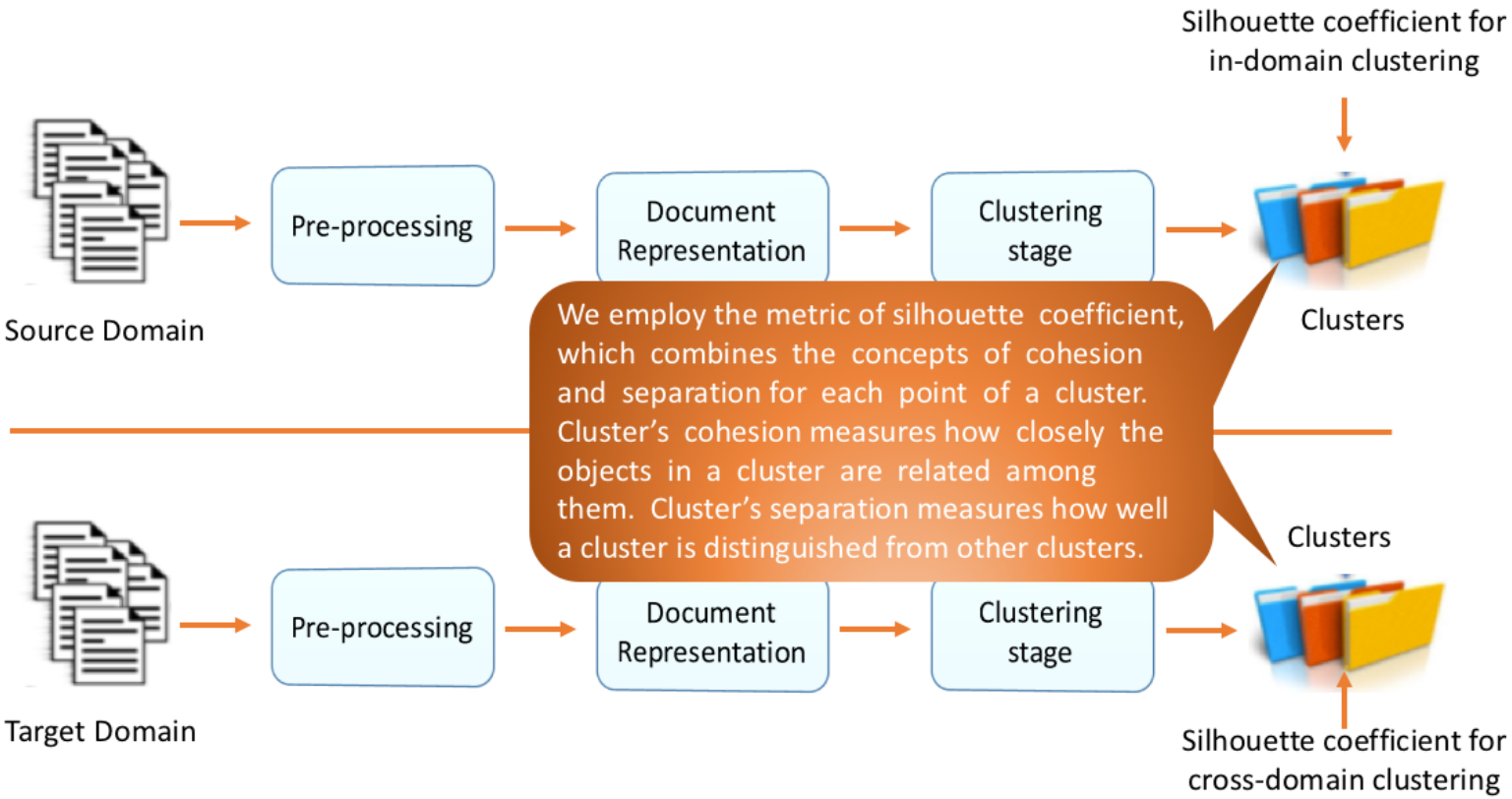



Silhouette coefficient for in-domain clustering
Pre-processing
Document Representation
Clustering stage
Source Domain
The TP depicts a frequency value that divides the vocabulary of a document into those terms with high and low frequency values respectively. those terms whose frequency value is closer to the tp factor, are strongly related to the main conceptual content of a document, and therefore are good elements for its representation.
Each document of the target domain is represented in terms of the source domain's vocabulary
Pre-processing
Document Representation
stage
Target Domain
Silhouette coefficient for cross-domain clustering

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.


Silhouette coefficient for in-domain clustering
Source Domain
Pre-processing
Document Representation
Clustering stage
Clusters
We employed k-means algorithm, wich is an iterative approach that executes two basic steps: first, assigns documents to existing centroids and second, reduces the divergence by the reestimation of the centroids based on the current assignment of documents.
Target Domain
Pre-processing
Document Representation
Clustering stage
Clusters
Silhouette coefficient for cross-domain clustering


Silhouette coefficient for in-domain clustering
Pre-processing
Document Representation
Clustering stage
Clusters
Source Domain
We employ the metric of silhouette coefficient, which combines the concepts of cohesion and separation for each point of a cluster. Cluster's cohesion measures how closely the objects in a cluster are related among them. Cluster's separation measures how well a cluster is distinguished from other clusters.
Clusters
Pre-processing
Document Representation
Clustering stage
Target Domain
Silhouette coefficient for cross-domain clustering

1. Replab 2013.
2. Documents are very short texts, Twitter posts.
3. Tweets in Spanish, English and another languages like Portuguese.

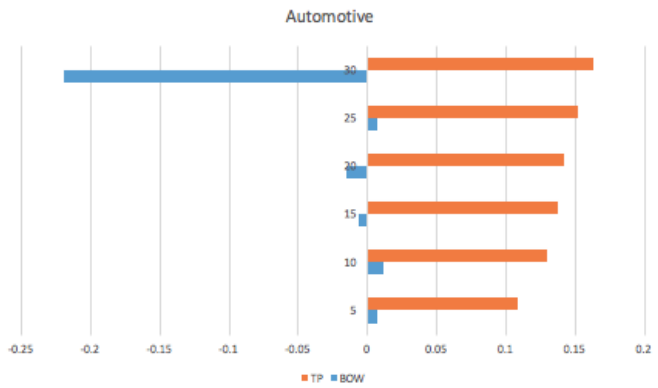
Corpus

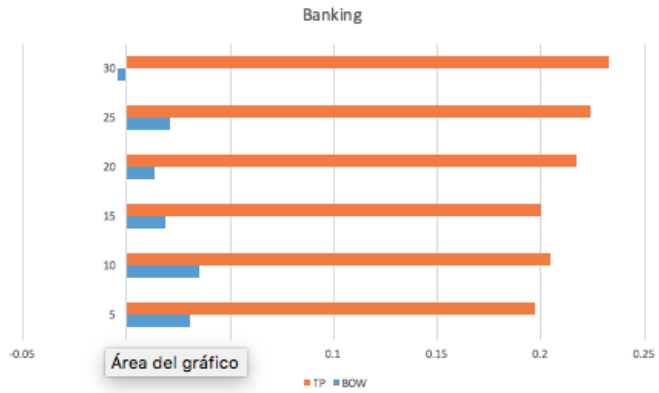
Stadistics regarding the corpora in English and Spanish

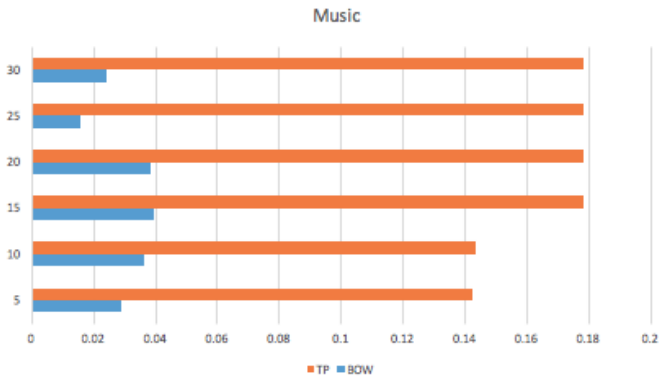
	Spanish				***English***			
	Au	*Ba*	*Mu*	*Un*	*Au*	*Ba*	*Mu*	*Un*
Avg. words per tweet	9.9	10.1	9.9	10.0	10.6	11.3	10.0	10.9
Avg. vocabulary per tweet	9.5	9.8	9.5	9.7	10.0	10.8	9.5	10.4
Avg. tweet length (chars.)	63.8	69.5	62.2	68.7	60.6	66.4	56.6	65.9
Avg. words length per tweet	6.4	6.9	6.3	6.9	5.7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

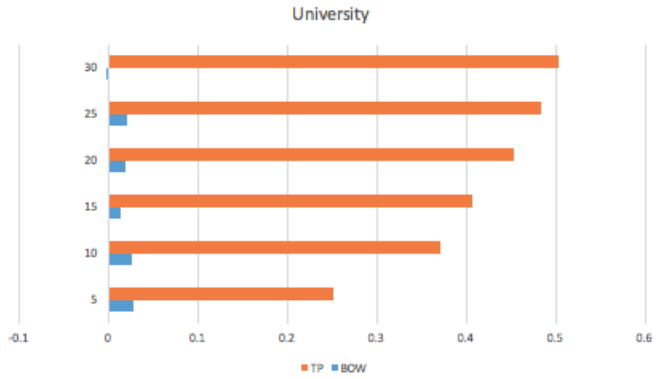
Experimental results

Silhouette coefficient values for **in-domain clustering** for tweets in **Spanish**.



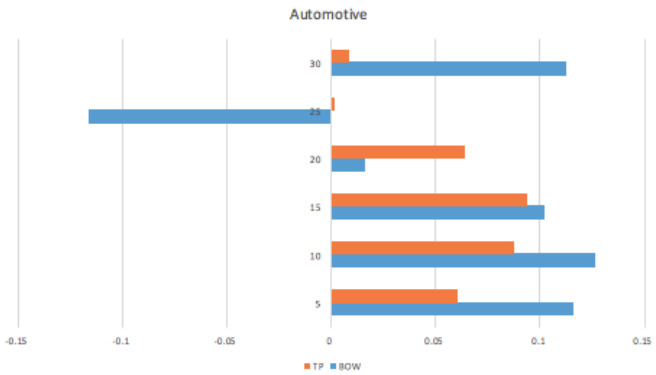


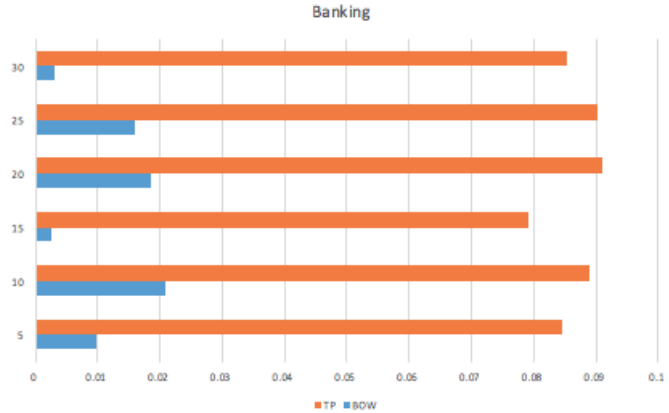


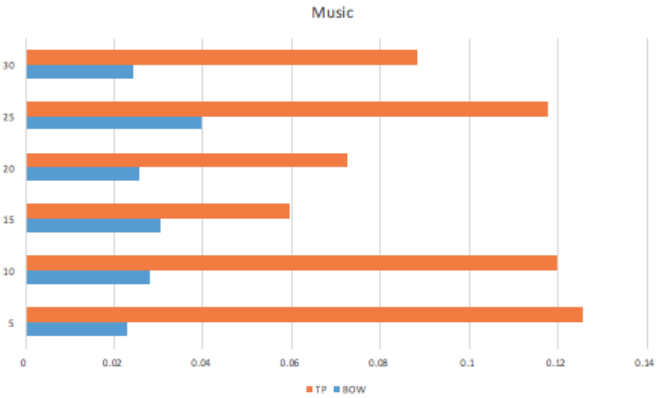


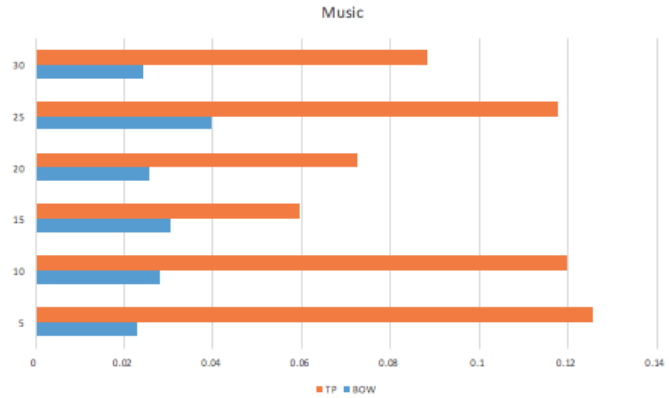
Experimental results

Silhouette coefficient values for **in-domain clustering** for tweets in **English**.



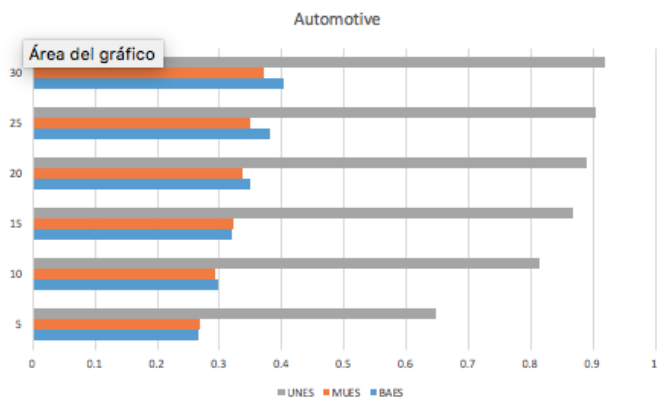


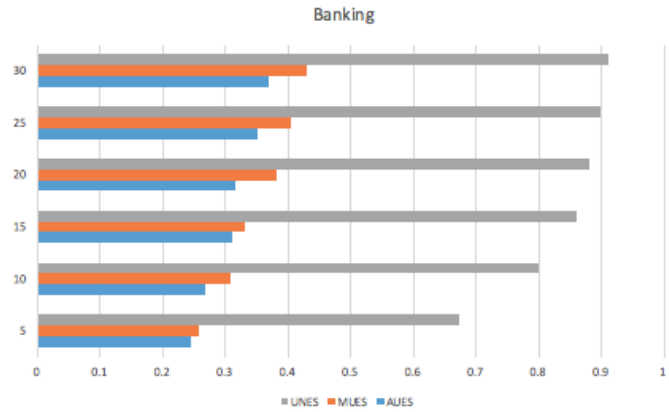


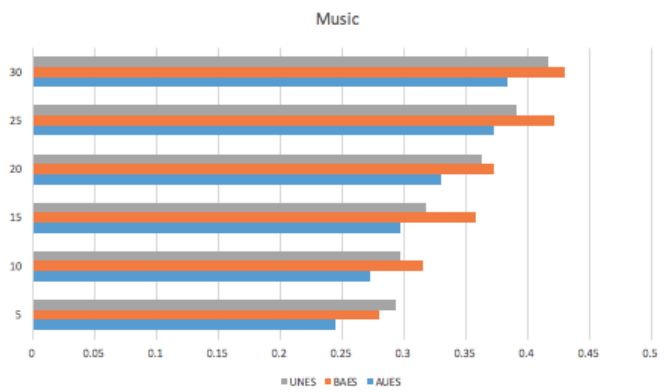


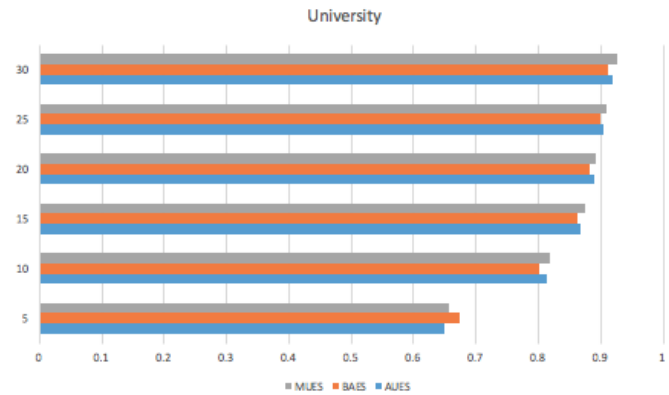
Experimental results

Silhouette coefficient values for **cross-domain clustering** for tweets in **Spanish**.



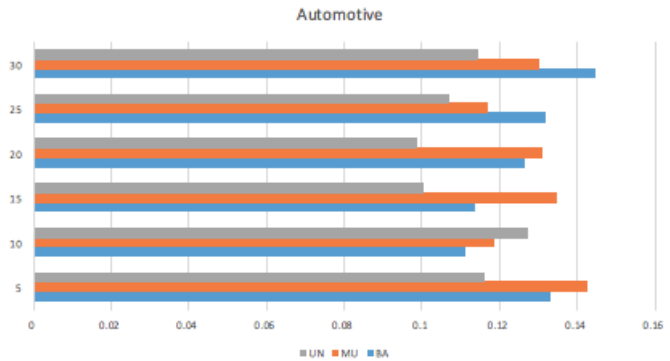


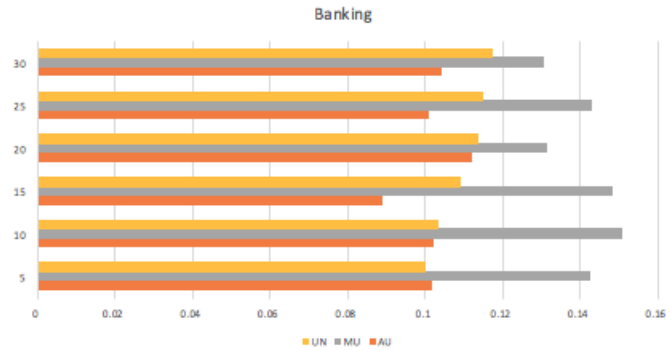


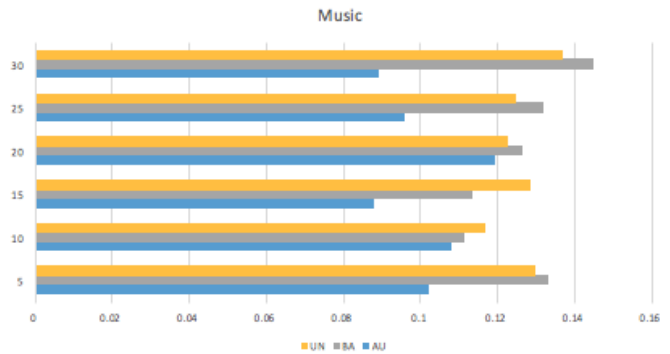


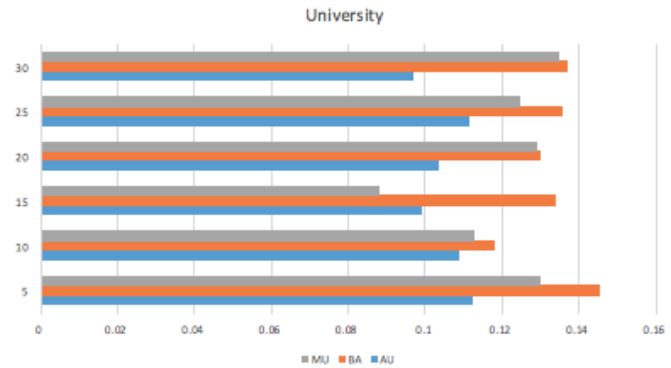
Experimental results

Silhouette coefficient values for **cross-domain clustering** for tweets in **English**.



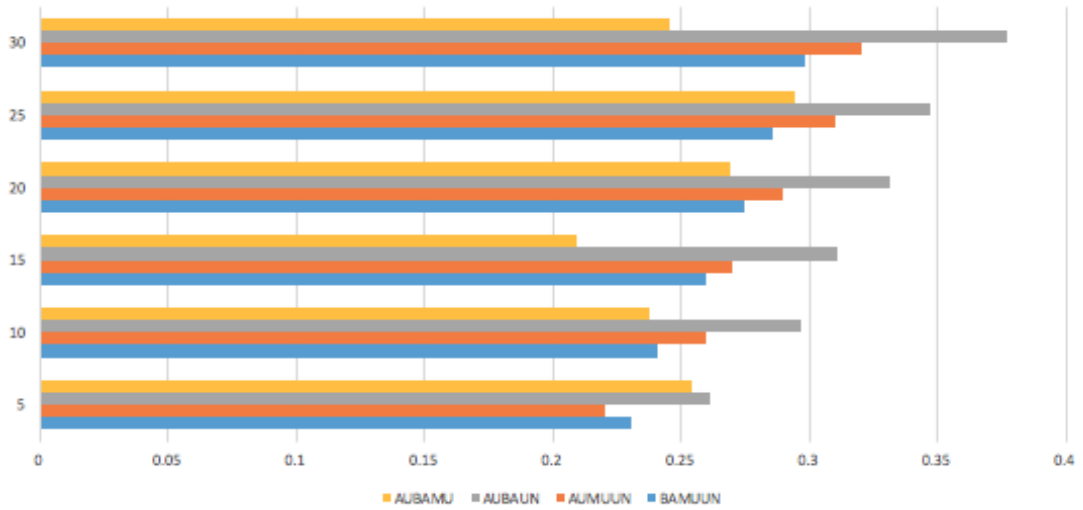






Experimental results

Silhouette coefficient values for **cross-domain clustering** of three source domains for **Spanish**.



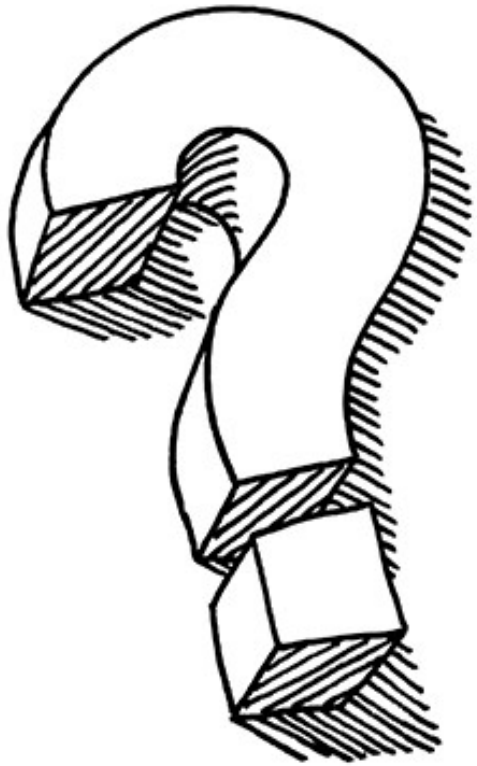
Conclusions

1. The performed experiments showed that the proposed **compact representation** allows to produce **high quality groups**, particularly for tweets in Spanish language.
2. Our experiments showed that the proposed methodology produces **high quality groups under a cross-domain scenario**, specially for tweets in Spanish.
3. 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.

Future work

1. We want to **explore the sensitivity** of the proposed compact representation to the number of selected terms by the TP technique.
2. 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.
3. 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.

Thank you !



Contact information :

- Alba Núñez,

  email : ar.nunezreyes@gmail.com

- Dr. Esaú Villatoro,

  email : evillatoro@correo.cua.uam.mx

  URL : http://ccd.cua.uam.mx/~evillatoro/

  Twitter : @EsauVT

  Twitter : @LyR_UAMC

Special thanks to...

- The Sociedad Mexicana de Inteligencia Artificial and to the Red Temática en Tecnologías del Lenguaje