Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 129-130 recibido 16-01-09, aceptado 02-03-09

On Clustering and Evaluation of Narrow Domain Short-Text Corpora*

Agrupamiento y Evaluación de Corpora de Textos Cortos y de Dominios Restringidos

David Eduardo Pinto Avendaño
Natural Language Engineering Lab., DSIC
Universidad Politécnica de Valencia

Facultad de Ciencias de la Computación, BUAP
dpinto@cs.buap.mx

Resumen: Tesis doctoral en Informática realizada por David Eduardo Pinto Avendaño y dirigida por los doctores Paolo Rosso (Univ. Politécnica de Valencia) y Héctor Jiménez (Univ. Autónoma Metropolitana, México). El acto de defensa de tesis tuvo lugar en Valencia en Julio de 2008 ante el tribunal formado por los doctores Manuel Palomar Sanz (Univ. de Alicante), Alfonso Ureña López (Univ. de Jaén), Eneko Agirre (Univ. del País Vasco), Benno Stein (Univ. de Weimar, Alemania) y Encarna Segarra Soriano (Univ. Politécnica de Valencia). La calificación obtenida fue *Sobresaliente Cum Laude*.
Palabras clave: Agrupamiento, Evaluación, Textos cortos, Dominios restringidos

Abstract: PhD thesis in Computer Science written by David Eduardo Pinto Avendaño under the supervision of Paolo Rosso (Univ. Politécnica de Valencia) and Héctor Jiménez (Univ. Autónoma Metropolitana, México). The author was examined in July 2008 in Valencia by the following committee: Manuel Palomar Sanz (Univ. de Alicante), Alfonso Ureña López (Univ. de Jaén), Eneko Agirre (Univ. del País Vasco), Benno Stein (Weimar Univ., Germany) and Encarna Segarra Soriano (Univ. Politécnica de Valencia). The grade obtained was *Sobresaliente Cum Laude*.
Keywords: Clustering, Evaluation, Narrow Domain Short-text corpora

1. Introduction

In this Ph.D. thesis we investigate the problem of clustering a particular set of documents namely *narrow domain short texts*.

To achieve this goal, we have analysed datasets and clustering methods. Moreover, we have introduced some corpus evaluation measures, term selection techniques and clustering validity measures in order to study the following problems:

1. To determine the relative hardness of a corpus to be clustered and to study some of its features such as *shortness*, *domain broadness*, *stylometry*, *class imbalance* and *structure*.
2. To improve the state of the art of clustering narrow domain short-text corpora.

The research work we have carried out is partially focused on "short-text clustering". We consider this issue to be quite relevant, given the current and future way people use "small-language" (e.g. blogs, snippets, news and text-message generation such as email or chat). Moreover, we study the domain broadness of corpora. A corpus may be considered to be *narrow* or *wide* domain if the level of the document vocabulary overlapping is high or low, respectively. In fact, in the categorization task, it is very difficult to deal with narrow domain corpora such as scientific papers, technical reports, patents, etc.

The aim of this research work is to study possible strategies to tackle the following two problems: **a)** the low frequencies of vocabulary terms in short texts, and **b)** the high vocabulary overlapping associated to narrow domains.

Each problem alone is challenging enough, however, the clustering of narrow domain short-text corpora is considered one of the most difficult tasks of unsupervised data analysis.

* This PhD thesis was supported by the BUAP-701 PROMEP/103.5/-05/1536 grant.

2. Thesis overview

In this thesis, we deal with the treatment of narrow domain short-text collections in three areas: *evaluation*, *clustering* and *validation* of corpora.

The document is structured as follows:

In Chapter 1, we introduce basic concepts and we summarize the major contributions of the research work carried out.

Chapter 2 gives an overview of the clustering methods, clustering measures, term selection techniques and datasets used in this study.

In Chapter 3, we analyse the implications of clustering narrow domain short-text corpora, studying the role of the term selection process as well as the instability of a term selection technique based on the selection of mid-frequency terms. We also make a comparison of different clustering methods in the narrow domain short-text framework. Finally, we evaluate the performance of the term selection techniques on a standard narrow domain short-text corpus.

Chapter 4 proposes the use of several measures (most of which are introduced in this work) to assess different corpus features. These measures are tested on several corpora and implemented in the Watermarking Corpora On-line System (WaCOS)[1,2].

Chapter 5 presents a new methodology (based on term co-occurrence) for improving document representation for clustering narrow domain short texts. The self-term expansion methodology, which is independent of any external knowledge resource, greatly improves the results obtained by using classical document representation. This fact was confirmed in the practical task of word sense induction whose obtained results are shown in Chapter 6.

In Chapter 7, we study the impact of internal clustering validity measures by using narrow domain short-text corpora.

Finally, in Chapter 8 we draw the conclusions of the research that we have carried out. In this last chapter we also discuss some interesting research directions, which are derived from the obtained results of this Ph.D. thesis and which we consider to be useful for future work.

3. Thesis contributions

The major contributions of the investigations carried out are:

1. The study and introduction of evaluation measures to analyse the following features of a corpus: *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure*.
2. The development of WaCOS for the assessment of corpus features.
3. A new unsupervised methodology (which does not use any external knowledge resource) for dealing with narrow domain short-text corpora. This methodology suggests first applying self-term expansion and then term selection.

We analysed different corpus features as evidence of the relative hardness of a given corpus with respect to clustering algorithms. In particular, the degree of *shortness*, *domain broadness*, *class imbalance*, *stylometry* and *structure* were studied.

We introduced some (un)supervised measures in order to assess these features. The supervised measures were used both to evaluate the corpus features and, even more importantly, to assess the gold standard provided by experts for the corpus to be clustered. The unsupervised measures evaluate the document collections directly (i.e., without any gold standard) and, therefore, they may also be used for other purposes, for instance, to adjust clustering methods while being executed in order to improve the results.

The most successful measures were compiled in a freely functional web-based system that allows linguistics and computational linguistics researchers to easily assess the quality of corpora with respect to the aforementioned features.

The experiments conducted confirmed that the clustering of narrow domain short-text corpora is a very challenging task. However, the contributions of this research work are proof that it is possible to deal with this difficult problem. The aim is now to investigate subjective scenarios such as the blogsphere.

[1] http://nlp.cs.buap.mx/watermarker/
[2] http://nlp.dsic.upv.es:8080/watermarker/