A Particle Swarm Optimizer to Cluster Parallel Spanish-English Short-text Corpora*

Un Optimizador basado en Cúmulo de Partículas para el Agrupamiento de Textos Cortos de Colecciones Paralelas en Español-Inglés

Diego Ingaramo, Marcelo Errecalde, Leticia Cagnina
LIDIC Research Group
Universidad Nacional de San Luis
Ej. de los Andes 950
5700 San Luis, Argentina.
{daingara,merreca,lcagnina}@unsl.edu.ar

Paolo Rosso
Natural Language Engineering Lab.
ELiRF, DSIC
Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, España.
prosso@dsic.upv.es

Resumen: El agrupamiento de textos cortos es actualmente un área importante de investigación debido a su aplicabilidad en la recuperación de información desde la web, generación automática de resúmenes y minería de texto. Estos textos con frecuencia se encuentran disponibles en diferentes lenguajes y en colecciones paralelas multilingüe. Algunos trabajos previos han demostrado la efectividad de un algoritmo optimizador basado en Cúmulo de Partículas, llamado CLUDIPSO, para el agrupamiento de colecciones monolingües de documentos muy cortos. En todos los casos considerados, CLUDIPSO superó la prestación de diferentes algoritmos representativos del estado del arte en el área. Este artículo presenta un estudio preliminar mostrando la prestación de CLUDIPSO en colecciones paralelas en Español-Inglés. La idea es analizar cómo la información bilingüe puede ser incorporada al algoritmo CLUDIPSO y en qué medida esta información puede mejorar los resultados del agrupamiento. Con el objetivo de adaptar CLUDIPSO al ambiente bilingüe, se proponen y evalúan algunas alternativas. Los resultados fueron comparados considerando CLUDIPSO en ambos ambientes, bilingüe y monolingüe. El trabajo experimental muestra que la información bilingüe permite obtener resultados comparables con aquellos obtenidos con colecciones monolingües. Se requiere más trabajo de forma tal de hacer un uso efectivo de esta clase de información.
Palabras clave: Agrupamiento de Textos Cortos, Colecciones Paralelas en Español-Inglés, Optimizador basado en Cúmulo de Partículas.

Abstract: Short-texts clustering is currently an important research area because of its applicability to web information retrieval, text summarization and text mining. These texts are often available in different languages and parallel multilingual corpora. Some previous works have demonstrated the effectiveness of a discrete Particle Swarm Optimizer algorithm, named CLUDIPSO, for clustering monolingual corpora containing very short documents. In all the considered cases, CLUDIPSO outperformed different algorithms representative of the state-of-the-art in the area. This paper presents a preliminary study showing the performance of CLUDIPSO on parallel Spanish-English corpora. The idea is to analyze how this bilingual information can be incorporated in the CLUDIPSO algorithm and to what extent this information can improve the clustering results. In order to adapt CLUDIPSO to a bilingual environment, some alternatives are proposed and evaluated. The results were compared considering CLUDIPSO in both environments, bilingual and monolingual. The experimental work shows that bilingual information allows to obtain just comparable results to those obtained with monolingual corpora. More work is required to make an effective use of this kind of information.
Keywords: Clustering of Short Texts, Parallel Spanish-English Corpora, Particle Swarm Optimizer.

1 Introduction

Vast amounts of information are actually available on internet in documents such as news, academic works, web-repositories, etc. many of which are in a short-text format. Document clustering groups automatically a large set of documents into different clusters. In this context, the clustering of short-text corpora, is one of the most difficult tasks in natural language processing due to the low frequencies of terms in the documents.

In document clustering, the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that in realistic document clustering problems, results can not usually be evaluated with typical *external* measures like *F*-Measure and Entropy, because the correct categorizations specified by a human expert are not available. Therefore, the quality of the resulting groups is evaluated with respect to *structural* properties expressed in different *Internal Clustering Validity Measures* (ICVMs). Classical ICVMs used as cluster validity measures include the Dunn and Davies-Bouldin indexes, the *Global Silhouette* (GS) coefficient and, new graph-based measures such as the *Expected Density Measure* and the λ-Measure (see (Ingaramo et al., 2008) for detailed descriptions of these ICVMs).

The use of ICVMs has not been limited to the cluster evaluation stage. Different ICVMs have also been used as explicit *objective functions* that the clustering algorithm attempts to optimize *during* the grouping process. This approach has been adopted, for example, in CLUDIPSO, a discrete Particle Swarm Optimizer (PSO) which obtained in previous work (Ingaramo et al., 2009) interesting results on small short-text collections. This algorithm uses the unsupervised measure GS as objective function to be optimized.

CLUDIPSO, and other techniques to cluster short documents (see for example (Alexandrov, Gelbukh, and Rosso, 2005; Pinto, Benedí, and Rosso, 2007; He et al., 2007; Carullo, Binaghi, and Gallo, ; Hu et al., 2009)), have obtained good results with documents written in the same language, i.e. monolingual environments. However, nowadays, there are many linguistics resources which make available information written in different languages. These resources include for example, parallel and aligned parallel corpora which have been used in different applications such as extraction of word translation equivalents (Ribeiro and Lopes, 2000) and studies of lexical semantics (Sharoff, 2002), among others. However, little effort has been dedicated to analyze if this multilingual information could help to improve the results that classical text analysis methods like text categorization and clustering obtain on monolingual corpora.

In this work, information from parallel Spanish-English corpora is used in short-text clustering and two main research questions are addressed: 1) how this bilingual information can be incorporated in the CLUDIPSO algorithm, and 2) to what extent this information can improve the clustering results. The first aspect is addressed in Section 2.1, where some modifications are introduced in CLUDIPSO to incorporate information from parallel Spanish-English short-text corpora. The second one is analyzed in Section 3 which includes results of CLUDIPSO with Spanish and English documents (taken separately) and results with approaches that simultaneously consider documents written in both languages.

The remainder of the paper is organized as follows. Section 2 describes CLUDIPSO, the PSO-based algorithm under study and the proposed alternatives to bilingual document clustering. Section 3 describes some general features of the corpora used in the experiments, the experimental setup and the analysis of the results obtained from the empirical study. Finally, some general conclusions are drawn and present and future work is discussed in Section 4.

2 The CLUDIPSO Algorithm

CLUDIPSO (CLUstering with a DIscrete Particle Swarm Optimization), is based on a PSO (Eberhart and Kennedy, 1995) algorithm that operates on a population of par-

* This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i, and by UPV as part of the PAID-02-10 programme (grant. no. 2257).

ticles. Each particle, in the basic version of PSO, is a real numbers vector which represents a position in the search space defined by the variables corresponding to the problem to solve. The best position found so far for the swarm (*gbest*) and the best position reached by each particle (*pbest*) are recorded at each cycle (iteration of the algorithm). The particles evolve at each cycle using two updating formulas, one for velocity (Equation (1)) and another for position (Equation (2)).

$$v_{id} = w(v_{id} + \gamma_1(pb_{id} - par_{id}) + \gamma_2(pg_d - par_{id})) \quad (1)$$

$$par_{id} = par_{id} + v_{id} \quad (2)$$

where par_{id} is the value of the particle i at the dimension d, v_{id} is the velocity of particle i at the dimension d, w is the inertia factor (Eberhart and Shi, 1998) whose goal is to balance global exploration and local exploitation, γ_1 is the personal learning factor, and γ_2 the social learning factor, both multiplied by 2 different random numbers within the range [0, 1]. pb_{id} is the best position reached by the particle i and pg_d is the best position reached by any particle in the swarm.

In the discrete version CLUDIPSO, each valid clustering is represented with a particle. The particles are n-dimensional integer vectors, where n is the number of documents in the corpus. Since the task was modeled with a discrete approach, a new formula was developed for updating the positions (shown in Equation (3)).

$$par_{id} = pb_{id} \quad (3)$$

where par_{id} is the value of the particle i at the dimension d and pb_{id} is the best position reached by the particle i until that moment. This equation was introduced with the objective of accelerate the convergence velocity of the algorithm (principal incoming of discrete PSO models). It is important to note that in this approach the process of updating particles is not as direct as in the continuous case (basic PSO algorithm). In CLUDIPSO, the updating process is not carried out on all dimensions at each iteration. In order to determine which dimensions of a particle will be updated the following steps are performed: 1) all dimensions of the velocity vector are normalized in the [0, 1] range, according to the process proposed by Hu et al. (Hu, Eberhart, and Shi, 2003) for a discrete PSO version; 2) a random number $r \in [0, 1]$ is calculated; 3) all the dimensions (in the velocity vector) higher than r are selected in the position vector, and updated using the Equation (3).
A Dynamic mutation operator (Cagnina, Esquivel, and Gallard, 2004) is applied with a *pm*-probability calculated with the total number of iterations in the algorithm (*cycles*) and the current cycle number: $pm = max\_pm - \frac{max\_pm - min\_pm}{max\_cycle} * current\_cycle$. Where max_pm and min_pm are the maximum and minimum values that pm can take, max_cycle is the total number of cycles and the current cycle in the iterative process is $current\_cycle$. The mutation operation is applied if the particle is the same that its own *pbest*, as was suggest by (Hu, Eberhart, and Shi, 2003). The mutation operator swaps two random dimensions of the particle and in that way avoids premature convergence.

Global Silhouette (GS) Coefficient was used as an *objective function* $f(p_i)$, because gives a reasonable estimation of the quality of the obtained groups. The optimization of GS drives the entire CLUDIPSO process.
The GS measure combines two key aspects to determine the quality of a given clustering: *cohesion* and *separation*. Cohesion measures how closely related are the objects in a same cluster whereas separation quantifies how distinct (well-separated) a cluster from other clusters is. The GS coefficient of a clustering is the average cluster silhouette of all the obtained groups. The cluster silhouette of a cluster C also is an average silhouette coefficient but, in this case, of all objects belonging to C. Therefore, the fundamental component of this measure is the formula used for determining the silhouette coefficient of any arbitrary object i, that we will refer as $s(i)$ and that is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (4)$$

with $-1 \leq s(i) \leq 1$. The $a(i)$ value denotes the average dissimilarity of the object i to the remaining objects in its own cluster, and $b(i)$ is the average dissimilarity of the object i to all objects in the nearest cluster. From this formula it can be observed that negative values for this measure are undesirable and

that for this coefficient values as close to 1 as possible are desirable.

2.1 CLUDIPSO adaptations to bilingual contexts

The idea of using bilingual information in a clustering algorithm is motivated by similar reasons to those giving origin to *ensemble methods* (Dietterich, 2000): the combined use of information (about the same problem) coming from different sources, can be more effective than consider this information separately. In the context of our work, this intuitive idea consists in determining how the information obtained from the alignment of documents in parallel corpora can be used by a clustering algorithm instead of clustering the documents in the different languages separately.

CLUDIPSO allows to combine this kind of information in a relatively direct way: using the values that the evaluation function obtains with the documents in each language, and combining these values using different criteria. For example, the minimum (MIN), maximum (MAX) or average (AVG) value between the evaluation function's values obtained with the documents in each language could be used. Or simply taking the value that the evaluation function returns with documents in both languages, but alternating in each iteration the language used to evaluate this function. More formally:
Let D be a bilingual environment (parallel corpus) with Spanish-English documents. Then, each document $d_i \in D$ will be represented by an English text, $d_{i_{EN}}$ and the corresponding representation of d_i in Spanish language, $d_{i_{ES}}$. Let D_{EN} and D_{ES} be the documents in D in its English and Spanish representation respectively.

We will denote as $CLUDIPSO_{MULTI}$ the CLUDIPSO version that incorporates D_{EN} and D_{ES} information in the clustering process. $CLUDIPSO_{MULTI}$ uses the available bilingual information by adapting the CLUDIPSO *evaluation function step* (see section 2). The four alternatives that were considered to incorporate the bilingual information in the evaluation function are described below.

Let p be a particle representing a possible solution (clustering) and let $f(p_{en})$ and $f(p_{es})$ be the fitness values of p with respect to D_{EN} and D_{ES} respectively. Then, the fitness value for:

1. $CLUDIPSO_{MULTI-MAX}$ is defined as: $f(p) = max(f(p_{en}), f(p_{es}))$.

2. $CLUDIPSO_{MULTI-MIN}$ is defined as: $f(p) = min(f(p_{en}), f(p_{es}))$.

3. $CLUDIPSO_{MULTI-AVG}$ is defined as: $f(p) = \frac{f(p_{en})+f(p_{es})}{2}$.

4. $CLUDIPSO_{MULTI-ALT}$ in the iteration i is $f(p_{en})$ if i is odd and $f(p_{es})$ in other case.

Thus, for example, if the Silhouette Coefficient is used as evaluation function, $CLUDIPSO_{MULTI-AVG}$ will use as fitness value the average value obtained from the Silhouette value for the clustering p_i using the English documents in D and the Silhouette value for the same clustering, but using in this case the Spanish documents.

3 Experimental Setting and Analysis of Results

For the experimental work, two collections with different levels of complexity with respect to the size, length of documents and vocabulary overlapping were selected: SEPLN-CICLing and JRC-Acquis. Table 1 shows some general features of these corpora: corpus size (CS), number of categories and documents ($|\mathcal{C}|$ and $|\mathcal{D}|$ respectively), total number of terms in the collection ($|\mathcal{T}|$), vocabulary size ($|\mathcal{V}|$) and average number of terms per document ($\overline{\mathcal{T}}_d$).

The first one, SEPLN-CICLing, is a small collection based on CICLing-2002[1] scientific abstract corpus which has been intensively used in different works (Ingaramo et al., 2009; Ingaramo, Errecalde, and Rosso, 2010; Errecalde, Ingaramo, and Rosso, 2010). This corpus was enriched with bilingual abstracts (in Spanish and English) of the SEPLN[2].

The SEPLN-CICLing$_{EN}$ corpus was composed by the English abstracts of CICLing-2002 and SEPLN. The SEPLN-CICLing$_{ES}$ was obtained adding to the Spanish version of SEPLN abstracts the manual translation of the CICLing abstracts.

[1] http://www.cicling.org/2002/
[2] http://www.sepln.org/

JRC-Acquis refers to a sub-collection of the Acquis (Steinberger et al., 2006), a popular multilingual collection with legal documents and laws corresponding to different countries of the European Union. For this work, we selected 563 documents in the English and Spanish versions, denoted JRC-Acquis$_{EN}$ and JRC-Acquis$_{ES}$ respectively.

Corpora	CS	$\|\mathcal{C}\|$	$\|\mathcal{D}\|$
SEPLN-CICLing$_{EN}$	25	4	48
SEPLN-CICLing$_{ES}$	21	4	48
JRC-Acquis$_{EN}$	798	6	563
JRC-Acquis$_{ES}$	870	6	563
Corpora	$\|\mathcal{T}\|$	$\|\mathcal{V}\|$	$\hat{\mathcal{T}}_d$
SEPLN-CICLing$_{EN}$	3143	1169	65.48
SEPLN-CICLing$_{ES}$	2129	904	44.35
JRC-Acquis$_{EN}$	110887	7391	196.96
JRC-Acquis$_{ES}$	121953	7903	216.61

Table 1: Features of the collections used in the experimental work.

Due to the fact the gold standard is known for each of the two sub-collections, the quality of the results was evaluated by using the classical (external) F-measure. Each algorithm generated 50 independent runs per collection after performing 10,000 iterations (stopping condition). The reported results in Table 2, correspond to the minimum (F_{min}), maximum (F_{max}) and average (F_{avg}) F-measure values obtained by the different the algorithm of CLUDIPSO. The values highlighted in bold, indicate the best obtained results.

	SEPLN-CICLing		
Algorithm	F_{avg}	F_{min}	F_{max}
$CLUDIPSO-EN$	**0.75**	**0.63**	0.85
$CLUDIPSO-ES$	0.7	0.6	0.83
$CLUDIPSO_{MULTI-ALT}$	0.52	0.39	0.68
$CLUDIPSO_{MULTI-AVG}$	0.7	**0.63**	0.79
$CLUDIPSO_{MULTI-MAX}$	0.71	0.62	0.83
$CLUDIPSO_{MULTI-MIN}$	0.7	0.62	**0.87**
	JRC-Acquis		
Algorithm	F_{avg}	F_{min}	F_{max}
$CLUDIPSO-EN$	**0.29**	**0.26**	**0.33**
$CLUDIPSO-ES$	**0.29**	0.21	0.31
$CLUDIPSO_{MULTI-ALT}$	**0.29**	**0.26**	0.32
$CLUDIPSO_{MULTI-AVG}$	0.28	**0.26**	0.31
$CLUDIPSO_{MULTI-MAX}$	**0.29**	0.25	0.32
$CLUDIPSO_{MULTI-MIN}$	**0.29**	0.25	0.32

Table 2: F-measures values per collection.

With the smaller SEPLN-CICLing collection it is observed that the version $CLUDIPSO-EN$ obtained the best F_{avg} value and $CLUDIPSO_{MULTI}$ the best F_{max} value. Similar results can be observed with the F_{min} values in both cases. It should be noted that $CLUDIPSO_{MULTI}$ slightly overcomes the algorithm $CLUDIPSO-ES$ but not the $CLUDIPSO-EN$ although both algorithms have a similar performance no matter the language used.

With respect to the obtained results with the larger collection JRC-Acquis, $CLUDIPSO_{MULTI}$ gets similar values to $CLUDIPSO-EN$ and a minimum improvement compared to $CLUDIPSO-ES$ (like in SEPLN-CICLing). It should be noted that $CLUDIPSO_{MULTI}$ improves F_{min} in all the cases, excluding $CLUDIPSO_{MULTI-ALT}$. Experiments carried out show an improvement related to $CLUDIPSO-ES$ but results are similar to $CLUDIPSO-EN$. However, in JRC-Acquis results needs to be improved for both languages.

4 Conclusions and Future Work

This work presents a preliminary study of performance of different versions of $CLUDIPSO_{MULTI}$, a novel bilingual PSO-based clustering algorithm. The results obtained by $CLUDIPSO_{MULTI}$ on Spanish-English corpora indicate that the approach is an alternative to solve bilingual clustering of small short-text corpora, although no significant improvement was obtained so far with respect to the monolingual PSO-based version CLUDIPSO. $CLUDIPSO_{MULTI}$ was also tested with a larger size collection and the performance was comparable to its predecessor monolingual CLUDIPSO, possibly the lack of improvement it is due to the limitations derived by a wide search space in large document collection.

Future works include text-enrichment of documents, combining both documents representations by using a term selection technique and also, including bilingual information into a novel on going version named $CLUDIPSO^{\star}$ considering newer mechanisms to incorporate the bilingual knowledge.

The proposed algorithm was tested with Spanish-English corpora although other bilingual corpora could be used in a future.

In order to tackle the problem of the size of the particle that $CLUDIPSO_{MULTI}$ suffers with collection such as JRC-Acquis,

we aim at investigating the possibility of dividing the particle in two: a part of the particle would deal with the representation in English and the other one with the Spanish representation.

References

Alexandrov, M., A. Gelbukh, and P. Rosso. 2005. An approach to clustering abstracts. In Andrés Montoyo, Rafael Muñoz, and Elisabeth Metais, editors, *Natural Language Processing and Information Systems*, volume 3513 of *LNCS*. Springer Berlin / Heidelberg, pages 1–10.

Cagnina, L., S. Esquivel, and R. Gallard. 2004. Particle swarm optimization for sequencing problems: a case study. In *Congress on Evolutionary Computation*, pages 536–541.

Carullo, M., E. Binaghi, and I. Gallo. An online document clustering technique for short web contents. *Pattern Recognition Letters*, 30.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Int. Workshop on Multiple Classifier Systems*, pages 1–15. Springer-Verlag.

Eberhart, R. and J. Kennedy. 1995. A new optimizer using particle swarm theory. In *Proc. of the Sixth International Symposium on Micro Machine and Human Science, MHS'95*, pages 39–43, Nagoya, Japan.

Eberhart, R. and Y. Shi. 1998. A modified particle swarm optimizer. In *International Conference on Evolutionary Computation*. IEEE Service Center.

Errecalde, M., D. Ingaramo, and P. Rosso. 2010. Itsa*: an effective iterative method for short-text clustering tasks. In *Proc. of the 23rd Int. Conf. on Industrial Engineering and other Applications of Applied Intelligent Systems, IEA/AIE 2010*, pages 550–559, Berlin, Heidelberg. Springer-Verlag.

He, H., B. Chen, W. Xu, and J. Guo. 2007. Short text feature extraction and clustering for web topic mining. In *Proc. of the Third Int. Conf. on Semantics, Knowledge and Grid*, pages 382–385, Washington, DC, USA. IEEE Computer Society.

Hu, X., R. Eberhart, and Y. Shi. 2003. Swarm intelligence for permutation optimization: a case study on n-queens problem. In *Proc. of the IEEE Swarm Intelligence Symposium*, pages 243–246.

Hu, X., N. Sun, C. Zhang, and T. Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proc. of the 18th ACM Conf. on Information and knowledge management*, pages 919–928, New York, NY, USA. ACM.

Ingaramo, D., M. Errecalde, L. Cagnina, and P. Rosso, 2009. *Computational Intelligence and Bioengineering*, chapter Particle Swarm Optimization for clustering short-text corpora, pages 3–19. IOS press.

Ingaramo, D., M. Errecalde, and P. Rosso. 2010. A general bio-inspired method to improve the short-text clustering task. In *Proc. of CICLing 2010*, LNCS 6008, pages 661–672. Springer-Verlag.

Ingaramo, D., David Pinto, P. Rosso, and M. Errecalde. 2008. Evaluation of internal validity measures in short-text corpora. In *Proc. of the International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2008*, volume 4919 of *Lecture Notes in Computer Science*, pages 555–567. Springer-Verlag.

Pinto, D., J. M. Benedí, and P. Rosso. 2007. Clustering narrow-domain short texts by using the Kullback-Leibler distance. In *Proc. of the International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 611–622. Springer-Verlag.

Ribeiro, A. and G. Pereira Lopes. 2000. Extracting portuguese-spanish word translations from aligned parallel texts. *Procesamiento del lenguaje natural*, 26:73–80.

Sharoff, S. 2002. Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proc. of Language Resources and Evaluation Conference (LREC02)*, pages 447–452.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genoa, Italy.