The University of Alicante at MultiLing 2015: approach, results and further insights

Marta Vicente
University of Alicante
Apdo. de correos 99
E-03080 Alicante, Spain
mvicente@dlsi.ua.es

Óscar Alcón
University of Alicante
Apdo. de correos 99
E-03080 Alicante, Spain
oalcon@dlsi.ua.es

Elena Lloret
University of Alicante
Apdo. de correos 99
E-03080 Alicante, Spain
elloret@dlsi.ua.es

Abstract

In this paper we present the approach and results of our participation in the 2015 MultiLing Single-document Summarization task. Our approach is based on the Principal Component Analysis (PCA) technique enhanced with lexical-semantic knowledge. For testing our approach, different configurations were set up, thus generating different types of summaries (i.e., generic and topic-focused), as well as testing some language-specific resources on top of the language-independent basic PCA approach, submitting a total of 6 runs for each selected language (English, German, and Spanish). Our participation in MultiLing has been very positive, ranking at intermediate positions when compared to the other participant systems, showing that PCA is a good technique for generating language-independent summaries, but the addition of lexical-semantic knowledge may heavily depend on the size and quality of the resources available for each language.

1 Introduction

Currently, the amount of on-line information generated per week reaches the same quantity of data that the one produced in the Internet between its inception and 2003, time of the Social Network emergency (Cambria and White, 2014). Moreover, the production of such volume of data is delivered in multiple languages, and accessing the relevant content of information or extracting the main features of documents in a competitive time is more and more challenging. Therefore, automatic tasks that can help processing all this information, such as multilingual text summarization techniques, are now becoming essential.

Back in 2011, the Text Analysis Conference MultiLing Pilot task[1] was first introduced as an effort of the community to promote and support the development of multilingual document summarization research. Considering the impact of this shared tasks in the progress of natural language processing technologies, a mutlilingual summarization workshop was also organized in 2013[2].

Nowadays, in 2015, we take part in the 3rd MultiLing event[3]. In this edition, new tasks have been added in order to adapt to social requirements. There were the traditional Multilingual Multi-document and Single-document Summarization (MMS and MSS), coming from previous events, but also new summarization tasks related to Online Fora (OnForumS) - on how to deal with reader comments- and Call Center Conversation (CCCS) - from spoken conversations to textual synopses.

Taking into consideration the interest that multilingual summarization approaches is gaining among the research community, and the positive impact and benefits it may have for the society, the objective of this paper is to present a multilingual summarization approach within the MultiLing 2015 competition, discussing its potentials and limitations, and providing some insights of the future of this type of summarization based on the average results obtained by us and other participants as well.

The remaining of the paper is organized as follows. In Section 2 we review the most relevant multilingual summarization approaches, some of them participating in previous MultiLing events. In Section 3, we explain our multilingual summarization approach and the required language-dependent knowledge. Section 4 describes the

[1] http://www.nist.gov/tac/2011/Summarization/

[2] http://multiling.iit.demokritos.gr/pages/view/662/multiling-2013

[3] http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015

Proceedings of the SIGDIAL 2015 Conference, pages 250–259,
Prague, Czech Republic, 2-4 September 2015. ©2015 Association for Computational Linguistics

task in which we participated, and the experiments performed. Furthermore, the results together with their discussion and comparison to other participants are provided in Section 5, followed by an analysis of the potentials and limitations of our approach in Section 6. Finally, the main conclusions are outlined in Section 7.

2 Related work

Eight teams participated in the Multilingual Pilot task in 2011, five of them testing their approaches for all the proposed languages (Arabic, Czech, English, French, Greek, Hebrew, and Hindi) (Giannakopoulos et al., 2011). Two systems are worth mentioning. On the one hand, the CLASSY system (Conroy et al., 2011) that ranked 2nd or 3rd in 5 out 7 languages. The main feature of this approach was that a model was first trained on a corpus of newswire taken from Wikinews, and then term scoring was limited to the naive Bayes term weighting. The final process of sentence selection was performed using non-negative matrix factorization and integer programming techniques. On the other hand, the best system on average was the one in (Steinberger et al., 2011), performing the 1st in five of the seven languages, and 4th in the two remaining ones. This approach did not used any language-dependent resources, apart from a stopword list for each language, and it relied on Latent Semantic Analysis and Singular Value Decomposition.

In the 2013 MultiLing edition, four teams participated submitting six systems to the task (Giannakopoulos, 2013). For their assessment in (Kubina et al., 2013), they were denoted as MUSE, MD, AIS and LAN. We briefly reviewed these approaches. MUSE (Litvak and Last, 2013), is a supervised learning approach that scores sets of sentences by means of a genetic algorithm. MD (Conroy et al., 2013) developed techniques both for MMS and MSS, examining the impact of dimensionality reduction and offering different weighting methods in the experiments: either considering the frequency of terms or applying a variant of TextRank, among others. Adapting their techniques to Arabic and English languages, the LAN team (El-Haj and Rayson, 2013) implements a system that recovers the most significant sentences for the summary using word frequency and keyness score, introducing a statistic approach that extracts those sentences with the maximum sum of log likelihood. Contrary to the previously described systems, mostly based in frequency of terms, AIS (Anechitei and Ignat, 2013) presented an approach based on the analysis of the discourse structure, exploiting, therefore, cohesion and coherence properties from the source articles. Although some of these participants performed well, achieving similar results as the ones obtained by human summaries, the WBU approach (Steinberger, 2013), was again the best performing summarization system in this MultiLing edition, reaching the first position in 5 of the 10 languages. Specifically, it was an improved version of the best-performing approach in MultiLing 2011 (Steinberger et al., 2011).

Outside the MultiLing competitions, other research works have been recently proposed, obtaining better results than existing commercial multilingual summarizers. An example of this can be found in (Lloret and Palomar, 2011) were three different approaches were analyzed and tested: i) one using language-independent techniques; ii) one with language-dependent resources; and iii) one using machine translation to monolingual summarization. The results obtained showed that having high-quality language specific resources often led to the best results; however, a simple language-independent approach based on term frequency was competitive enough, avoiding the effort needed to develop and/or obtain the particular resources for each language, when they were not available.

Having revised different multilingual summarization approaches, the main contribution of our paper is to propose a novel approach based on the Principal Component Analysis (PCA) technique, studying the influence of lexical-semantic knowledge to the base approach. To the best of our knowledge, although PCA has been already used for text summarization (for instance, in (Lee et al., 2003)), it has never been tested with the addition of semantic knowledge, nor in the context of multilingual summarization. Given that it bears some relation to LSA and SVD techniques, and it has been shown that such techniques are very competitive, MultiLing 2015 is the perfect context to test it.

3 The UA-DLSI Approach

In this Section, we present our proposed multilingual summarization approach (i.e., UA-DLSI approach).

As it was previously mentioned, the main technique that characterise the UA-DLSI approach is the Principal Component Analysis (PCA). PCA is a statistical technique focused on the synthesis of information to compress and interpret the data (Estellés Arolas et al., 2010).

As a method for developing summarization systems, PCA provides a way to determine the most relevant key terms of a document. It has been often employed in conjunction with other data mining techniques, such as Semantic Vector Space model (Vikas et al., 2008) or Singular Value Decomposition (Lee et al., 2005), using term-based frequency methods. Our main difference with respect to other summarization PCA-based approaches is the incorporation of lexical-semantic knowledge into the PCA technique, since it is necessary to go beyond the terms, and determine the meaningful sentences. Moreover, to finish the process, some strategies for selecting relevant information (in our case, choosing the most relevant sentences) needs to be defined as well.

For developing our UA-DLSI approach, we relied on the summary process stages outlined in (Sparck-Jones, 1999): 1) *interpretation*, 2) *transformation* and, finally, 3) the *summary generation*.

Interpretation. The first stage of our approach includes a linguistic and lexical-semantic processing (this latter part is optional). For the linguistic processing, sentence segmentation, tokenization and stopwords removal is applied. For the lexical-semantic processing, a named entity recognizer (*Standford Named Entity Recognizer*[4]) and semantic resources, such as *WordNet* (Miller, 1995) and *EuroWordNet* (Vossen, 2004) are employed. Whereas named entity recognizers mainly provide the identification of person, organization and place names in a document (Tjong et al., 2003), the semantic resources used comprises a set of synonyms grouped by means of the *synsets* that allow us to work with concept better than just with terms. In this manner, we group a set of synonyms under the same concept. For instance, *detonation* and *explosion* are different words but their share the same synset (*07323181*), so we would keep them as a single concept. For identifying concepts, we relied on the most frequent sense approach, and therefore, the process searches for the first synset of each word in the document, which corresponds to its most probable meaning. If two words have the same first synset, we will assume that they are synonyms and their occurrences will be added together.

The result of this stage is to build an initial lexical-semantic matrix, where for each sentence (rows in our matrix), we identify the units that will be later taken into account (i.e., terms, named entities, and/or concepts) which will correspond to the columns.

Transformation. It is in the transformation stage that we use the PCA method. In our approach, PCA is applied using the PCA_transform Java library[5] to process the covariance matrix that is computed from the lexical-semantic matrix obtained in the previous stage. Once PCA has been applied over the covariance matrix, the principal components (eigenvectors) and its corresponding weight (eigenvalue) are obtained. The eigenvectors are composed by the contribution of each variable, which determines the importance of the variable in the eigenvector. Moreover, the eigenvectors are derived in decreasing order of importance. In this manner, an eigenvector with high eigenvalue carries a great amount of information. Therefore, the first eigenvectors collect the major part of the information extracted from the covariance matrix, and they will be used for determining the most important sentences in the document, as it will be next shown.

Summary generation. In this final stage, the relevant sentences are selected and extracted, thus producing an extractive summary. Since from the previous stage, only the key elements (e.g., concepts) were determined, it is necessary to define some strategies for deciding which sentences containing these elements will be finally taking part in the summary.

Two strategies were proposed for selecting and ordering the most relevant sentences from the document, leading to two types of summaries: one generic and one topic-focused. In this manner, taking into account the element with the highest value for each eigenvector from the PCA matrix, we select and extract:

- one sentence (searching in order of appearance in the original text) in which

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

[5] https://github.com/mkobos/pca_transform

such concept[6] appears. During this process, if a sentence had been already selected by a previous concept to take part in the summary, we would select and extract the following sentence in which the concept appears (generic summary).

- all the sentences (searched in order of appearance in the original text) in which such concept appears (topic-focused summary).

Regarding these strategies, it is worth mentioning that if we found different concepts with the same highest value for the same eigenvector, we would extract the corresponding sentences for all these concepts. In the same manner, if a synset is represented by several synonyms, we would extract the corresponding sentences for each of these synonyms.

4 Experimental Setup

This section describes the MultiLing 2015 task in which we participated, together with the dataset employed, and the explanation of the different variants of our approach submitted to the competition.

4.1 MSS - Multilingual Single-Document Summarization Task

The Multilingual Single Document Summarization task was initially proposed in MultiLing 2013, targeting the same goal in the current edition: to evaluate the performance of participant systems whose work is focused on generating a single document summary for all the given Wikipedia articles in some of the languages provided (at least the participants should select three languages). In the context of MultiLing 2015, two datasets were provided for the MSS task: a training dataset, containing 30 articles for each of the 38 available languages with their corresponding human-generated summaries; and a test dataset, which contains the same number of documents per language, but different from the training dataset, the human summaries were not provided. For both datasets, the character length that the target automatic summaries should aim was also provided (i.e., the *target length*), which coincided with the length of the human summaries that will be later used in the evaluation. Each automatic summary had to be as close to the target length provided as possible, and summaries exceeding the given target length were truncated to it.

In order to prove the adequacy of our approach to select the relevant sentences from a document, we decided to start testing it within small goals to be able to analyze and further improve the proposed approach. This was the main reason for participating in the MSS task rather than in the MMS, which had implied more complexity.

Concerning the language choice, since one of our main objectives was to evaluate the impact of lexical-semantic knowledge in the summary generation, some language-dependent resources were necessary (e.g. WordNet and EuroWordNet). The availability of these resources also conditioned the languages that were chosen for testing our apporach, in our case: English, German, and Spanish.

For each language considered, we computed the average length of the Wikipedia articles in the test corpus, both in characters and words. These figures are shown in Table 1. In addition, we also provide the target summary length (in characters) and the compression ratio for the summaries. As it can be seen, the length of the summaries compared to the original length of the Wikipedia articles (i.e., compression ratio) is very short, always below 10%. This means that generated summaries have to be very concise and precise in selecting the most relevant information.

	English	**Spanish**	**German**
Characters	25850	39202	38905
Words	4223	6271	5245
Target length	1858	2044	1071
Compression ratio	7.19%	5.21%	2.75%

Table 1: Average length (words and characters) of the test dataset, and target length and compression ratio for the summaries

4.2 Configuring the UA-DLSI approach to the MSS task

Having provided the information about the general multilingual summarization process in Section 3, and since each participant in the MSS task was allowed to submit up to six approaches, different versions of our approach were set to participate in MultiLing 2015.

Apart of the two types of summaries that could

[6] Concepts here refer to the possible elements that the matrix can have, e.g. named entities, synsets, or terms

be generated with our approach (*T1: generic summary*; *T3: topic-focused summary*), the incorporation of lexical-semantic knowledge was an optional substage, so we decided to test our approach also without any type of semantic knowledge, other than a list of stopwords for each language (*LI: language-independent; LEX: using lexical knowlege (named entity recognition); SEM: using semantic knowledge (i.e., WordNet and EuroWordNet)*). This way the performance of a fully language-independent summarization approach based on PCA could be also analyzed. Moreover, due to the nature of the test dataset (Wikipedia articles), all documents included headings for structuring different sections within them, so we opt for taking advantage of this information, considering only the words in these headings for the matrix construction (*OWFH*), instead of working with all words in the document, except stopwords (*AW*). Headings usually contain important concepts that reflect the main topic of the section that follows. Considering only this words, we also reduce the amount of information we have to process by 99% of the PCA matrix.

Therefore, given the impossibility to test all the variations taking into account these issues, our submitted approaches for MultiLing 2015, specifying also their priority, were the following:

- *T1_LI_AW* (UA-DLSI-*lang*-1): generic language-independent summarizer considering all words in the documents.
- *T1_LI_OWFH* (UA-DLSI-*lang*-3): generic language-independent summarizer considering only the words included in the headings of the documents.
- *T1_LEXSEM_AW* (UA-DLSI-*lang*-4): generic summarizer, including lexical-semantic knowledge into the interpretation stage, and considering all words in the documents.
- *T3_LI_OWFH* (UA-DLSI-*lang*-5): topic-focused language-independent summarizer considering only the words included in the headings of the documents.
- *T3_LEXSEM_AW* (UA-DLSI-*lang*-6): topic-focused summarizer, including lexical-semantic knowledge into the interpretation stage, and considering all words in the documents.
- *T3_LEXSEM_OWFH* (UA-DLSI-*lang*-2): topic-focused summarizer, including lexical-semantic knowledge into the interpretation stage, but considering only the words included in the headings of the documents.

5 Results and Analysis

After all participants submitted their runs to the MultiLing 2015 MSS task over the test dataset, the summaries were evaluated via automatic methods. ROUGE tool (Lin, 2004) was employed for automatic content evaluation, which allows the comparison between automatic and model summaries based on different types of n-grams. Specifically the ROUGE 1 (unigrams), 2 (bigrams), 3 (trigrams), and 4 (quadrigrams), ROUGE-SU4 (bigram similarity skipping unigrams) scores were computed. The files contain the overall and individual summary scores.

Moreover, two additional systems were proposed by the organizers. On the one hand, a system called *"Lead"*, which was the baseline summary used for the evaluation process. This approach selects the leading substring of the article's body text having the same length as the human summary of the article. On the other hand, a system called *"Oracles"* was also developed, where sentences were selected from the body text to maximally cover the tokens in the human summary using as few sentences as possible until its size exceeded the human summary, upon which it was truncated.

In this edition, five systems participated in the *MSS* task (details about their implementation have not made available yet). Three of them were applied to 38 languages, including English, Spanish and German. They are named as *CCS* - that implements five variations for each language- *LCS-IESI* and *EXB*. The fourth one, *BGU-SCE* has been proven for Arabic and Hebrew, besides English.

Table 2, Table 3, and Table 4 show the results obtained by all participants, and the two methods proposed by the organizers in the MultiLing 2015 competition for English, German, and Spanish. Due to size constraints, only the average results for the recall, precision and F-measure metrics of ROUGE 1 are shown, since this ROUGE metric takes into account the common vocabulary between the automatic and the human summaries, without taking into account stopwords.

Focusing only on the analysis of our six versions of our approach (UA-DLIS-*lang-priority*),

System	R1 recall	R1 precision	R1 F-measure
UA-DLSI-en-1	0.45488	0.45827	0.45605
UA-DLSI-en-2	0.42111	0.43774	0.42703
UA-DLSI-en-3	0.37175	0.49104	0.40551
UA-DLSI-en-4	0.45641	0.45673	0.45627
UA-DLSI-en-5	0.41994	0.43334	0.42419
UA-DLSI-en-6	0.42439	0.43093	0.42727
BGU-SCE-M-en-1	0.49195	0.48354	0.48744
BGU-SCE-M-en-2	0.47826	0.47953	0.47868
BGU-SCE-M-en-3	0.45955	0.46053	0.45974
BGU-SCE-M-en-4	0.46819	0.46651	0.46713
BGU-SCE-M-en-5	0.49982	0.48813	0.49361
BGU-SCE-P-en-1	0.46247	0.44367	0.45269
BGU-SCE-P-en-2	0.49420	0.47512	0.48425
BGU-SCE-P-en-3	0.46546	0.45039	0.45753
CCS-en-1	0.49507	0.47662	0.48539
CCS-en-2	0.49041	0.47299	0.48132
CCS-en-3	0.49130	0.47455	0.48255
CCS-en-4	0.48849	0.47211	0.47986
CCS-en-5	0.48689	0.47600	0.48117
EXB-en-1	0.49471	0.46692	0.48022
LCS-IESI-en-1	0.45556	0.46144	0.45811
NTNU-en-1	0.45585	0.46966	0.46213
Lead-en-1	0.43381	0.42495	0.42907
Oracles-en-1	0.61917	0.60114	0.60983

Table 2: Average results for English (recall, precision and F-measure ROUGE 1 (R1) values.

we observe that our approach with priority 3 is one of our best performing approaches considering the precision for the three tested languages. This version corresponds to *T1_LI_OWFH* approach - generic language-independent summarizer considering only the words included in the headings of the documents, and this means that the title headings of the Wikipedia articles do contain enough meaningful information of the documents. This is an interesting finding, because we are reducing the amount of information to be processed by almost 99%. Moreover, this also outlines the potential of the studied PCA technique for developing completely language-independent summarizers.

Other versions of our proposed approach, such as the ones submitted as priority 4, and priority 1 may obtained also competitive results for some languages. Again, the submission with priority 1 correspond to a generic language-independent summarizer considering all words in the documents (*T1_LI_AW*). It can be shown that when considering all words in the documents, instead of only the words in the headings, recall values improve, but for some languages, e.g. German, to take into account all the words does not have a positive influence in general. Regarding the submission with priority 4 (*T1_LEXSEM_AW*), the inclusion of lexical-semantic knowledge has been beneficial for the English results, but not for the other languages. This may be due to the type of semantic knowledge that is being used. WordNet for English is much bigger in size than for German and Spanish, and therefore, this could influence the results, not obtaining the expected improvements that were expected by using language-dependent resources. Generally speaking, from our approaches, apart from the previously mentioned findings, we can also observe that when summarizing Wikipedia articles, generic summarization has been shown to be more appropriate.

Analyzing all the results achieved by the other participants, we can observe that German is the language, among the three analyzed languages within our scope, that obtains poorer ROUGE results. This could occur since the summaries had a compression ratio lower than 3%, which is a

System	R1 recall	R1 precision	R1 F-measure
UA-DLSI-de-1	0.33993	0.34401	0.34110
UA-DLSI-de-2	0.33207	0.34331	0.33725
UA-DLSI-de-3	0.36126	0.36448	0.36236
UA-DLSI-de-4	0.33492	0.35565	0.34317
UA-DLSI-de-5	0.33023	0.33927	0.33437
UA-DLSI-de-6	0.34401	0.34807	0.34553
CCS-de-1	0.40140	0.36441	0.38163
CCS-de-2	0.40025	0.36601	0.38203
CCS-de-3	0.40257	0.37118	0.38575
CCS-de-4	0.40587	0.37234	0.38803
CCS-de-5	0.39356	0.38055	0.38665
EXB-de-1	0.37909	0.35621	0.36692
LCS-IESI-de-1	0.34844	0.36285	0.35504
Lead-de-1	0.33010	0.31562	0.32230
Oracles-de-1	0.54342	0.51331	0.52759

Table 3: Average results for German (recall, precision and F-measure ROUGE 1 (R1) values.

very low compression ratio for the summarization task. Moreover, it can be seen from the tables, that all systems overperformed the *"Lead"* baseline, but none of them surpassed the *"Oracles"* system. This was expected since the *"Oracles"* system was kind of upper boundary for the MSS task. Among the systems, the best performing ones taking into account the ROUGE 1 F-measure value were: the *BGU-SCE* team with their submission *BGU-SCE-M-en-5* for English; *CCS* team, with *CCS-de-4* for German; and again *CCS* team with *CCS-es-3* for Spanish. Taking into account the different submissions, our versions were not among the best performing approaches, despite obtaining results in line of the other participants. In general, there were not very big differences in results between the teams. In this sense, according to ROUGE 1 F-measure, we ranked[7] 15th out of 22nd for English with our *UA-DLSI-en-4* submission; 7th out of 13th for German with our *UA-DLSI-de-3* submission; and 8th out of 13th with our *UA-DLSI-es-1* submission. As it was previously discussed, for German and Spanish, the best submissions were the ones without using any type of lexical-semantic knowledge, whereas for English the use of a named entity recognizer, and a semantic knowledge base led to an improvement over the language-independent approach.

6 Potentials and Limitations of the UA-DLSI Approach

From our participation in MultiLing 2015, we have tested our approach in a real competition and compared its performance with respect to state-of-the-art multilingual summarizers. Although in general terms, the best versions of our approach ranked at intermediate positions, the participation and evaluation process has been a positive issue for learning from errors, as well as gaining some insights into potentials and limitations that our approach and in general the multilingual summarization task may have.

After analyzing the performance of the different system configurations, it becomes clear that some of our assumptions need to be reviewed. Nevertheless, good positions were achieved when reducing the words to compute the PCA algorithm, which let us infer that article section headings contain enough information to produce accurate and precise summaries, while decreasing the amount of information to be processed by the system. Moreover, our results indicate that using PCA present advantages when language independent processing is required.

On the other hand, the limitations encountered are mostly related to inclusion of lexical-semantic knowledge. As it requires the use of external resources, the system performance becomes dependent of some aspects such as their quality, availability and size. The version of the system tak-

[7] The two systems provided by the organization has not been taken into account for the ranking.

System	R1 recall	R1 precision	R1 F-measure
UA-DLSI-es-1	0.48273	0.49799	0.48977
UA-DLSI-es-2	0.46191	0.48250	0.47141
UA-DLSI-es-3	0.45203	0.50965	0.46979
UA-DLSI-es-4	0.47795	0.49211	0.48454
UA-DLSI-es-5	0.46748	0.48820	0.47691
UA-DLSI-es-6	0.46657	0.47827	0.47193
CCS-es-1	0.52817	0.50834	0.51783
CCS-es-2	0.53135	0.51065	0.52057
CCS-es-3	0.52430	0.50440	0.51388
CCS-es-4	0.53234	0.51121	0.52126
CCS-es-5	0.52410	0.51321	0.51835
EXB-es-1	0.53018	0.49760	0.51310
LCS-IESI-es-1	0.50057	0.50575	0.50213
Lead-es-1	0.46826	0.46419	0.46599
Oracles-es-1	0.62557	0.60875	0.61691

Table 4: Average results for Spanish (recall, precision and F-measure ROUGE 1 (R1) values.

ing into account this kind of background obtains better results in English language, for which resources as WordNet have reached a state of maturity higher than for other languages. In addition, and regarding the format of the source documents (Wikipedia articles), topic-focused summaries have been shown to be less adequate than generic summarization.

Concerning the multilingual summarization task from a broader perspective, it is worth stressing that this is a challenging task. On the one hand, language-independent methods exist, and they offer more capabilities to be employed for a wide range of languages; however, this type of techniques do not take into account any semantic analysis, so it is difficult that only with these techniques, abstractive summaries can be produced, thus limiting mostly to extractive summarization.

In the context of the MSS task, the summary compression ratio was extremely low, compared to the length of the original documents. This posed the task even more challenging, since the generated summaries had to be very concise as well as precise. Nevertheless, it is of great value to organize this type of events and have the possibility to participate in order to advance the state of the art, addressing difficult summarization challenges necessary in the current society.

7 Conclusions

In this paper we described our participation in MultiLing 2015 - Multilingual Single-document Summarization task, presenting our approach and comparing and discussing the results obtained with respect to the other participants in the task.

Our initial development was focused on the application of the PCA technique, given its suitability for developing language-independent approaches. Although some related work has been done on summarization, we contributed to the state of the art extending the PCA scope by the inclusion of lexical and semantic knowledge in its implementation and testing it in a multilingual scenario.

Our approach was tested in three languages, English, German, and Spanish, and six different configurations were submitted to the competition, obtaining average results when compared to other participants.

From our participation in MultiLing 2015, and the further analysis of our PCA based approach given the results obtained, three main conclusions can be drawn: i) PCA is a good technique for generating language-independent summaries; ii) generic summaries were more appropriate for the type of documents dealt with (i.e., Wikipedia documents); and iii) the title headings of Wikipedia articles were meaningful enough to build the PCA matrix in the summarization process, discarding the remaining words of the document. Although this version of our approach worked with very few content, it was shown to be one of our best performing approaches.

Acknowledgments

This research work has been partially funded by the University of Alicante, Generalitat Valenciana, Spanish Government and the European Commission through the projects, “Tratamiento inteligente de la información para la ayuda a la toma de decisiones” (GRE12-44), “Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario” (GRE13-15), DIIM2.0 (PROMETEOII/2014/001), ATTOS (TIN2012-38536-C03-03), LEGOLANG-UAGE (TIN2012-31224), and SAM(FP7-611312).

References

Daniel Anechitei and Eugen Ignat, 2013. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, chapter Multilingual summarization system based on analyzing the discourse structure at MultiLing 2013, pages 72–76. Association for Computational Linguistics.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

John M. Conroy, Judith D. Schlesinger, and Jeff Kubina. 2011. CLASSY 2011 at TAC: Guided and Multi-lingual Summaries and Evaluation Metrics. In *Proceedings of the Text Analysis Conference (TAC 2011).*

John Conroy, T. Sashka Davis, Jeff Kubina, Yi-Kai Liu, P. Dianne O’Leary, and D. Judith Schlesinger, 2013. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, chapter Multilingual Summarization: Dimensionality Reduction and a Step Towards Optimal Term Coverage, pages 55–63. Association for Computational Linguistics.

Mahmoud El-Haj and Paul Rayson, 2013. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, chapter Using a Keyness Metric for Single and Multi Document Summarisation, pages 64–71. Association for Computational Linguistics.

Enrique Estellés Arolas, Fernando González Ladrón De Guevara, and Antonio Falcó Montesinos. 2010. Principal Component Analysis for Automatic Tag Suggestion. Technical report.

George Giannakopoulos, Mahmoud El-Haj, Benoit Favre, Litvak Marina, Josef Steinberger, and Vaduseva Varma. 2011. TAC2011 MultiLing Pilot Overview. In *Proceedings of the Text Analysis Conference (TAC 2011).*

George Giannakopoulos. 2013. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jeff Kubina, John M Conroy, and Judith D Schlesinger. 2013. Acl 2013 multiling pilot overview. *Proceedings of MultiLing 2013 Workshop on Multilingual Multi-document Summarization, Sofia, Bulgaria*, pages 29–38.

Chang Beom Lee, Min Soo Kim, and Hyuk Ro Park. 2003. Automatic Summarization Based on Principal Component Analysis. *Progress in Artificial Intelligence*, pages 409–413.

Chang B. Lee, Hyukro Park, and Cheolyoung Ock. 2005. Significant Sentence Extraction by Euclidean Distance Based on Singular Value Decomposition. In *Proceedings of the Natural Language Processing-IJCNLP 2005*, pages 636–645.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Marie-Francine Moens, S. S., editor, Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Marina Litvak and Mark Last. 2013. Multilingual single-document summarization with muse. *MultiLing 2013*, page 77.

Elena Lloret and Manuel Palomar. 2011. Finding the Best Approach for Multi-lingual Text Summarisation: A Comparative Analysis. In *International Conference Recent Advances in Natural Language Processing*, pages 194–201.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Karen Sparck-Jones. 1999. Automatic summarising : factors and directions. *Advances in automatic text summarisation*, pages 1–21.

Josef Steinberger, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Marco Turchi, and Vanni Zaravella. 2011. JRC’s Participation at TAC 2011: Guided and MultiLingual Summarization Tasks. In *Proceedings of the Text Analysis Conference (TAC 2011).*

Josef Steinberger. 2013. The uwb summariser at multiling-2013. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 50–54, Sofia, Bulgaria, August. Association for Computational Linguistics.

Erik Tjong, Kim Sang, and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *7th conference on Natural language learning at HLT-NAACL 2003*, volume 4, pages 142–147.

Om Vikas, Akhil K Meshram, Girraj Meena, and Amit Gupta. 2008. Multiple Document Summarization Using Principal Component Analysis Incorporating Semantic Vector Space Model. *Computational Linguistics and Chinese Language Processing*, 13(2):141–156.

Piek Vossen. 2004. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography Vol.17*, 2:161–173.