Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües

Studying the influence of adding lexical-semantic knowledge to Principal Component Analysis technique for multilingual summarization

Óscar Alcón
Universidad de Alicante
oalcon@dlsi.ua.es

Elena Lloret
Universidad de Alicante
elloret@dlsi.ua.es

Resumen

El objetivo de la generación automática de resúmenes es reducir la dimensión de un texto y a su vez mantener la información relevante del mismo. En este artículo se analiza y aplica la técnica de Análisis de Componentes Principales, que es independiente del idioma, para la generación de resúmenes extractivos mono-documento y multilingües. Dicha técnica se estudiará con el objetivo de poder evaluar su funcionamiento cuando se incorpora (o no) conocimiento léxico-semántico, a partir del uso de recursos y herramientas dependientes del idioma. La experimentación propuesta se ha realizado en base a dos corpus de diferente naturaleza: noticias periodísticas y artículos de la Wikipedia en tres idiomas (alemán, español e inglés) para verificar el uso de esta técnica en varios escenarios. Los enfoques propuestos presentan resultados muy competitivos comparados con generadores de resúmenes multilingües existentes, lo que indica que, aunque exista un claro margen de mejora respecto a la técnica y el tipo de conocimiento incorporado, ésta tiene una gran potencial para ser aplicada en otros contextos e idiomas.

Palabras clave

PCA, Análisis de Componentes Principales, generación de resúmenes, multilingües, extractivos, entidades nombradas, identificación de conceptos

Abstract

The objective of automatic text summarization is to reduce the dimension of a text keeping the relevant information. In this paper we analyse and apply the language-independent Principal Component Analysis technique for generating extractive single-document multilingual summaries. This technique will be studied to evaluate its performance with and without adding lexical-semantic knowledge through language-dependent resources and tools. Experiments were conducted using two different corpora: newswire and Wikipedia articles in three languages (English, German and Spanish) to validate the use of this technique in several scenarios. The proposed approaches show very competitive results compared to multilingual available systems, indicating that, although there is still room for improvement with respect to the technique and the type of knowledge to be taken into consideration, this has great potential for being applied in other contexts and for other languages.

Keywords

PCA, Principal Component Analysis, automatic summarization, multilingüal summarization, extractive summarization, NER, concept identification

1 Introducción

Actualmente, el tratamiento y gestión de la información es una tarea difícil de abordar para el ser humano. En un contexto donde cada vez la cantidad de información y la heterogeneidad de la misma aumentan a un ritmo considerable, cobran una mayor importancia las técnicas automáticas de análisis y reducción de volumen para la detección y extracción de la información relevante.

Además, se dispone de una gran cantidad de información por lo que se hace necesario el desarrollo de técnicas de Procesamiento de Lenguaje Natural (PLN) para poder procesar, clasificar, extraer y resumir la información del texto. Respecto a la tarea de generación de resúmenes, cuyo objetivo es obtener una versión reducida del documento o documentos fuentes, reduciendo su contenido pero sin perder información clave (Spärck Jones, 2007), no siempre resulta sencillo determinar qué información es la más rele-

vante, debido a la variedad de factores que podemos tener en cuenta (por ejemplo, preferencias del usuario, necesidades de información, finalidad del resumen, etc.). Así, se han establecido diferentes tipologías de resúmenes (Mani & Maybury, 1999; Spärck Jones, 2007). Entre los tipos más comunes se encuentra la distinción entre resúmenes *mono-documento* (el resumen se genera a partir de un único documento de entrada) vs. *multi-documento* (varios documentos de entrada); *extractivos* (el resumen simplemente se limita a realizar una selección de las frases más relevantes) vs. *abstractivos* (el resumen contiene información expresada de distinta manera con respecto al documento fuente); *genéricos* vs. *centrados en un tema concreto*; así como también *monolingües* vs. *multilingües*, si el resumidor funciona únicamente para un idioma o para varios.

Para abordar la generación automática de los distintos tipos de resúmenes anteriormente comentados se han utilizado distintas técnicas (Nenkova & McKeown, 2011): desde técnicas superficiales que determinan la relevancia de información según el peso de las unidades que forman las frases, como por ejemplo la frecuencia de palabras y aproximaciones derivadas (McCargar, 2005), hasta enfoques que se basan en el uso de técnicas de análisis del discurso, que implican un procesamiento más profundo del texto. En este último caso encontramos como ejemplo, la técnica de las cadenas léxicas (Barzilay & Elhadad, 1999), o la técnica de la estructura retórica del discurso (RST) (Uzêda et al., 2010).

En el contexto actual, donde no solamente hay grandes cantidades de información y el ritmo de crecimiento de la misma es exponencial, sino que dicha información está disponible en una gran variedad de idiomas, es necesario investigar en técnicas de generación de resúmenes multilingües que consigan determinar la información clave sea cuál sea el idioma en el que se haya escrito. Para ello, es necesario o bien recurrir a técnicas totalmente independientes del idioma como la frecuencia de términos (Teng et al., 2008), o bien que la técnica o el conocimiento aplicado estén disponibles para varios idiomas. Una técnica independiente del idioma es el Análisis de Componentes Principales (Principal Component Analysis, PCA), que se puede aplicar a la detección y extracción de palabras clave en un texto. Dada la naturaleza de la misma, esta técnica puede ser adecuada para la generación de resúmenes multilingües, y por tanto, será la que estudiaremos en este trabajo.

Por consiguiente, el principal objetivo de este articulo es analizar la técnica PCA para la generación de resúmenes extractivos monodocumento y multilingües. Esta técnica se analizará y evaluará por un lado de forma independiente (técnica base), y por otro con el enriquecimiento mediante la incorporación de conocimiento léxico-semántico, obtenido a partir del reconocimiento de entidades nombradas (Named Entity Recognition, NER) y la identificación de conceptos sinónimos, con el fin de medir la influencia de estas técnicas sobre la técnica base, y determinar si pueden ser beneficiosas en el proceso de generación de resúmenes multilingües diseñado. Además, una vez obtenidas las palabras clave a partir del uso de la técnica PCA en sus diversas variantes, se proponen y analizan cuatro heurísticas para la selección de las frases relevantes del documento fuente, dando lugar a distintos tipos de resúmenes extractivos.

El artículo se estructura del siguiente modo: la sección 2 recoge los trabajos realizados hasta el momento acerca de resúmenes multilingües y del uso de la técnica PCA para resumir textos. En la sección 3 se explica el método implementado para la generación de resúmenes mono-documento y multilingües con PCA. La sección 4 alberga la información de los corpus empleados para la experimentación y de las medidas de evaluación que se utilizarán en la sección 5. En la sección 6 se recogen y analizan los resultados que serán comparados con sistemas previos en la sección 7. Finalmente, se exponen las conclusiones obtenidas en la sección 8.

2 Estado de la cuestión

La generación de resúmenes multilingües es una tarea aún en desarrollo dada la dificultad de poder abarcar las características particulares de cada idioma.

En (Gupta & Lehal, 2010) se recoge una serie de enfoques en relación a la tarea de resúmenes multilingües entre los que se incluye (Cowie et al., 1998), quienes parece que iniciaron la investigación en esta temática cuando presentaron MINDS, un sistema que incluye soporte para la creación de resúmenes de documentos en inglés, español, ruso y japonés. El núcleo del sistema se generó empleando técnicas como análisis estadístico, sintáctico y de estructura de documentos. Más tarde, Hovy y Lin (1997) se adentraron en la materia con SUMMARIST, un sistema que permite la generación de resúmenes tanto abstractivos como extractivos en distintos idiomas, empleando técnicas de procesamiento del lenguaje natural, junto con bases de conocimiento. En (Patel et al., 2007) se propuso también

un método, independiente del idioma, para resumir textos de distintos idiomas. Estaba basado en factores estadísticos y de estructura del documento, tales como posición en el texto o identificación de nombres comunes y propios. No obstante, utilizaba el lexema de las palabras y filtraba el documento para eliminar las palabras carecientes de contenido léxico-semántico (stopwords). El sistema se testó con documentos en inglés, hindi, gujarati y urdu. En (Lloret & Palomar, 2011) se analizaron tres enfoques distintos, empleando: 1) técnicas independientes del idioma; 2) técnicas específicas de cada idioma; y 3) aplicando traducción automática a resúmenes monolingües. Dichos enfoques se orientaron a la producción de resúmenes extractivos en cuatro idiomas diferentes (español, inglés, alemán y frances).

La competición bienal MultiLing[1] se creó en 2011 con motivo de fomentar el trabajo sobre la generación de resúmenes multilingües. Se presentan enfoques de distintos equipos de investigación para mostrar el estado del arte en la materia y enfocar las investigaciones futuras (Giannakopoulos et al., 2011; Kubina et al., 2013). En uno de los trabajos de esta competición se presentó el sistema MUSE (Litvak & Last, 2013). Para el desarrollo de este sistema emplearon un algoritmo genético para la optimización lineal de diversas medidas de clasificación de frases. Otro enfoque fue presentado en (Conroy et al., 2013), donde se describía el uso del Análisis de la Semántica Latente (Latent Semantic Analysys, LSA) para la generación de resúmenes multi-documento para 10 idiomas distintos. Cabe destacar como en la última edición de MultiLing (Multiling 2013), algunos de los sistemas participantes alcanzaron unos resultados similares a los obtenidos con resúmenes manuales (Giannakopoulos, 2013). Concretamente, el sistema WBU (Steinberger, 2013) fue el que mejor resultados consiguió, basándose en la técnica LSA. Este sistema fue probado en 10 idiomas, y en los que en 5 de ellos quedó en primera posición.

La técnica PCA, al igual que la técnica LSA, se encuentra englobada dentro de las técnicas de minería de datos, pero se diferencian en la manera de calcular la matriz, teniendo menos dispersión en el caso de la técnica PCA. Esta técnica se ha utilizado con anterioridad para la tarea de resumir texto en Lee, Kim y Park (2003), donde se propuso un sistema para la extracción de frases de un texto que representen la información relevante, y se empleó la técnica PCA para extraer las palabras clave del documento, seleccionando las frases según la cantidad de palabras clave que incluían, siendo la más relevante aquella que albergara mayor cantidad de palabras clave. Obtuvieron buenos resultados (medida F = 0.416) para textos en coreano. Vikas et al., (2008) desarrollaron otro enfoque en donde se expone un sistema para resumir textos mono-documento y multi-documento, utilizando un Modelo Espacial de Vectores Semánticos (Semantic Vector Space Model, SVSM) para modelar el conjunto de documentos. La técnica PCA se empleó para extraer características referentes al tema del documento sobre un conjunto de textos de distinta temática en inglés. Más recientemente, la técnica PCA se ha aplicado con éxito a la generación automática de *hashtags*, en la que se logran unos resultados cercanos al 60 % (Estellés Arolas et al., 2010). Los *hashtags* de un tuit se pueden equiparar a las palabras clave que pueden resumir el texto expresado en un tuit, y por tanto, se demuestra la utilidad de la técnica PCA para los nuevos géneros textuales surgidos con la Web 2.0.

Revisados los trabajos previos en éste ámbito, la técnica PCA no ha sido investigada con anterioridad para generar resúmenes multilingües, a pesar de ser una técnica independiente del idioma, ni tampoco se ha analizado la influencia de incorporar información léxico-semántica al proceso. En base a esto, nuestra contribución en este artículo es el estudio de dicha técnica para la producción de resúmenes extractivos monodocumento y multilingües, así como el análisis de la influencia de incorporar conocimiento léxico-semántico, dependiente del idioma, a la técnica base. En nuestra propuesta, la técnica PCA se utilizará para extraer las palabras clave de los textos, que serán posteriormente empleadas para escoger las frases mas relevantes que conformarán el cuerpo del resumen final.

3 Aplicación de la técnica PCA para la generación de resúmenes multilingües

Los procesos de generación de resúmenes automáticos se han caracterizado por seguir un flujo genérico que engloba tres fases claramente diferenciadas (Sparck-Jones, 1999): i) interpretación; ii) transformación; y iii) generación de resúmenes.

Partiendo de esa base se ha formulado una propuesta de método para la generación de resúmenes aplicando la técnica PCA para textos multilingües, reflejado en la Figura 1. Como se puede observar, la fase de interpretación (sección 3.1) será en la cual se realice un preprocesado para obtener la información de interés y prepararla para la siguiente fase. En la fase de trans-

[1] http://multiling.iit.demokritos.gr/

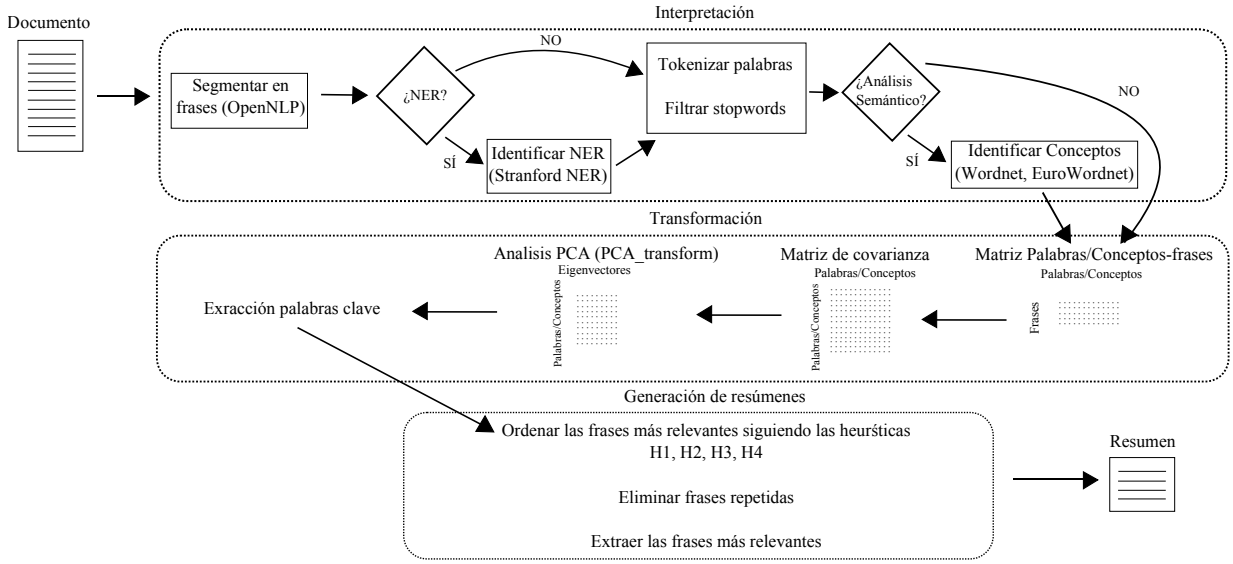

Figura 1: Flujo de acción de nuestro método de generación de resúmenes basado en la técnica PCA.

formación (sección 3.2) será en donde se aplique la técnica PCA para obtener las palabras clave del documento mediante el procesamiento de la información del texto. Finalmente, en la fase de generación de resúmenes (sección 3.3) se definen una serie de heurísticas para la selección y extracción de las frases más relevantes y formar el resumen final.

3.1 Interpretación

El método desarrollado tiene como entrada un texto al cual se le aplica el siguiente preprocesamiento lingüístico: i) segmentación en frases (OpenNLP[2]); ii) tokenización; iii) eliminación de palabras carecientes de contenido semántico (stopwords). Se incluye la opción de añadir conocimiento léxico-semántico mediante la identificación de entidades nombradas, explicado en la sección 3.1.1, y/o la identificación de conceptos, posteriormente explicado en la sección 3.1.2.

3.1.1 Reconocimiento de Entidades Nombradas

El Reconocimento de Entidades Nombradas (Named Entity Recognition, NER) consiste en etiquetar el texto de entrada para reconocer secuencias de palabras que sean nombres de personas, organizaciones y lugares, llamadas entidades (Tjong et al., 2003).

Para nuestro enfoque utilizamos como reconocedor de entidades la herramienta *Stanford Named Entity Recognizer*[3] que funciona para varios idiomas (español, inglés y alemán, entre otros).

[2] https://opennlp.apache.org/

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

3.1.2 Identificación de conceptos

La identificación de conceptos en nuestro enfoque se basa en la detección de sinónimos, considerando cada conjunto de sinónimos como un único concepto y agrupando sus apariciones a lo largo del texto. Para ello, se ha utilizado *WordNet* (Miller, 1995) y *EuroWordnet* (Vossen, 2004) ya que estos recursos recogen conocimiento léxico-semántico para distintos idiomas - *Wordnet* para inglés, y *EuroWordnet* para un conjunto de idiomas europeos (entre ellos, el español y el alemán) - agrupando las palabras por conjuntos de sinónimos y almacenando las relaciones entre los mismos. La razón por la que se utilizó EuroWordnet frente a recursos que pudieran estar más actualizados para cada idioma, como Mutlilingual Central Repository[4] para el castellano o el GermaNet[5] fue para validar en una primera versión de la investigación realizada si el uso de este tipo de conocimiento integrado en la técnica PCA era apropiado o no.

Para identificar los conceptos, nos basamos en el sentido más frecuente como algoritmo de desambiguación[6], puesto que la relación resultados-coste computacional es aceptable en el estado de la cuestión (los resultados de esta aproximación están alrededor del 50 % (McCarthy, 2011)). Por tanto, en esta fase, se utiliza esta aproximación para buscar el primer synset de cada palabra en el documento. *Wordnet* y *EuroWordnet* estiman como primer synset de cada

[4] http://adimen.si.ehu.es/web/MCR

[5] http://www.sfs.uni-tuebingen.de/GermaNet/index.shtml

[6] La tarea de desambiguación del sentido de las palabras no es objeto de este artículo.

palabra el significado más frecuente de la misma y por tanto, el más probable. Si dos palabras tienen el mismo primer synset, serán consideradas como sinónimos y sus apariciones en el texto serán sumadas. Por ejemplo, los términos *"devastar"* y *"destrozar"*, aunque distintos en su morfología, se considerarían englobados en el mismo concepto, puesto que el primer synset de ambas palabras es el *"00260311"*.

3.2 Transformación: la técnica PCA para la detección de palabras clave

la técnica PCA se basa en un algoritmo matemático que reduce la dimensionalidad de los datos conservando la mayor parte de la variación en el conjunto de datos (Ringnér, 2008). Ante un gran volumen de datos con distintas variables, el objetivo de este algoritmo es encontrar una serie de patrones o tendencias dentro del conjunto de datos de entrada y transformar linealmente el conjunto de variables original en un conjunto considerablemente menor de variables incorreladas (Dunteman, 1989).

3.2.1 Generación de la matriz de componentes principales

A partir de la información obtenida en la etapa anterior, en primer lugar se genera la matriz de palabras/conceptos-frases, en la que para cada palabra/concepto se recoge el número de ocurrencias de esa palabra/concepto en el texto. A partir de esta matriz, se genera la matriz de covarianza para descifrar las relaciones existentes entre las palabras/conceptos del texto. Esta matriz será utilizada para obtener las componentes principales (eigenvectores) y su correspondiente valor propio (eigenvalue) mediante la aplicación de la técnica PCA, utilizando la librería de Java PCA_transform[7].

La aplicación de la técnica PCA devuelve una matriz en la cual las columnas son los eigenvectores (ordenados en orden descendiente determinado por el eigenvalue asociado) y las filas serían las variables (que este caso serían las palabras/conceptos del texto). Cada eigenvector se conforma con la contribución de cada variable, que determina la importancia de dicha variable en el eigenvector. De cada eigenvector extraemos la palabra(s) o concepto(s) que presente mayor contribución, considerándolas como palabras clave del texto, que posteriormente serán empleadas con el fin de seleccionar las frases más relevantes, explicado en la sección 3.3.

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

3.3 Generación de resúmenes

En esta fase se define la estrategia para escoger las frases en función de los valores obtenidos de la técnica PCA. Con las palabras clave extraídas, se plantean diferentes heurísticas para la extracción de las frases más relevantes del texto para la realización del resumen automático:

H1: Se selecciona, por orden de aparición en el texto, la frase que contiene el término extraído del eigenvector, realizándose este proceso para todas las palabras clave determinadas por la técnica PCA. En el caso de que la frase ya esté seleccionada, se escogería la siguiente frase donde aparece.

H2: Se selecciona, por orden de aparición en el texto, sólo la primera frase que contiene el término extraído del eigenvector, prosiguiendo para todas las palabras clave. En el caso de que la frase ya esté seleccionada se pasaría al siguiente término. Aunque esta heurística es similar a la anterior, la principal diferencia entre ambas radica en el número de frases que se pueden incluir para cada término extraído y el tratamiento de las frases que ya han sido incluidas anteriormente para formar parte del resumen. Mientras que en H1 se puede seleccionar más de una frase que contenga el término, si la frase que se va a seleccionar ya se ha incluido en el resumen anteriormente, según la estrategia definida para H2, sólo seleccionamos la primera frase que contiene el término y cuando nos encontramos con una frase que ya ha sido incluida, se finaliza con ese término y se pasa al siguiente para seguir con el proceso de selección de frases.

H3: Se seleccionan todas las frases donde aparecen las palabras clave extraídas, siguiendo el orden determinado por la técnica PCA.

H4: Se buscan las frases en las que aparecen las palabras clave y se escogen aquellas frases en las que haya incluidos al menos dos términos. Serán ordenadas en el resumen según la importancia de dichas palabras clave incluidas.

Las frases seleccionadas para cada heurística serán las conformantes del resumen final, eliminando previamente las frases repetidas. Cabe mencionar que si existen dos o más palabras con el mismo valor máximo dentro de un mismo eigenvector, se extraería las frases correspondientes para cada palabra. Del mismo modo y en el caso de utilizarse conocimiento léxico-semántico, cuando un concepto está representado por varios

sinónimos, se extraería las frases correspondientes para cada sinónimo.

4 Entorno de evaluación

En esta sección se explican los corpus que serán utilizados para testar el funcionamiento del sistema (sección 4.1). Además, los resúmenes generados serán evaluados mediante las medidas determinadas en la sección 4.2.

4.1 Corpus

Para probar el funcionamiento de la técnica desarrollada se han empleado dos corpus multilingües de distinta naturaleza: corpus JRC y corpus de entrenamiento de MultiLing 2015 y concretamente, nos vamos a centrar en 3 idiomas: inglés, español y alemán.

4.1.1 Corpus JRC

El corpus JRC[8] dispone de un conjunto de noticias periodísticas en 7 idiomas, donde para cada idioma hay 20 documentos agrupados en 4 temas: genética; conflicto Israel-Palestina; malaria; ciencia y sociedad. Así mismo, proporciona un conjunto de resúmenes humanos que se emplearán como modelos para la evaluación, contando con 4 resúmenes modelo extractivos para cada uno de los documentos. Los documentos en inglés, español y alemán tienen de media 820, 927 y 836 palabras, respectivamente.

4.1.2 Corpus de entrenamiento de MultiLing 2015

El corpus de entrenamiento de MultiLing 2015[9] está formado por artículos extraídos de la Wikipedia, donde para cada idioma hay 30 documentos. Se cuenta también con un resumen modelo abstractivo para cada documento, que puede ser empleado para realizar una evaluación automática. Los documentos en inglés, español y alemán tienen de media 3973, 6311 y 4248 palabras, respectivamente.

4.2 Medidas de evaluación

La evaluación de los resúmenes se realiza de forma cuantitativa, centrándonos exclusivamente en el contenido de los resumenes generados, y para ellos utilizamos la herramienta ROUGE (Lin, 2004), por ser una de las más utilizadas en este campo. Esta herramienta permite la evaluación automática de resúmenes mediante la comparación del número de n-gramas coincidentes de dichos resúmenes con respecto a unos resúmenes modelo. Partiendo de esta premisa, ROUGE implementa diferentes métricas, teniendo en cuenta: la similitud de unigramas (ROUGE-1); la similitud de bigramas (ROUGE-2); la secuencia común más larga (ROUGE-L) y la similitud de bigramas evitando unigramas (ROUGE-SU4). Además, para cada uno de los indicadores antes mencionados, ROUGE devuelve las siguientes medidas: Precisión, Recall y medida F.

5 Experimentación

En esta sección se describen los experimentos realizados sobre el método propuesto para comprobar y analizar su funcionamiento. Nuestro objetivo es determinar la idoneidad de la técnica PCA aplicada a resúmenes multilingües y analizar la influencia de la incorporación de conocimiento léxico-semántico de forma gradual. La experimentación se va a realizar para tres idiomas (inglés, español y alemán), puesto que existen recursos que nos permiten obtener el tipo de información léxico-semántica que necesitamos.

Para cada una de las cuatro heurísticas formuladas, se plantean los siguientes enfoques para la incorporación de conocimiento léxico-semántico de forma gradual:

(a) *PCA_base*: no se utiliza conocimiento léxico-semántico y se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras del documento (excepto stopwords).

(b) *PCA_base+CPT*: se incorpora al método base la identificación de conceptos, de tal manera que se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras y los conceptos (excepto stopwords).

(c) *PCA_base+CPT+NER*: se enriquece el método anterior con un proceso de NER, y por tanto, se incluyen en la matriz obtenida a partir de la técnica PCA todas las palabras, los conceptos y las NER identificadas (excepto stopwords).

6 Resultados y discusión

Los resultados de los experimentos realizados se muestran en la Tabla 1 y la Tabla 2, correspon-

[8] http://optima.jrc.it/Resources/2010_JRC_multilingual-summary-evaluation.zip.

[9] http://users.iit.demokritos.gr/~ggianna/MultiLing2015/multilingMss2015Training.tar.gz

		(a) PCA_base				(b) PCA_base+CPT				(c) PCA_base+CPT+NER			
		R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4
Inglés	H1	0.54200	0.33613	0.51832	0.36249	**0.54576**	**0.33841**	**0.52220**	**0.36528**	0.56159	0.36577	0.53736	0.39008
	H2	0.53714	0.32581	0.51184	0.35312	0.54101	0.32523	0.51509	0.35496	0.54760	0.34377	0.52402	0.36977
	H3	**0.57797**	0.39244	**0.55502**	0.41626	0.53348	0.33485	0.50800	0.36249	0.56370	0.37574	0.53888	0.40023
	H4	0.57774	**0.39444**	0.55475	**0.41821**	0.52967	0.33088	0.50299	0.35851	**0.56614**	**0.37775**	**0.54167**	**0.40269**
Español	H1	0.58221	0.37045	0.55315	0.39808	**0.59845**	**0.39288**	**0.57178**	**0.41868**	0.59827	**0.39197**	**0.57054**	0.41400
	H2	0.58105	0.36419	0.55157	0.39277	0.59449	0.38367	0.56499	0.41041	**0.59878**	0.39173	0.57024	**0.41580**
	H3	**0.58786**	**0.38836**	**0.56577**	**0.41942**	0.58710	0.38716	0.56487	0.41842	0.57595	0.37670	0.55273	0.40714
	H4	0.58850	0.38738	0.56571	0.41780	0.58357	0.38151	0.56046	0.41247	0.57913	0.37658	0.55423	0.40712
Alemán	H1	0.52992	0.35991	0.50932	0.36927	0.52527	0.35010	0.50142	0.36028	0.52156	0.35081	0.49985	0.35968
	H2	**0.53300**	**0.36391**	**0.51279**	**0.37096**	**0.53364**	**0.36400**	**0.50996**	**0.37128**	**0.54070**	**0.37632**	**0.51825**	**0.38161**
	H3	0.49642	0.32025	0.47508	0.33613	0.50753	0.33865	0.48716	0.35300	0.51446	0.34086	0.49254	0.35533
	H4	0.49835	0.32219	0.47705	0.33787	0.50750	0.33725	0.48654	0.35138	0.52221	0.35199	0.50156	0.36509

Tabla 1: Resultados ROUGE corpus JRC (Medida F)

		(a) PCA_base				(b) PCA_base+CPT				(c) PCA_base+CPT+NER			
		R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4	R-1	R-2	R-L	R-SU4
Inglés	H1	**0.43991**	**0.11488**	**0.34974**	**0.16889**	0.43633	**0.11173**	0.34399	0.16504	0.43255	0.10784	0.34147	0.16328
	H2	0.43812	0.11062	0.34873	0.16589	**0.43785**	0.11116	**0.34496**	**0.16552**	**0.43633**	**0.11044**	**0.34607**	**0.16603**
	H3	0.41745	0.09944	0.33273	0.15724	0.41749	0.10070	0.33370	0.15799	0.40548	0.09234	0.32157	0.14997
	H4	0.41866	0.09941	0.33315	0.15751	0.41757	0.10068	0.33316	0.15803	0.40510	0.09219	0.32094	0.14981
Español	H1	0.43477	**0.12031**	0.35097	0.17376	0.43940	**0.11739**	0.35430	**0.17348**	0.43491	0.11601	0.34709	0.17016
	H2	**0.44031**	0.11863	**0.35482**	**0.17461**	**0.44024**	0.11617	**0.35614**	0.17259	**0.44125**	**0.12012**	**0.35578**	**0.17509**
	H3	0.41901	0.10849	0.33738	0.16820	0.41978	0.10963	0.33900	0.16925	0.41949	0.10567	0.33620	0.16499
	H4	0.41878	0.10833	0.33692	0.16809	0.42030	0.10995	0.33937	0.16951	0.42004	0.10601	0.33600	0.16539
Alemán	H1	**0.26939**	**0.04422**	**0.19030**	**0.07323**	**0.27202**	**0.04266**	**0.19197**	**0.07352**	**0.26873**	0.04078	**0.19231**	**0.07103**
	H2	0.26782	0.04403	0.19020	0.07242	0.26662	0.04111	0.18846	0.07135	0.25946	**0.04116**	0.18835	0.06889
	H3	0.25689	0.03099	0.18424	0.06621	0.25689	0.03099	0.18424	0.06621	0.25641	0.03017	0.18391	0.06576
	H4	0.25780	0.03099	0.18462	0.06638	0.25684	0.03100	0.18414	0.06622	0.25636	0.02996	0.18380	0.06553

Tabla 2: Resultados ROUGE corpus de entrenamiento de MultiLing 2015 (Medida F)

dientes al corpus JRC y corpus de entrenamiento de MultiLing 2015 respectivamente. Resulta lógico que los resultados del corpus JRC sean notablemente superiores a los del corpus MultiLing 2105, debido a la naturaleza distinta de cada corpus y a los resúmenes utilizados como modelo para la evaluación, dado que los del corpus JRC son de tipo extractivo, mientras que los del MultiLing 2015 son abstractivos.

La aportación de conocimiento léxico-semántico es muy dependiente de los propios textos, de la cantidad de entidades y de los sinónimos que alberguen. Es por ello que se han estudiado los dos corpus para ver las características de los documentos (número de palabras por documento (sin stopwords); número de NER identificadas; y número de conceptos sinónimos identificados).

Estas características, reflejadas en la Tabla 3, nos pueden servir para sacar conclusiones de la relevancia de la adición de conocimiento léxico-semántico al proceso. Cabe destacar la escasa identificación de conceptos sinónimos en los corpus, siendo los documentos de entrenamiento del Multiling en español en los que más sinónimos se han identificado (1.76 %). A partir de un análisis más en profundidad de los corpus utilizados, se ha comprobado que efectivamente no abunda el uso de conceptos sinónimos y por el contrario, predominan más las referencias a entidades nombradas, sobre todo a entidades de tipo lugar (Inglaterra, Egipto, Japón, Estados Unidos, Europa, entre otras) y de tipo persona (Jane Austen, Harris Bigg-Wither, Thomas Blanchard, Woo-Suk Hwang, Presidente Bush, entre otras).

	Idioma	**PPD**	**NER**	**Conceptos sinónimos**
JRC	Inglés	372.10	3.52 %	0.48 %
	Español	454.15	3.33 %	0.40 %
	Alemán	376.65	2.60 %	0.18 %
MultiLing	Inglés	1979.46	4.36 %	1.27 %
	Español	3054.90	6.38 %	1.76 %
	Alemán	1999.76	3.84 %	0.43 %

Tabla 3: Valores medios de las estadísticas de los documentos. PPD: Palabras por documento (sin stopwords)

Como se puede apreciar, no existe una técnica concreta que represente los mejores resultados para los tres idiomas, por lo que es difícil generalizar y determinar el mejor enfoque. No obstante, se destaca el enfoque sin análisis semántico (*PCA_base*) ya que presenta muy buenos resultados, siendo muy interesante dado que es totalmente independiente del idioma, al contrario que los métodos con conocimiento léxico-semántico, en los que hay que disponer de sistemas de reconocimiento de entidades y recursos para identificar conceptos específicos para cada idioma, cuya repercusión y correcto funcionamiento condicionan en gran medida los resultados obtenidos.

Por otro lado, los resultados (ROUGE-1) para el idioma español son interesantes ya que ofrece su mejor valor con H2 cuando incorporamos co-

nocimiento léxico-semántico para ambos corpus, obteniendo los mejores resultados en comparación con el resto de idiomas.

En general, la aportación de conocimiento léxico-semántico en ambos corpus es mínima, dando lugar a que su contribución no sea excesivamente notable. En el caso del corpus MultiLing 2015 aunque en porcentaje la aportación de NER es mayor que para el corpus JRC, ese porcentaje tiene menor efecto dada la extensión total de los documentos a resumir. Es por ello que el enfoque que incluye NER mejora en algunos casos para el corpus JRC (sobre todo para las heurísticas H1 y H2), pero no para el corpus MultiLing 2015.

En comparación con el enfoque independiente del idioma, las mejorar con la incorporación de conocimiento léxico-semántico no son las esperadas. Esto puede deberse, en parte, a que el uso de recursos y herramientas externas para identificar tanto NERs como conceptos sinónimos pueda dar lugar a la presencia de errores cometidos por las propias herramientas, o factores como la no correcta asignación de un sentido a una palabra, ya que se realiza el tipo de desambiguación más básica. Por ello, de los resultados obtenidos, así como del análisis en detalle de algunos resumenes generados se han identificado una serie de limitaciones no contempladas inicialmente en nuestro enfoque, y que podrían afectar negativamente a la calidad de los resúmenes. La primera de ellas es el uso de conocimiento semántico sin ninguna técnica de desambiguación. Debido a que la tarea de desambiguación es muy compleja y que no era el objetivo principal de este trabajo, optamos por realizar la identificación de conceptos según su sentido más frecuente, y por lo tanto, no teniendo en cuenta el contexto en el que se está utilizando el término en cuestión. A pesar de que los resultados para el sentido más frecuente giran en torno al 50 % (McCarthy, 2011), esto puede dar lugar a que: i) no se estén identificando correctamente algunos conceptos; y ii) se cometan errores en la agrupación de los conceptos. Técnicas más recientes y precisas en la tarea de desambiguación (Agirre et al., 2014), así como el uso de recursos más actualizados y con mayor cobertura, como Multilingual Central Repository o GermaNet podrían contribuir a mejorar los resultados.

Por otro lado, una vez calculada la matriz obtenida a partir de la técnica PCA, estamos teniendo en cuenta todas las palabras que integran cada uno de los conceptos identificados para realizar la selección de las frases hasta que se alcanza una determinada longitud. Hubiera sido interesante analizar y aplicar alguna técnica para realizar una segunda selección de entre todas esas frases para que los resúmenes generados recogieran una mayor variedad de conceptos, ya que debido a la longitud impuesta por los resúmenes modelo, éstos se generaron con el tamaño especificado.

Tal y como se comenta en la sección 8, se plantea abordar estos y otros aspectos en los trabajos futuros para determinar si solventando estas limitaciones, la aportación de conocimiento léxico-semántico tiene una influencia positiva mayor en la técnica PCA o, si por el contrario, la influencia es ngativa.

7 Comparativa con respecto a sistemas existentes

Una vez analizados los resultados, en esta sección vamos a realizar una comparativa de los mejores resultados de nuestros métodos con respecto a algunos sistemas de resúmenes multilingües existentes. Los sistemas utilizados se han seleccionado dada su disponibilidad y accesibilidad para la generación de resúmenes multilingües con ambos corpus. En concreto, dichos sistemas son:

- Open Text Summarizer (OTS)[10]. El enfoque implementado en este sistema identifica las palabras clave mediante la ocurrencia de las palabras. Emplea algunos recursos específicos por idioma tales como analizadores lingüísticos y listas de stopwords para más de 25 idiomas.
- Resumidor integrado en Microsoft Word 2007 (MS Word)[11]. Dado que es un sistema comercial, los detalles de implementación no son públicos.
- Essential Summarizer (Essential)[12]. Este sistema es una versión comercial del presentado por (Lehmam, 2010). Se basa en técnicas lingüísticas para realizar análisis semántico, teniendo en cuenta los elementos discursivos del texto.

Además, dado que el corpus JRC fue previamente utilizado por (Lloret & Palomar, 2011), se incluye en la comparativa sus mejores resultados obtenidos por el enfoque dependiente del idioma (LS), que emplea recursos específicos para cada idioma (etiquetador gramatical, NER e identificación de conceptos), siendo similar a nuestro método (c) *PCA_base+CPT+NER*. La Tabla 4 muestra la comparativa realizada.

[10] http://libots.sourceforge.net/

[11] https://support.office.com/en-nz/article/Automatically-summarize-a-document-b43f20ae-ec4b-41cc-b40a-753eed6d7424

[12] https://essential-mining.com/

	Sistema	Inglés	Español	Alemán
JRC	Mejor (a)	**0.57797**	0.58786	0.53300
	Mejor (b)	0.54576	0.59845	0.53364
	Mejor (c)	0.56614	0.59878	**0.54070**
	LS	0.56530	**0.62351**	0.52614
	OTS	0.55732	0.60591	0.53451
	MS Word	0.53591	0.57396	0.48427
	Essential	0.52622	0.53978	0.43727
MultiLing	Mejor (a)	**0.43991**	0.44031	0.26939
	Mejor (b)	0.43785	0.44024	**0.27202**
	Mejor (c)	0.43633	**0.44125**	0.26873
	OTS	0.43090	0.41345	0.26293
	MS Word	0.43382	0.40501	0.27096
	Essential	0.41382	0.39131	0.24127

Tabla 4: Comparativa (R-1, medida F) con diferentes enfoques - (a)*PCA_base*; (b)*PCA_base+CPT*; (c)*PCA_base+CPT+NER*.

Como se puede observar, nuestros enfoques presentan los mejores resultados para el corpus de entrenamiento MultiLing 2015, superándolos para todos los idiomas. Con respecto al corpus JRC, nuestros enfoques mejoran los resultados para alemán e inglés pero no consiguen superar a los de OTS y LS para el idioma español, a pesar de quedarse muy cercanos. La razón por la que nuestro método no haya sido capaz de obtener mejores resultados que el método LS de (Lloret & Palomar, 2011) puede deberse a que en nuestro enfoque no se utiliza ningún método para la desambiguación del sentido de las palabras, mientras que en el trabajo de referencia analizan los documentos en español mediante el analizador Freeling (Padró & Stanilovsky, 2012), que procesa y anota semánticamente un texto utilizando algoritmos de desambiguación como UKB (Agirre & Soroa, 2009), que han demostrado superar a la aproximación del sentido más frecuente. Cabe destacar el enfoque (c) ya que la adición de conocimiento léxico-semántico se hace de manera similar a LS, consiguiéndose mejorar los resultados para el inglés y alemán. El enfoque base (a) presenta buenos resultados y puede ser generalizable para cualquier idioma dado que su ejecución es invariante del idioma. Por lo tanto, se puede concluir que los métodos propuestos presentan resultados muy competitivos comparados con sistemas comerciales.

8 Conclusiones y trabajos futuros

En este artículo se ha realizado un estudio de la técnica PCA para la generación de resúmenes extractivos mono-documento y multilingües, analizando la influencia de introducir conocimiento léxico-semántico a la técnica base (reconocimiento de entidades e identificación de conceptos sinónimos) e investigando cuatro heurísticas diferentes para seleccionar la frase a partir de las palabras clave determinadas con la técnica PCA.

Para la experimentación se utilizaron dos corpus de diferente naturaleza (noticias periodísticas y artículos de la Wikipedia) y se generaron resúmenes automáticos para tres idiomas (inglés, español y alemán), evaluando la relevancia de la información seleccionada con respecto a resúmenes modelos utilizando la herramienta ROUGE.

Como conclusión general, la calidad de los resúmenes con conocimiento léxico-semántico es muy dependiente de las características de los textos que hay que resumir dado que la cantidad de entidades y conceptos que posean afecta en gran medida a los resultados. Los resúmenes generados con la técnica sin ningún tipo de conocimiento (*PCA_base*) presentan muy buenos resultados en comparación con otros sistemas existentes, siendo un enfoque atractivo dada su independencia del idioma con que se trabaje.

Para futuros trabajos, se propone mejorar el enfoque propuesto incluyendo el etiquetado gramatical, que podría mejorar notablemente el rendimiento de la etapa de identificación de conceptos, y replicar la experimentación utilizando recursos léxico-semánticos más actualizados, como el recurso Multilingual Central Repository para el castellano y el recurso GermaNet para el alemán. También planteamos redefinir la estrategia de selección de palabras clave, para realizar un filtrado intermedio de palabras clave relevantes y evitar así considerar todas las palabras de la matriz obtenida con la técnica PCA. Finalmente, sería interesante analizar los textos de los corpus para poder realizar y orientar el resumen que más se pueda adecuar, dado que cada heurística da lugar a un tipo de resumen diferente.

Agradecimentos

Esta investigación se ha realizado gracias a la financiación recibida en los proyectos: DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0 (PROMETEOII/2014/001) de la Generalitat Valenciana; SAM (FP7-611312) de la Comisión Europea; "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) y "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224)), del Ministerio de Economía y Competitividad del Gobierno de España; "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15), de la Universidad de Alicante.

Referencias

Agirre, Eneko, Oier López de Lacalle & Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Comput. Linguist.* 40(1). 57–84.

Agirre, Eneko & Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* EACL '09, 33–41. Stroudsburg, PA, USA: Association for Computational Linguistics.

Barzilay, Regina & Michael Elhadad. 1999. Using lexical chains for text summarization. En *Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization, 111–122. MIT Press.

Conroy, John M, Sashka T Davis, Jeff Kubina, Yi-Kai Liu, Dianne P. O'Leary & Judith D. Schlesinger. 2013. Multilingual Summarization: Dimensionality Reduction and a Step Towards Optimal Term Coverage. En *MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 55–63.

Cowie, Jim, Kavi Mahesh, Sergei Nirenburg & Remi Zajac. 1998. MINDS - Multi-lingual INteractive Document Summarization. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization* 122–123.

Dunteman, George H. 1989. *Principal components analysis* 69. Sage.

Estellés Arolas, Enrique, Fernando González Ladrón De Guevara & Antonio Falcó Montesinos. 2010. Principal Component Analysis for Automatic Tag Suggestion. Relatório técnico. http://dspace.ceu.es/bitstream/10637/6327/1/Principal%20component%20analysis%20for%20automatic%20tag%20suggestion.pdf.

Giannakopoulos, G, M El-Haj, J Steinberger, B Favre, M Litvak & V Varma. 2011. TAC 2011 MultiLing Pilot Overview. *TAC 2011 Workshop* .

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

Gupta, Vishal & Gurpreet Singh Lehal. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence* 2(3). 258–268.

Hovy, Eduard & Chin-yew Lin. 1997. Automated Text Summarization in SUMMARIST. En *ACL Workshop on Intelligent, Scalable Text Summarization*, 18–24.

Kubina, Jeff, John M Conroy & Judith D Schlesinger. 2013. ACL 2013 MultiLing Pilot Overview. En *MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 29–38.

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

Lehmam, Abderrafih. 2010. Essential summarizer: innovative automatic text summarization software in twenty languages. En *RIAO '10: Adaptivity, personalization and fusion of heterogeneous information*, 216–217. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

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

Litvak, Marina & Mark Last. 2013. Multilingual Single-Document Summarization with MUSE. En *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, 77–81. Association for Computational Linguistics.

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

Mani, Inderjeet & Mark T. Maybury. 1999. *Advances in automatic text summarization.* The MIT Press. ISBN 0-262-13359-8.

McCargar, Victoria. 2005. Statistical approaches to automatic text summarization. *Bulletin of the American Society for Information Science and Technology* 30(4). 21–25.

McCarthy, Diana. 2011. Word sense disambiguation. Seminar.

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

Nenkova, Ani & Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval* 5(2-3). 103–233.

Padró, Lluís & Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey: ELRA.

Patel, Alkesh, Tanveer Siddiqui & U. S. Tiwary. 2007. A language independent approach to multilingual text summarization. *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)* .

Ringnér, Markus. 2008. What is principal component analysis? *Nature biotechnology* 26(3). 303–304.

Sparck-Jones, Karen. 1999. Automatic summarising: factors and directions. *Advances in Automatic Text Summarization* 1–21.

Spärck Jones, Karen. 2007. Automatic summarising: The State of the Art. *Information Processing & Management* 43(6). 1449–1481.

Steinberger, Josef. 2013. The uwb summariser at multiling-2013. En *Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization*, 50–54. Sofia, Bulgaria: Association for Computational Linguistics.

Teng, Zhi, Ye Liu, Fuji Ren, Seiji Tsuchiya & Fuji Ren. 2008. Single document summarization based on local topic identification and word frequency. En *Proceedings of the Seventh Mexican International Conference on Artificial Intelligence*, 37–41. Washington, DC, USA: IEEE Computer Society.

Tjong, Erik F, Kim Sang & Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. En *7th conference on Natural language learning at HLT-NAACL 2003*, vol. 4, 142–147. Association for Computational Linguistics.

Uzêda, Vinícius Rodrigues, Thiago Alexandre Salgueiro Pardo & Maria Das Graças Volpe Nunes. 2010. A comprehensive comparative evaluation of rst-based summarization methods. *ACM Trans. Speech Lang. Process.* 6(4). 4:1–4:20.

Vossen, Piek. 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.