Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 37-44 recibido 11-11-2016 revisado 12-01-2017 aceptado 07-02-2017

Un detector de la unidad central de un texto basado en técnicas de aprendizaje automático en textos científicos para el euskera*

A Machine Learning based Central Unit Detector for Basque Scientific Texts

Kepa Bengoetxea, Aitziber Atutxa y Mikel Iruskieta
IXA Group. University of the Basque Country
{kepa.bengoetxea,aitziber.atucha,mikel.iruskieta}@ehu.eus

Resumen: En este artículo presentamos el primer detector de la Unidad Central (UC) de resúmenes científicos en euskera basado en técnicas de aprendizaje automático. Después de segmentar el texto en unidades de discurso elementales, la detección de la unidad central es crucial para anotar de forma más fiable la estructura relacional de textos bajo la Teoría de la Estructura Retórica o *Rhetorical Structure Theory* (RST). Además, la unidad central puede ser explotada en diversas tareas como resumen automático, tareas de pregunta y respuesta o análisis del sentimiento. Los resultados obtenidos demuestran que las técnicas de aprendizaje automático superan a las técnicas basadas en reglas a pesar del pequeño tamaño del corpus y de la heterogeneidad de los dominios que éste muestra, dejando todavía lugar para mejoras y desarrollo.
Palabras clave: Unidad central, tópico principal, RST, aprendizaje automático

Abstract: This paper presents an automatic detector of the discourse central unit (CU) in scientific abstracts based on machine learning techniques. After segmenting a text in its elementary discourse units, the detection of the central unit is a crucial step on the way to robustly build discourse trees under the *Rhetorical Structure Theory* (RST). Besides, CU detection may also be useful in automatic summarization, question answering and sentiment analysis tasks. Results show that the CU detection using machine learning techniques for Basque scientific abstracts outperform rule based techniques, even on a small size corpus on different domains. This leads us to think that there is still room for improvement.
Keywords: Central unit, main topic, RST, machine learning

1 Introducción

Saber cuál es el tema principal o la idea global del texto es una tarea relativamente fácil siempre que se domine la lengua; aunque también es cierto que dicha tarea puede complicarse en algunos textos que no exponen la idea principal explícitamente, para conseguir un efecto comunicativo o simplemente porque los textos no están bien redactados.

El tema principal puede ser representado de diferentes formas: *i*) por elementos o palabras clave (desde una única palabra a una lista de palabras), *ii*) por proposiciones u oraciones completas.

Según Iruskieta, Diaz de Ilarraza, y Lersundi (2014) la detección del tema principal o unidad central (UC)[1] es de gran ayuda en la anotación de la estructura retórica, ya que conocer de antemano cúal es la UC permite mejorar el ratio de acuerdo entre anotadores en la Rhetorical Structure Theory (RST) de Mann y Thompson (1988). Teniendo en cuenta esos resultados, pensamos que un analizador discursivo automático podría

* Agradecemos tanto a Kike Fernandez como a Esther Miranda todo el trabajo técnico para poder analizar y visibilizar los resultados de este trabajo. Este trabajo a sido financiado en parte por el siguiente proyecto: TIN2015-65308-C5-1-R (MINECO/FEDER).

[1] La Unidad Central (UC) es un concepto asociado con los árboles de la RST y es la unidad discursiva elemental (UDE) más importante del árbol que tiene la función de ser el principal núcleo del árbol, aunque puede constar de múltiples UDEs en el caso de parataxis.

ISSN 1135-5948

ofrecer resultados más fiables si detectara la unidad central tras la segmentación discursiva automática (Iruskieta y Zapirain, 2015). Además, podría ser utilizado en tareas del Procesamiento del Lenguaje Natural (PLN), aquellas como, resumen automático, análisis del sentimiento o búsqueda de respuestas.

El objetivo de este artículo es construir un detector automático de unidades centrales en textos científicos para el euskera construyendo un clasificador del tipo *Multivariate Bernoulli Naive Bayes*.[2]

Para entrenar y evaluar el detector automático de la unidad central hemos utilizado el corpus[3] *Basque RST Treebank* (Iruskieta et al., 2013), previamente anotado para otros propósitos y tareas (y el único accesible para el euskera).[4]

En el Ejemplo (1) presentamos un texto de ese corpus anotado manualmente: con los segmentos enumerados y la UC en negrita.

(1) [Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.]$_1$ ["Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da,]$_2$ [tamainu, kokapena eta iraunkortasuna aldakorra izanik.]$_3$ [Honen etiologia eztabaidagarria da.]$_4$ [Ultzera mingarri batzu bezala agertzen da,]$_5$ [hauek periodiki beragertzen dira.]$_6$ [**Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.**]$_7$ GMB03013[5]

El texto del Ejemplo (1) se ha segmentado en 7 Unidades de Discurso Elementales (UDE)[6] y la unidad central es la última de ellas, la UDE$_7$.

Según Paice (1980) existen algunos indicadores que facilitan detectar automáticamente las ideas principales. Basándonos en esos indicadores y otros que hemos desarrollado en este estudio, la UDE$_7$ muestra los siguientes:

i) *Lan honetan* 'en este trabajo', el nombre *lan* 'trabajo' junto al demostrativo *hau* 'este' junto con el sufijo -*n* (inesivo) de lugar, se refiere al trabajo que el autor presenta en el resumen.

ii) *Garrantzitsuena* 'el más importante', el adjetivo *garrantzitsu* 'importante' y el superlativo −*en*− 'el más' indican que el elemento modificado por el adjetivo está resaltado de alguna forma en la oración.

iii) *Analizatu dugu* 'hemos analizado', el verbo *analizatu* 'analizar' es común para expresar la acción principal que se realiza en trabajos de investigación (Iruskieta, Diaz de Ilarraza, y Lersundi, 2014) y el pronombre adjunto al verbo auxiliar −*gu* 'nosotros', indica que el la acción la han desarrollado los autores del artículo.

Aunque los indicadores[7] por si solos pueden ser ambiguos, ya que pueden utilizarse en otras UDEs que no son unidades centrales, nuestra hipótesis es que podemos detectar la unidad central de resúmenes científicos de una forma aceptable, utilizando adecuadamente todos estos indicadores con técnicas de aprendizaje automático.

En lo que sigue del artículo, explicamos en la Sección 2 los trabajos relacionados en los que nos hemos basado. En la Sección 3 la metodología que hemos empleado para construir el detector de la unidad central. En la Sección 4 presentamos el sistema y en la Sección 5 los resultados obtenidos. Finalmente, exponemos en la Sección 6 las conclusiones y

[2]Los textos utilizados son relativamente complejos teniendo en cuenta la disposición discursiva de la unidad central, ya que la unidad central puede estar en diferentes posiciones en el texto: al principio, en la mitad o al final del texto.

[3]Este corpus puede ser consultado en `http://ixa2.si.ehu.eus/diskurtsoa/`.

[4]Aunque en este trabajo nos hemos basado en la RST, pensamos que la detección de la unidad central podría ser aplicable también en otras teorías.

[5]Texto literalmente traducido: [La Estomatitis Aftosa Recurrente (I): Epidemiología, etiopatogenia y aspectos clínicopatológicos.]$_1$ [La estomatitis aftosa recurrente es una de las patologías orales más frecuentes,]$_2$ [de tamaño, localización y duración variable.]$_3$ [Su etiología es todavía controvertida.]$_4$ [Se caracteriza por la aparición de úlceras dolorosas,]$_5$ [estas recidivan periodicamente.]$_6$ [En este trabajo analizamos las principales características epidemiológicas, etiopatogénicas y clinicopatológicas de esta común patología oral.]$_7$

[6]Las UDEs son los bloques o segmentos más pequeños de los que consta una estructura en árbol discursivo (Carlson, Marcu, y Okurowski, 2001). En general, las UDEs son enunciados independientes o adverbiales.

[7]Otros indicadores en este texto aunque más complejos son: *i*) las palabras o lemas repetidos del título: *epidemiologia* 'epidemiología', *etiopatogenia* y *klinikopatologia* 'clinicopatología', *ii*) los sinónimos como *aspektu* 'aspecto' y *ezaugarri* 'característica', y *iii*) la relación de anafora entre *Estomatitis Aftosa Recurrente* y *patologia arrunt honetan* 'esta patología común'.

el trabajo futuro.

2 Trabajos relacionados

La extración de la unidad más relevante de un texto se ha estudiado con diferentes propósitos y aplicando distintas técnicas. Luhn (1958) hace uso de información estadística sobre una lista de palabras significativas o clave para la extracción de las sentencias más relevantes en resúmenes literarios en inglés. Mientras que Neto et al. (2000) aplica la técnica TF-ISF (*Term Frequency-Inverse Sentence Frequency*) para generar de forma automática resúmenes de textos. En Pardo, Rino, y Nunes (2003) emplean ambas técnicas para extraer la oración más importante de textos científicos tanto en inglés como en portugués de Brasil y obtienen mejores resultados haciendo un ranking de sentencias basado en palabras clave y la posición de la oración.

La unidad central también se puede extraer automáticamente de aquellos analizadores que obtienen la estructura relacional del discurso en forma de árboles jerárquicos. Por ejemplo, se puede extraer del analizador CODRA[8] para el inglés (Joty, Carenini, y Ng, 2015), ya que ésta sería la UDE situada en la raiz del árbol.

Nuestro trabajo es similar al trabajo realizado por Burstein et al. (2001), que emplea un clasificador Bayesiano para identificar la oración temática del texto. El clasificador se sirve como características de la posición, de una lista de palabras clave y ciertas características discursivas basadas en el analizador RST de Marcu (2000). Para extraer la lista de palabras clave, hemos tomado como punto de partida el trabajo de Iruskieta et al. (2015) basado en reglas, para detectar la UC en resúmenes científicos de euskera.

En la sección 5, los resultados del presente experimento en el que se aplican técnicas de aprendizaje automático se comparan con aquellos obtenidos en Iruskieta et al. (2015) a partir de aplicación de reglas.

3 Metodología

3.1 Etapas

Las etapas para desarrollar nuestro detector de UCs basado en técnicas de aprendizaje automático han sido las siguientes:

i. Corpus. Se ha reutilizado el mismo corpus de Iruskieta et al. (2015) que consta de 100 resúmenes científicos en euskera segmentados y con las UCs anotadas manualmente.
ii. Indicadores. Se han utilizado los indicadores de Iruskieta et al. (2015).
iii. Optimización. Se ha elegido y optimizado el algoritmo de aprendizaje automático.
iv. Evaluación. Se ha evaluado el detector automático de UCs.

3.2 El corpus

El corpus sobre el que hemos realizado este estudio está conformado por 100 textos de 5 dominios diferentes (medicina (GMB), terminologia (TERM), ciencia (ZTF), ciencias de la salud (OSA) y de la vida (BIZ)), catalogados por UZEI[9] y la *Udako Euskal Unibertsitatea* (UEU).[10] El corpus de 100 textos contiene 15.168 palabras, cada texto con su unidad central. Presentamos el corpus con mayor detalle en la Tabla 1.

Dominio	Textos	Palabras	UDEs	UCs
GMB	20	2.753	247	29
TERM	20	5.398	523	37
ZTF	20	6.646	548	27
OSA	20	4.964	454	21
BIZ	20	5.407	572	23
Total	**100**	**15.168**	**2.344**	**137**

Tabla 1: Descripción del Corpus

Hemos empleado los dominios GMB, TERM y ZTF para entrenar nuestro sistema y generar el modelo de aprendizaje (incluyendo la selección características y la optimización hiperparamétrica), y los dominios OSA y BIZ para validar los resultados. El corpus de entrenamiento se ha dividido en 10 partes para realizar una validación cruzada. En la Tabla 2 hemos calculado si ambos corpus muestran la misma dificultad en la detección de la unidad central de este modo:

$Dificultad = \frac{UCs}{UDEs}$ cuanto más cerca de 1 es más fácil de determinar la UC.

Corpus	UDEs	UCs	Dificultad
Train	1.318	93	0,07050
Test	1.026	44	0,04288

Tabla 2: Dificultad para elegir la UC

[8] CODRA se puede probar muy facilmente aquí: http://alt.qcri.org/demos/Discourse_Parser_Demo/.

[9] http://www.uzei.eus/.

[10] http://www.ueu.eus/.

Según la información de la Tabla 2 detectar la UC en el corpus de validación (*test*) es más difícil. Los resultados obtenidos en (Iruskieta et al., 2015) también señalan que el resultado fué peor en esa parte del corpus.

El tamaño de este corpus (a nivel de número de textos) es similar al que se ha utilizado en trabajos ya mencionados anteriormente, como el de Paice (1980) con un corpus de 32 textos y el de Burstein et al. (2001) con 100 textos.

3.3 El método de anotación

El corpus fué anotado con la herramienta RSTTool[11] por dos linguistas expertos de RST, en tres fases:

i) Los anotadores segmentaron el texto en UDEs.

ii) Ambos anotadores determinaron cual o cuales de las UDEs formaban la UC.

iii) La anotación de la UC fue evaluada y harmonizada para obtener un *gold standard*.

3.4 Acuerdo entre anotadores

Dos anotadores anotaron manualmente las UDEs y las UCs.[12]

El acuerdo entre el anotador-1 (A1) y el anotador-2 (A2) con el coeficiente Kappa (κ) (Siegel y Castellan, 1988) fue del 0,796 (de un total de 2.344 UDEs). Este grado de acuerdo que está entre los valores del 0,8 κ (acuerdo muy alto) y del 0,6 κ (buen acuerdo) es aceptable, según Krippendorff (2004). También es comparable al acuerdo obtenido en trabajos similares como el de Burstein et al. (2001) con un acuerdo entre dos anotadores de 0,733 κ (de un total de 2.391 oraciones) en un corpus compuesto por 100 textos.[13]

3.5 Extracción de características

El corpus ha sido enriquecido con información morfosintáctica utilizando un analizador morfológico (Aduriz, 2000) y el desambiaguador morfológico (Ezeiza et al., 1998). Se ha creado una lista de palabras clave o significativas para la extracción de la unidad central, una vez que se han analizado las características que mejor indican las UCs en el corpus de entrenamiento. Tomando como referencia el trabajo de Paice (1980), hemos analizado qué verbos, nombres, pronombres y palabras claves (*bonus words*) permiten identificar la UC en nuestro corpus, incluyendo las características que fueran necesarias. Un resumen de las características que se utilizan aprendizaje automático puede verse en la Tabla 3.

Caract.	Descripción
Nombres	Lista de nombres relacionados con la UC
Verbos	Lista de verbos relacionados con la UC
Clave/bonus	Lista de adjetivos y adverbios
Ver. Auxiliares	Lista de verbos con la primera persona del plural
Determinantes	Del tipo *hau* 'este' y *hemen* 'aqui'
Pronombres	Primera persona del plural *gu* 'nosotros'
Combinaciones	Nombres + determinantes, pronombres + nombres y verbos + verbos auxiliares
Verbos principales	Si contiene un verbo principal
Título	Listas de palabras que aparecen en el texto del título
Posición	Posición del segmento en el texto
Posición UDE con verb. aux.	Orden del segmento entre los que incluyen un verbo auxiliar
Condional	Si contiene un verbo condicional
Lista de palabras de parada	Lista de palabras carentes de significado para las UCs

Tabla 3: Características para detectar la UC

3.6 Medidas de evaluación

Para evaluar el detector de la UC, el corpus se ha separado en dos partes. Una parte para el entrenamiento y otra para la prueba final de validación.

Se ha utilizado la misma separación de datos de entrenamiento y validación de Iruskieta et al. (2015) para poder comparar los resultados de ambos trabajos. Los experimentos se han realizado aplicando la técnica de *10-fold cross-validation* sobre los datos de entrenamiento y finalmente se ha evaluado sobre los datos de validación. Para evaluar el sistema se han utilizado las medidas habituales: Exhaustividad (*Recall*), Precisión, y los valores de ambas métricas combinadas en una media armónica denominada valor-F (*F-score* o F_1).

También se ha llevado a cabo un análisis de errores a nivel de texto, para entender como funciona el detector de la UC y ver si hay

[11] http://www.isi.edu/licensed-sw/RSTTool/.

[12] El *gold standard* de estos ficheros pueden ser consultados en http://ixa2.si.ehu.es/diskurtsoa/en/segmentuak.php.

[13] Los desacuerdos más comunes y el proceso de armonización para obtener un gold standard se describen en Iruskieta et al. (2015).

$$log(P(UC|UDE)) = log(P(UC)) + \sum_i \begin{cases} log(P(Ai|UC)/P(A_i)), \\ \text{Si UDE contiene } A_i \\ log(P(\overline{A_i}|UC)/P(\overline{A_i})), \\ \text{Si UDE no contiene } A_i \end{cases}$$

Tabla 4: Fórmula *Bernoulli multivariante*

lugar para mejoras.

4 El detector automático de UCs

Como se ha mencionado previamente, para crear un clasificador que detecte aquellos segmentos de un resumen que tienen mayor probabilidad para ser etiquetados como UC, se ha experimentado con diferentes algoritmos de clasificación como *Multinomial Naive Bayes*, *Multivariate Bernoulli Naive Bayes*, *Support Vector Machines (SVM)* con polinomios de grado 2 y 3, *Radial Basis Functions (RBF)* y *Single Perceptron*, utilizando tanto características basadas en frecuencia como binarias. Finalmente se ha optado por *Multivariate Bernoulli Naive Bayes* por las siguientes razones:

- Los parámetros necesarios para el clasificador se pueden estimar con corpus de entrenamiento pequeños.
- Ha sido utilizado con éxito en tareas similares: para identificar oraciones temáticas (Burstein et al., 2001) o para clasificar textos cortos (McCallum y Nigam, 1998).
- Puede ser empleado tanto como modelo predictivo como descriptivo.
- La aplicación de este clasificador es la que mejores resultados nos ha brindado sobre el corpus de entrenamiento.

La distribución de Bernoulli a la hora de clasificar tiene en cuenta tanto la ausencia como la presencia de las caraterísticas. Para enriquecer el modelo, hemos valido de las características que se muestran en la Tabla 3.

Empleando la fórmula de la Tabla 4, *Bernoulli multivariante*, se obtiene la probabilidad logarítmica que tiene una UDE para pertenecer a la clase UC. El rendimiento mejora si utilizamos el estimador de Laplace para hacer frente a los casos en que las estimaciones de probabilidad de ciertas características que son iguales a cero.

En la fórmula de la Tabla 4: *i*) $P(UC)$ es la probabilidad a priori para que una UDE pertenezca a la clase UC, *ii*) $P(A_i|UC)$ es la probabilidad condicional para que una UDE que pertenece a UC tenga la característica A_i, y *iii*) $P(A_i)$ es la probabilidad a priori para que una UDE contenga la característica A_i, *iv*) $P(\overline{A_i}|UC)$ es probabilidad condicional de que una UDE que pertenece a UC no tenga la característica A_i, y *v*) $P(\overline{A_i})$ es la probabilidad a priori para que una UDE no contenga la característica A_i.

4.1 Elección de un subconjunto de características usando un método Wrapper

Como los algoritmos ingenuos de Bayes sufren con las características reduntantes o correlacionadas, después de seleccionar el algoritmo de aprendizaje con todas las características de entrada, hemos aplicado un wrapper que nos permite seleccionar el mejor subconjunto de características para el clasificador seleccionado.

Para aplicar wrapper necesitamos definir los siguientes criterios:

- Operaciones en el Espacio de Búsqueda. Las operaciones puede ser "añadir característica" o "eliminar característica" o ambas. El término de "selección hacia delante" se refiere a realizar la búsqueda usando el operador "añadir característica", mientras que el término "selección hacia atrás" se refiere a realizar la búsqueda usando el operador "eliminar característica". Mientras que término "*step-wise*" usa ambos operadores. En nuestros experimentos hemos usado únicamente el operador "eliminar característica".
- Estimador de exactitud. Para medir la exactitud de cada operación hemos usado *ten-fold cross-validation* con la función de estimación *F-score*.
- El algoritmo de búsqueda. Para condu-

cir la búsqueda se puede usar diferentes algoritmos. En nuestro experimentos hemos usado el algoritmo de búsqueda *hill-climbing* con la "selección hacia atrás". El algoritmo empieza con todo el conjunto de características y progresivamente elimina una característica y en cada iteración genera sucesores del mejor nodo (aquel que ha obtenido el mayor *F-score*). La condición de terminación será cuando todos los sucesores de la iteracción actual no mejoren el valor de *F-score* de la iteracción anterior.

El wrapper resuelve que el subconjunto óptimo de características que mejor resultado ha obtenido es el siguiente: nombres, verbos, *bonus*, determinantes, pronombres, palabras del título, posición, verbos auxiliares y 3 combinaciones (nombres + determinantes, pronombres + nombres y verbos + verbos auxiliares).

4.2 Post-proceso estadístico

Finalmente, se ha realizado un post-proceso estadístico para los casos en los el clasificador no elija ninguna UDE como UC. En este caso, el post-proceso selecciona el primer candidato más probable de todos ellos, ya que el clasificador nos devuelve un valor de probabilidad para cada UDE.

4.3 Demo para detectar la UC

Una vez realizadas estas tareas, hemos desarrollado una demo, para que pueda ser utilizada por la comunidad científica. De esta forma, la demo pide un texto plano de entrada y ofrece dos formatos de salida diferentes: *i*) Formato web, para utilizar en tareas de PLN. *ii*) Formato RSTTool (RS3), para poder correguir la segmentación o la unidad central y seguir con la tarea manual de la anotación de las relaciones RST en euskera. La demo que puede ser consultada en `http://ixa2.si.ehu.es/CU-detector`.

5 Resultados

En la Tabla 5 se muestran varios resultados: *i*) *Rule Based*. En la primera fila se presenta el mejor resultado registrado en Iruskieta, Antonio, y Labaka (2016) utilizando métodos basados en reglas y aplicando la mejor heurística. *ii*) *ML*. En la segunda fila se pueden ver los resultados obtenidos con el clasificador Bernoulli Naive Bayes utilizando todas las características. *iii*) *ML + Wrap*. En la tercera fila aparecen los resultados obtenidos después de emplear el wrapper, y aplicando el mejor subconjunto de características obtenido. *iv*) *ML + Wrap + Post*. Y finalmente, en la cuarta fila se presentan los resultados después de aplicar el post-proceso estadístico. Obteniendo los mejores resultados en *F-score* de 0,54 con *10-fold cross-validation* y 0,57 con los datos de validación.

Sistema	Datos	Prec.	Rec.	F_1
Rule Based	Dev	0,43	0,51	0,47
	Test	0,70	0,40	0,51
ML	Dev	0,47	0,48	0,48
	Test	0,46	0,54	0,50
ML+Wrap	Dev	0,58	0,46	0,51
	Test	0,46	0,59	0,51
ML+Wrap+Post	Dev	0,56	0,53	**0,54**
	Test	0,48	0,70	**0,57**

Tabla 5: Tabla de resultados

5.1 Análisis de errores

Los diferentes tipos de acuerdos y desacuerdos que hemos observado en el análisis global (texto por texto) de errores que describimos en la Tabla 6 son los siguientes:

- Acuerdo total (coincidencia). El detector solamente ha etiquetado como UC, aquella UDE que se determina como UC en el *gold standard*.
- Acuerdo en UC, pero con falsos candidatos (exceso). Además de las UCs determinadas, el detector ha etiquetado otras UDEs que nos son UCs en el texto.
- Acuerdo parcial en UCs múltiples (falta). El detector ha detectado alguna UC del texto, pero ha dejado otras UCs sin etiquetar.
- Desacuerdo total (desacuerdo). El detector no ha detectado bien ninguna UC del texto.

	Coinc.	Exc.	Falta	Desac.
ML+Wrap	13	13	0	14
ML+Wrap+Post	16	13	2	9

Tabla 6: Análisis de errores

Si comparamos los resultados obtenidos con el método *ML+Wrap* y con el ML+Wrap+post de la Tabla 6, observamos que el postproceso mejora los resultados; ya que, hay mayor número de acuerdos: *i*) hay mayor 'coincidencia' y *ii*) hay mayor número de 'falta', que son acuerdos parciales, ya que

por lo menos una de las UCs ha sido etiquetada adecuadamente.

Hemos podido observar que las causas de los errores cometidos por el sistema en los resultados del post-proceso, son los siguientes:

- 'Exceso'. En 13 ocasiones se ha detectado la UC y otro falso candidato. En 10 ocasiones la primera UC detectada por el sistema es el único válido y en 7 de ellas es la UDE con más indicadores. En las otras 3 ocasiones, el sistema debería decantarse por el segundo candidato detectado con también con más indicadores.
- 'Falta'. En 2 ocasiones se ha detectado una sola UC de las UCs múltiples anotadas manualmente.
- 'Desacuerdo'. En 9 ocasiones el detector no ha sabido establecer correctamente la UC. En 2 ocasiones el texto no cuenta con indicadores suficientes para su detección. En otras 5 ocasiones la unidad central no se presenta como tema principal, sino como una definición o se anuncia mediante una catáfora. En las otras 2 el sistema ha fallado, porque no se han definido alguna otra característica, como por ejemplo la de darle importancia a que algunas caracteríticas estén unas detrás de otras.

Observando estos datos pensamos que hay lugar para mejorar resultados desarrollando técnicas para seleccionar candidatos en el postproceso basándonos en reglas.

6 Conclusiones y trabajo futuro

La mayor aportación de este trabajo es que se ha creado el primer detector de la unidad central (UC) de textos científicos para el euskera, que primero segmenta los textos en UDEs y después determina la UC utilizando únicamente técnicas de aprendizaje automático.[14] La UC se puede extraer del análisis automático que realizan otros analizadores de la RST, como por ejemplo del analizador CODRA, que está entrenado con textos periodísticos en inglés y no para abstracts científicos.

Ahora mismo estamos estudiando si es posible mejorar los resultados obtenidos de las siguientes formas:

- Combinando otras técnicas de aprendizaje automático.
- Combinando diferentes sistemas basados en reglas y en aprendizaje automático.

En el futuro también queremos medir la utilidad de este detector en tareas del PLN y adaptar este detector a otras lenguas y géneros textuales.

- Utilizar en tareas de búsqueda de respuestas (Aldabe et al., 2013) para preguntar sobre el tema principal del texto.
- Aplicar en tareas de análisis del sentimiento en euskera, ya que mejora resultados según Alkorta et al. (2015).
- Adaptar el detector a otras lenguas y evaluarlo con corpus anotados con RST, como pueden ser:
  - La *Spanish RST Treebank* (da Cunha et al., 2011) con 267 textos anotados.
  - La *RST Treebank* en inglés (Carlson, Okurowski, y Marcu, 2002) con 385 textos anotados.

Bibliografía

Aduriz, I. 2000. *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz.* Ph.D. tesis, Euskal Herriko Unibertsitatea, UPV/EHU, Donostia.

Aldabe, I., I. Gonzalez-Dios, I. Lopez-Gazpio, I. Madrazo, y M. Maritxalar. 2013. Two approaches to generate questions in basque. *Procesamiento del Lenguaje Natural*, (51):101–108.

Alkorta, J., K. Gojenola, M. Iruskieta, y A. Perez. 2015. Using relational discourse structure information in Basque sentiment analysis. En *5th Workshop RST and Discourse Studies", in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante.

Burstein, J., D. Marcu, S. Andreyev, y M. Chodorow. 2001. Towards automatic classification of discourse elements in essays. En *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, páginas 98–105. Association for Computational Linguistics.

[14] Este detector se puede probar en http://ixa2.si.ehu.es/CU-detector.

Carlson, L., D. Marcu, y M. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. En *2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, página 10, Aalborg, Denmark, 1-2 September. Association for Computational Linguistics.

Carlson, L., M. E. Okurowski, y D. Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

da Cunha, I., J.-M. Torres-Moreno, G. Sierra, L.-A. Cabrera-Diego, y B.-G. Castro-Rolón. 2011. The RST Spanish Treebank On-line Interface. En *International Conference Recent Advances in NLP*, Bulgaria, 12-14 September.

Ezeiza, N., I. Alegria, J.-M. Arriola, R. Urizar, y I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. *Proceedings and 17th International Conference on Computational Lingustics*, 1:380–384.

Iruskieta, M., J. Antonio, y G. Labaka. 2016. Detecting the central units in two different genres and languages: a preliminary study of brazilian portuguese and basque texts. *Procesamiento de Lenguaje Natural*, (56):65–72.

Iruskieta, M., M. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, y O. L. de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. En *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.

Iruskieta, M., A. Diaz de Ilarraza, G. Labaka, y M. Lersundi. 2015. The Detection of Central Units in Basque scientific abstracts. En *5th Workshop RST and Discourse Studiesïn Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN)*, Alicante.

Iruskieta, M., A. Diaz de Ilarraza, y M. Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. En *COLING*, páginas 466–475, Dublin. Dublin City University and ACL.

Iruskieta, M. y B. Zapirain. 2015. Euseduseg: a dependency-based edu segmentation for basque. *Procesamiento del Lenguaje Natural*, (55):41–48.

Joty, S., G. Carenini, y R. T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Sage.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Mann, W. C. y S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marcu, D. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

McCallum, A. y K. Nigam. 1998. A comparison of event models for naive bayes text classification. En *AAAI-98 workshop on learning for text categorization*, volumen 752, páginas 41–48.

Neto, J. L., A. D. Santos, C. A. Kaestner, y A. A. Freitas. 2000. Generating text summaries through the relative importance of topics. *Advances in Artificial Intelligence*, páginas 300–309.

Paice, C. D. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. En *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, páginas 172–191. Butterworth & Co.

Pardo, T., L. Rino, y M. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, páginas 196–196.

Siegel, S. y N. Castellan. 1988. The Friedman two-way analysis of variance by ranks. *Nonparametric statistics for the behavioral sciences*, páginas 174–184.