Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 165-175 recibido 29-11-2020 revisado 19-01-2021 aceptado 30-01-2021

Generación automática de meta-resúmenes para la evaluación del manejo de estructuras discursivas y coherencia en el alumnado

Automatic generation of meta-summaries for evaluation of the handling of discursive structures and coherence in students

Unai Atutxa[1], **Alejandro Molina-Villegas**[2], **Mikel Iruskieta**[1]
[1] HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain
[2] Conacyt - Centro de Investigación en Ciencias de Información Geoespacial, Mexico
atutxaunai@gmail.com, amolina@centrogeo.edu.mx, mikel.iruskieta@ehu.eus

Resumen: La técnica de crowd-sourcing puede ser una herramienta de gran ayuda tanto para evaluar los resúmenes de los alumnos como para poder ofrecerles un *feedback* que ayude a mejorar sus destrezas para resumir. En este trabajo, se propone un enfoque para la generación de meta-resúmenes en euskera, con el objetivo de diseñar y desarrollar una evaluación automática de los resúmenes de extracción. Se presenta un nuevo algoritmo que permite usar los meta-resúmenes generados con las siguientes finalidades: *i*) comparar los resúmenes elaborados por alumnos de diferentes edades y cursos educativos (primaria y universidad), *ii*) evaluar los resúmenes creados en clase (evaluación de la clase) y *iii*) evaluar a cada alumno (evaluación individual). Los resultados muestran que el método propuesto, el cual se ha elaborado basándose en aspectos cualitativos (estructura discursiva de la coherencia) y cuantitativos (kappa de Fleis y distancia de Hamming), es apto para comparar grupos e individuos.
Palabras clave: Evaluación de Resumen Automático, Análisis del discurso, PLN en Euskera.

Abstract: Crowd-sourcing can help teachers to evaluate student summaries and give them feedback to improve their summarization skills. In this paper, we propose an approach for meta-summaries generation, to design and develop the automatic evaluation of extractive summaries for the Basque language. We propose a novel algorithm that allows to use the generated meta-summaries to *i*) compare students meta-summaries at different ages and education stages (elementary and undergraduates), *ii*) evaluate classroom meta-summaries (classroom evaluation) and *iii*) evaluate each student (individual evaluation). The results show that our proposed method, based on qualitative (coherence discourse structure) and quantitative (Fleis kappa and Hamming distance) measures, is accurate to compare both: groups and individuals.
Keywords: Summarization Evaluation, discourse Analysis, Basque NLP.

1 Introducción

En la actualidad, el exceso de información puede conducirnos a no distinguir lo que es realmente relevante. En ese contexto, el resumen es de vital importancia en áreas como la educación, ya que un resumen muestra la capacidad de comprensión y de síntesis de quien lo ha elaborado. El hecho de que un alumno disponga de una gran cantidad de información no significa que vaya a entender más y mejor el tema de estudio, lo cual puede provocar que el proceso de aprendizaje no sea óptimo. Esto sugiere que trabajar el resumen de una manera eficaz en el aula puede ser de gran ayuda para lograr los objetivos curriculares.

Cuando se resume, primeramente se debe entender lo que se ha leído, y después, plasmar las ideas más importantes del texto adecuando el lenguaje al conocimiento propio.

ISSN 1989-7553. DOI 10.26342/2021-66-14

Las ideas más relevantes deben ser extraídas del texto, manteniendo la coherencia entre dichas ideas. Sin embargo, los resúmenes producidos por los estudiantes no siempre coinciden con los resúmenes producidos por los profesores o evaluadores. En el ámbito educativo, es habitual que los libros de texto (o la información de los sistemas de gestión de aprendizaje) empleados en la escuela no dispongan de actividades que desarrollen la destreza de resumir, al menos en la educación del País Vasco. En consecuencia, raramente se trabajan estrategias que ayuden a desarrollar la capacidad de resumen. Además, ni siquiera se trabaja lo suficiente el resumir los textos, poco comprensible, ya que como indica Sanz (2005) resumir en sí es una estrategia muy útil para mejorar la comprensión lectora y la expresión escrita. Ante esta situación, utilizar una evaluación basada en crowd-sourcing y emplear técnicas de Procesamiento de Lenguaje Natural (PLN) puede ser de gran ayuda para que profesores y estudiantes evalúen de manera automática los resúmenes y obtengan un *feedback* que les haga mejorar su capacidad de resumen.

El objetivo de este trabajo es lograr modelos fiables mediante crowd-sourcing que sirvan para: i) comparar los resúmenes elaborados por alumnos de diferentes edades y cursos educativos (primaria y universidad), ii) evaluar los resúmenes modelo creados en clase (evaluación colectiva de la clase) y iii) evaluar a cada alumno (evaluación individual). Según los resultados, cuando los alumnos de primaria resumen textos previamente trabajados en clase, la capacidad de resumir que muestran en los textos más fáciles es muy similar a la de los universitarios. Sin embargo, cuando se trata de resumir textos más complejos, queda patente que los universitarios muestran una capacidad superior. En cuanto al método, el cual se ha elaborado basándose en aspectos cualitativos (estructura discursiva de la coherencia) y cuantitativos (kappa de Fleis y distancia de Hamming), los resultados indican que el método empleado para comparar tanto grupos como individuos es adecuado.

En cuanto a la evaluación de resúmenes, este siempre ha sido un tema complejo y controvertido en la lingüística computacional (Saggion et al., 2010). ROUGE (CY, 2004), BLEU (Papineni et al., 2002), Pyramid (Nenkova y Passonneau, 2004) y SummTriver (Cabrera-Diego y Torres-Moreno, 2018), son métodos que permiten evaluar resúmenes de manera automática. Sin embargo, tal y como lo explican Molina y Torres (2015), muchas de estas métricas pueden acarrear ciertas desventajas, ya que resúmenes no gramaticales pueden obtener puntuaciones muy altas. Además, no se debe olvidar que algunos de estos métodos necesitan modelos de referencia hechos por expertos, lo cual es un problema a la hora de aplicarlo en un escenario real como puede ser la escuela. Por un lado, los profesores carecen de tiempo para poder crear resúmenes modelo de varios textos y después corregir todos los trabajos realizados (todos los resúmenes de 25-35 estudiantes). Por otro lado, como bien indican Radev y Tam (2003), existe más de un modelo adecuado para realizar un resumen, pudiendo ser dos resúmenes igualmente buenos aun estando elaborados con frases totalmente distintas. Esto muestra la necesidad de contar en el aula con varios modelos de un mismo resumen, labor que es difícilmente realizable por un único profesor. Por lo tanto, es necesario trabajar con varios modelos de referencia, para poder ofrecer un modelo tanto para la máquina, como para dar un *feedback* adecuado al docente y estudiante. La medida *Relative Utility* (Radev y Tam, 2003) utiliza más de un modelo de referencia para evaluar los resúmenes. Aun así, poder lograr varios modelos de referencia de todos los textos curriculares, es una tarea dificil para el docente, lo que deja en evidencia las limitaciones que puede llegar a tener dicho modelo.

En consecuencia, además de los desafíos habituales de la comprensión de un texto, es fundamental adaptar las herramientas, y así proporcionar un entorno propicio para el aprendizaje (de los estudiantes) y la evaluación (para los profesores). Para ello, es necesario tener en cuenta los recursos y las limitaciones (lingüísticas y no lingüísticas) que existen en las escuelas vascas si se quiere crear un entorno eficiente para desarrollar habilidades de síntesis y procedimientos de evaluación.

2 Estado de la cuestión

Como indica Molina-Villegas (2013b), una posible solución para recopilar resúmenes de referencia es recurrir a la participación de voluntarios no expertos en tareas científicas que no requieran experiencia en el tema. Invo-

lucrar ciudadanos en actividades científicas puede ser interesante y útil, ya que permite ahorrar recursos humanos; aumentar la cantidad y velocidad del procesamiento de datos; o simplemente acercar las personas a la ciencia. Molina-Villegas (2013b) desarrolló un sistema de crowd-sourcing para compilar un corpus de resumen automático en español donde se obtuvieron cerca de tres mil resúmenes en tan solo 10 semanas, lo que demuestra su utilidad para reunir un corpus amplio en un espacio de tiempo muy concreto. Siendo considerables las ventajas, es imprescindible centrarse en las posibles desventajas, para poder después contrarrestarlas. La principal desventaja es que los participantes pueden llegar a actuar al azar, o es posible que carezcan de criterios específicos. En consecuencia, en este artículo trataremos de solventar estas desventajas utilizando la prueba exacta de Fisher, para verificar si los participantes han respondido al azar, y el coeficiente kappa para comprobar si los participantes han respondido con criterios similares a los de un evaluador.

En cuanto al resumen automático, y en concreto al resumen de extracción, en palabras de Saggion y Poibeau (2013) la selección de oraciones de un resumen extractivo puede basarse en información estadística o en una teoría que tenga en cuenta la información lingüística y semántica. Por tanto, para realizar un resumen extractivo automático se utilizan técnicas superficiales y enfoques basados en el conocimiento. Molina-Villegas (2013a) por ejemplo, presenta un estudio sobre la comprensión de oraciones y propone un modelo de regresión que predice la eliminación de segmentos dentro de la oración con aplicación en la generación de resumen abstractivo. En cuanto a herramientas se refiere, a día de hoy es posible encontrar varias herramientas que permiten generar resúmenes de manera automática, a destacar MEAD (Radev et al., 2004) y SUMMA (Saggion, 2008). Estas herramientas, junto con otras, dan opción a trabajar con lenguas como el inglés o el chino, por ejemplo. Pero hasta donde nosotros sabemos, no hay ninguna herramienta de resumen automático que emplee información en euskera. Esto se debe a que estas herramientas no utilizan analizadores morfológicos, los cuales son necesarios para poder tratar con una lengua aglutinante como el euskera.

Así, Uno de los retos principales de este trabajo era reunir un corpus amplio de extracciones. Para lograrlo, se ha utilizado Compress-eus (Atutxa et al., 2017) que recopila el corpus de extracciones y abstracciones de los resúmenes realizados de textos en euskera. Además, esta herramienta ofrece información automática de las extracciones.

En cuanto a la evaluación humana, Atutxa (2018) propone criterios de evaluación para evaluar extracciones y abstracciones en euskera. Para ello se basan en la guía BABAR (Álvarez, 2004) y en la información que proporciona la estructura relacional de discurso (coherencia). La guía BABAR tiene como objetivo evaluar textos expositivos en inglés como lengua extranjera, y en ella se evalúan los siguientes apartados: contenido, organización, vocabulario, uso de la lengua y presentación.

Evaluar un resumen no es una tarea sencilla, ya que requiere evaluar un proceso complejo. Si la evaluación de resúmenes es compleja para los humanos, la evaluación automática lo es aún más, especialmente en lo que respecta a la fiabilidad. En la evaluación automática, Saggion y Poibeau (2013) diferencian dos tipos de evaluación:

i) Evaluación de un resumen automático comparándolo con *gold standards* creados manualmente: el resumen elaborado es evaluado comparándolo con un resumen de referencia creado por un humano. Por lo que nos consta, es este el método utilizado por Zipitria, Arruarte, y Elorriaga (2008) con LEA (una aplicación web para la evaluación de resúmenes en euskera). Zipitria et al. (2008) explica el método utilizado en su trabajo. Primero, observan cómo se realiza la evaluación humana y, luego, automatizan sus observaciones. 15 expertos tuvieron que resumir 5 textos para crear un modelo para la evaluación de los resúmenes. Los expertos estaban formados por cinco profesores de secundaria, cinco profesores de euskera de segunda lengua (L2) y cinco profesores universitarios.

ii) Evaluación de resúmenes automáticos sin *gold standard*: para evaluar se utiliza la información que contiene el propio resumen. Louis y Nenkova (2009) presentaron un marco para evaluar el contenido usando el input como referencia. Se basa en el hecho de que la distribución de palabras en el input (texto a resumir) y el resumen de ese *input* deben ser similares. En el caso del euskera, no se conoce

ningún trabajo relacionado con la evaluación de resúmenes automáticos sin ningún corpus *gold standard* para comparar.

3 Materiales y método

3.1 Compress-eus: una herramienta para recopilar resúmenes

Uno de los principales problemas para trabajar en una lengua como el euskera suele ser no contar con un corpus adecuado. Para ello, es muy útil contar con una herramienta de características similares a las de Compress-eus. Compress-eus es una interfaz que permite reunir resúmenes elaborados por alumnos y profesores. El texto original está segmentado en unidades elementales del discurso (EDU) y la idea más importante o la Unidad Central (UC) está anotada. El alumno realiza el resumen extractivo, eliminando del texto las EDUs que considere oportunas, es decir, los segmentos menos relevantes. Cuando finaliza la extracción, el usuario guarda el resumen. El uso de esta interfaz facilita la recopilación y el análisis del resúmenes. Además, facilita utilizar información jerárquica de la estructura discursiva de los textos originales, pero también de los resúmenes extractivos, lo que será de gran ayuda para este trabajo.

Compress-eus proporciona el seguimiento de todas las operaciones realizadas por el usuario (qué EDU se eliminó, si se eliminó la CU, por ejemplo) al realizar la extracción. Además, facilita la siguiente información sobre los textos a resumir: número de párrafos (en el texto); número de oraciones (en el texto y cada párrafo); número de EDUs (en el texto, cada párrafo y cada oración).

3.2 Corpus

Para este trabajo se ha reunido un corpus compuesto por 1036 resúmenes de extracción. 352 son de alumnos de quinto año de primaria (9-10 años), 88 estudiantes han resumido 4 textos cada uno. Los 684 resúmenes restantes han sido realizados por alumnos universitarios de la Facultad de Educación de Bilbao, en este caso han sido 171 alumnos quienes han resumido los 4 textos. Para tener un escenario real de enseñanza, fueron utilizados textos procedentes de libros de texto que se emplean en la escuela para llevar a cabo el programa curricular. De hecho, el alumno tuvo que resumir los textos seleccionados en las fechas reales programadas para ello. Los detalles de los textos utilizados para los experimentos se describen en la Tabla 1.

	Párrafos	Oraciones	EDUs	Pals.	UC
T1	5	11	17	121	1
T2	4	11	23	131	1
T3	10	17	37	218	2
T4	11	25	41	289	1

Tabla 1: Características del corpus.

3.3 Método

La metodología que se presenta para describir i) la evaluación entre etapas escolares, ii) la evaluación de la clase y iii) la evaluación individual se divide en tres pasos principales: 1) generación de meta-resúmenes (modelos de referencia), 2) armonización e inclusión de resúmenes (si fuese necesario) y 3) medidas de evaluación adecuadas.

Se propone el siguiente algoritmo para generar meta-resúmenes, lo cual permite comparar la diferencia que pueda existir entre los resúmenes elaborados por estudiantes de primaria y los universitarios.

Sea $E = \{\boldsymbol{e_1}, \ldots, \boldsymbol{e_n}\}$ el conjunto de n resúmenes de referencia (*e.g.* los resúmenes de los universitarios); tal que cada resumen está codificado por una tupla de m elementos binarios representando la presencia/ausencia de las m unidades discursivas del documento original: $\boldsymbol{e_i} = (e_{i1}, \ldots, e_{im})$. Es decir, para cada unidad discursiva del documento original hay una entrada en la tupla que tendrá asignado un 1 cuando la unidad se preservó en el resumen y 0 cuando la unidad se eliminó en el resumen. Note que el orden de las oraciones no cambia sino que algunas aparecerán en el resumen y algunas otras no. Siguiendo esta representación se han codificado tanto los resúmenes de estudiantes de primaria como los de estudiantes universitarios. La idea general es codificar un solo meta-resumen de referencia a partir de los n resúmenes de referencia, en principio distintos, pero hacerlo de manera tal que el meta-resumen generado maximice el acuerdo entre las referencias y por lo tanto se pueda considerar como un resumen modelo.

El **Algoritmo 1**, de complejidad $\mathcal{O}(n)$, resuelve el problema de encontrar un subconjunto $E^* \subseteq E$ tal que el acuerdo entre sus elementos sea máximo; donde el acuerdo está determinado por el coeficiente kappa κ de Fleiss (2013). Al inicializar el algoritmo se incluye solamente la primera tupla de

E. Luego, en cada iteración se calcula el valor de κ que se produciría al incluir una nueva tupla de E en el subconjunto óptimo E^*; si el nuevo elemento aumenta el acuerdo entre las referencias, entonces es incluido en el subconjunto óptimo. Una vez realizadas todas las comparaciones, el resumen modelo de referencia se construye a partir de la moda de las entradas de las tuplas que hayan sido incluidas en E^*.

Algorithm 1 Algoritmo de generación de resúmenes modelo

procedure CREATEMODEL(E)
  $E^* \leftarrow \{\boldsymbol{e_1}\}$
  $\kappa* \leftarrow 0{,}0$
  for $\boldsymbol{e_i}$ in $E - \{\boldsymbol{e_1}\}$ **do**
    $\kappa \leftarrow FleissKappa(E^* \cup \{\boldsymbol{e_i}\})$
    if $\kappa \geq \kappa*$ **then**
      $\kappa \leftarrow \kappa*$
      $E^* \leftarrow E^* \cup \{\boldsymbol{e_i}\}$
    end if
  end for
  $model \leftarrow (mode(e_{ij})$ for $\boldsymbol{e_i} \in M; \boldsymbol{e_i} = (e_{i1}, \ldots, e_{im}))$
  return $model$
end procedure

4 Resultados

4.1 Evaluación entre distintas etapas escolares

Una característica del algoritmo es que el meta-resumen generado puede ser diferente dependiendo del orden de lectura de las tuplas $\boldsymbol{e_i}$. A mayor variabilidad entre los criterios de los alumnos, habrá una mayor variedad de meta-resumenes posibles. La Tabla 2 muestra que el algoritmo ha creado 63 meta-resúmenes con los resúmenes de primaria y 60 con los de la universidad. Se han creado más meta-resúmenes en primaria teniendo una cantidad sensiblemente inferior de resúmenes (352 de primaria y 684 de la universidad). En cuanto a los textos, el algoritmo ha creado 85 meta-resúmenes que corresponden al Texto-4, y solamente 36 con el Texto-2. Esto refleja que la longitud y la estructura discursiva del texto tienen mayor impacto (en cuanto a cantidad se refiere) al crear meta-resúmenes. Además, los meta-resúmenes de primaria se distribuyen de manera más equilibrada. Se ha creado un mínimo de 14 meta-resúmenes (Texto-1 y Texto-2), y un máximo de 18 (Texto-3). Sin embargo, con los resúmenes de los universitarios, la cantidad de los meta-resúmenes creados tiene una variación sensiblemente superior. Con el Texto-1 (17 EDUs) y Texto-2 (23 EDUs) se han creado menos que con el Texto-3 (37 EDUs) y Texto-4 (41 EDUs). Esto refleja que cuantos más EDUs tenga el texto, más meta-resúmenes creará el algoritmo con los resúmenes de los universitarios.

	T1	**T2**	**T3**	**T4**	**Total**
Primaria	17	14	18	14	63
Universidad 2018-2019	10	5	19	25	59
Universidad 2019-2020	7	8	18	22	55
Universidad 2018-2020	10	9	17	24	60
Total	44	36	72	85	

Tabla 2: Cantidad de meta-resumenes de resúmenes extractivos creados por el Algoritmo-1.

Para analizar y comparar los meta-resúmenes de primaria y universidad, el método se basa en las siguientes tres variables: i) cantidad de resúmenes en un meta-resumen, ii) acuerdo kappa obtenido y iii) puntuación obtenida en la coherencia (acuerdo entre la estructura retórica descrita por el profesor y la estructura retórica que se obtiene de la extracción hecha por el estudiante).

La coherencia fue calculada con el siguiente procedimiento:

- Paso-1: el texto se segmentó en EDUs y la unidad que contiene la idea principal (UC) fue etiquetada según Iruskieta, Diaz de Ilarraza, y Lersundi (2014). A continuación, el texto fue anotado manualmente según la Teoría de la Estructura Retórica (RST) (Mann y Thompson, 1987).
- Paso-2: Cada texto resumido se dividió en 4 cuartiles (Q1 a Q4) siguiendo la estructura jerárquica del árbol RST. Se calcula la distancia de cada EDU contando cuántas relaciones se necesitan para llegar a la UC de los resúmenes. La EDU más cercana está a la distancia 0 (la misma UC) y la EDU más lejana está a 6 relaciones de distancia. Pero, en algunos textos, la distancia más lejana es de 4 relaciones. Una vez obtenida la distancia de cada EDU, se clasifican las EDUs en cuartiles. Por ejemplo, el árbol RST con una distancia máxima de 6 es el siguiente: *i*) Q1 representa la UC y todas

las EDUs a la distancia 1 (distancia 0-25 %). *ii*) Q2 representa todas las EDUs a 2 y 3 relaciones de distancia, lo que significa que se saltan 2 o 3 relaciones desde la EDU en que se sitúa la UC (distancia 26-50 %). *iii*) Q3 representa todas las EDUs a 4 y 5 relaciones de distancia (distancia 51-75 %). *iv*) Q4 representa el resto de las EDUs a 6 y 7 relaciones de distancia (distancia 76-100 %).

- Paso-3: Puntuación de la coherencia. Por un lado, para representar y comparar la calidad de los meta-resúmenes, todas las EDUs son clasificadas en grupos según su distancia respecto a la UC. Después, se ha calculado el porcentaje de EDUs mantenidas en cada grupo, para poder aplicar las siguientes reglas: *i*) regla aplicada a EDUs de nivel 1 (distancia 0): si el porcentaje de las ideas mantenidas es igual a 100 %, entonces 1 punto, si no 0. *ii*) Regla aplicada a EDUs de nivel 2 (distancia 1): si el porcentaje de las ideas mantenidas es igual o menor al porcentaje de las ideas del nivel previo (siendo este superior al 0 %), entonces 1 punto, si no 0. *iii*) Regla aplicada al resto de niveles: si el porcentaje de las ideas mantenidas es igual o menor al porcentaje de las ideas del nivel previo y menor al porcentaje de las ideas del nivel anterior al previo, entonces 1 punto, si no 0.

  Por otro lado, cada cuartil se ha ponderado en base a su relevancia: *i*) 0,4 para la puntuación obtenida en Q1. *ii*) 0,3 para el Q2. *iii*) 0,2 para el Q3. *iv*) 0,1 para el Q4.

De esta forma, es posible representar y visualizar los resultados empleando gráficos de burbuja. Estos gráficos ayudan a comparar las distintas etapas escolares.

La Figura 1 muestra la calidad de los meta-resúmenes creados en ambas etapas escolares. La gran mayoría tiene una puntuación superior al 0,6, lo cual es comprensible. Por una parte, es normal que los estudiantes universitarios no tengan grandes dificultades para trabajar textos que pertenecen a libros de texto de primaria. Por otra parte, los alumnos de primaria han trabajado con estos textos a lo largo de toda la unidad didáctica, lo cual ha podido facilitar la acción de resumir.

Las burbujas que representan a los alumnos de la universidad son más grandes, lo cual indica que la cantidad de resúmenes que contienen es superior. Se podría pensar que los estudiantes de la universidad resumen de manera mucho más similar entre ellos si se comparan con los estudiantes de primaria. Sin embargo, se han analizado 88 estudiantes de primaria y 171 universitarios; por lo tanto, era de esperar que se generasen burbujas más grandes en la universidad.

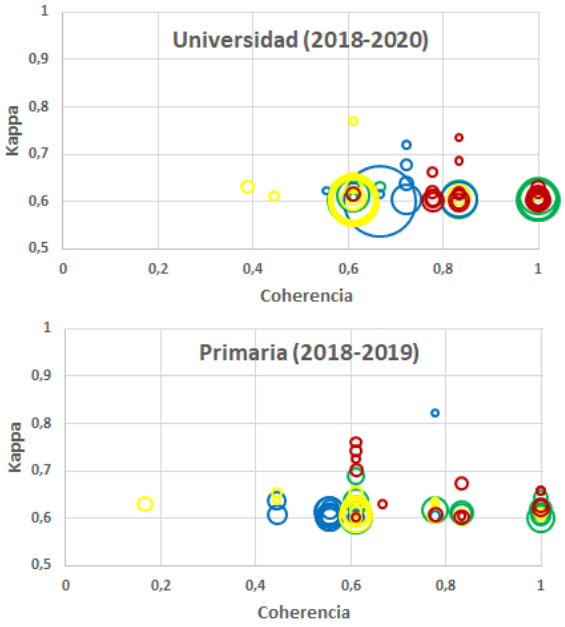

Figura 1: Cantidad y calidad de los meta-resúmenes creados por el algoritmo con estudiantes de universidad y primaria. El eje horizontal muestra la puntuación de coherencia (Coherencia) obtenida. El eje vertical muestra acuerdo entre los anotadores y resúmenes (Kappa). El tamaño de la burbuja refleja el número de resúmenes que contiene el meta-resumen. Las burbujas verdes representan los resultados del T1, azules T2, amarillas T3 y rojas T4.

Si se compara el tamaño de los meta-resúmenes creados en el Texto-2 (burbujas azules), la Figura 1 muestra que los universitarios tienen una manera muy concreta de resumir dicho texto. Este fenómeno no se da entre los alumnos de primaria. La gran cantidad de meta-resúmenes creados con el Texto-2 (burbujas azules), con un tamaño similar entre ellos, indica una gran diversidad a la hora de resumir este texto. Si se observa el eje de la coherencia, se puede intuir que los estudiantes (tanto en primaria como universidad)

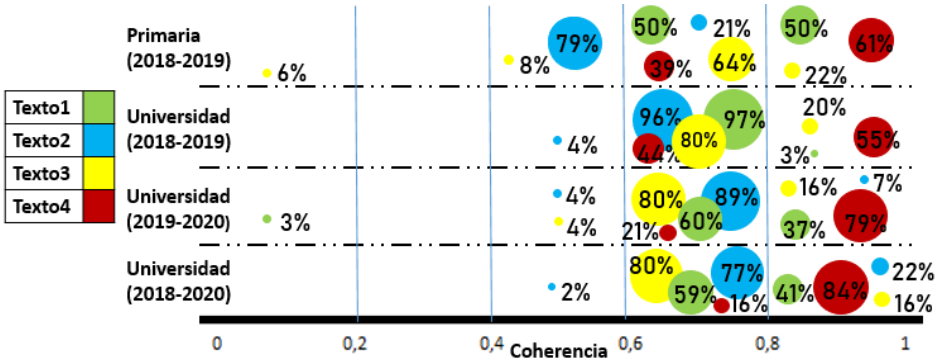

Figura 2: Calidad de los meta-resúmenes de los estudiantes de primaria y universidad. El eje horizontal muestra la puntuación de coherencia mientras que el eje vertical muestra los diferentes niveles escolares que se han analizado. Los tamaños de las burbujas muestran cuántos resúmenes hay en cada área de la cuadrícula.

han tenido mayor facilidad a la hora de resumir el Texto-1 y el Texto-4. Por el contrario, los meta-resúmenes creados con los textos 2 y 3 han obtenido una puntuación menor en general.

Para describir de manera más concisa esta diferencia, se utilizara el eje horizontal que muestra la Figura 2.

La Figura 2 indica que los estudiantes han tenido mayor facilidad con el Texto-1 y Texto-4, especialmente en este último; ya que todos los meta-resúmenes han logrado una puntuación de coherencia que se sitúa entre 0,6 y 0,8 o entre 0,8 y 1. En cambio, el algoritmo ha creado meta-resúmenes de menor calidad con los textos 2 y 3. En este caso, la diferencia entre los alumnos de universidad y primaria es considerable. En el Texto-3, la mayoría de los meta-resúmenes de los estudiantes tienen una puntuación de coherencia de 0,6 y 0,8, sin embargo, aparecen algunos con una puntuación más baja. Estos últimos pertenecen principalmente a los estudiantes de primaria, ya que como se muestra en la Figura 2 el 8 % de los estudiantes se ubica entre 0,4 y 0,6 y el 6 % entre 0 y 0,2. En cuanto al Texto-2, la mayoría de los estudiantes universitarios han obtenido una puntuación entre 0,6 y 0,8; sin embargo, la mayoría de los estudiantes de primaria se ubican entre 0,4 y 0,6. Esto podría deberse a que cuando los estudiantes resumen textos de poca complejidad, no existe una gran diferencia entre los resultados de los estudiantes de universidad y primaria. Pero cuando se trabaja con textos que son más difíciles de resumir, los estudiantes universitarios tienen más recursos para resumir el texto, a pesar de que los estudiantes de primaria hayan trabajado previamente estos textos.

Los estudiantes del curso 2019-2020 han logrado en general mejores resultados que los del curso 2018-2019, concretamente con los textos 1, 2 y 4. El ejemplo más evidente se da con el Texto-1. El 3 % de los meta-resúmenes pertenecientes al curso 2018-2019 han logrado una puntuación de coherencia entre 0,8 a 1; sin embargo, este porcentaje sube hasta el 37 % en el curso 2019-2020. Cuando se tienen en cuenta ambos cursos de la universidad, el porcentaje de estudiantes que se incluyen en un meta-resumen y obtienen una puntuación de coherencia de 0,8 a 1 sube hasta el 41 %. Este fenómeno se da en 3 de los 4 textos, lo que deja ver que cuanto mayor es el corpus mayor es la calidad de los meta-resúmenes en cuanto a la coherencia. Por lo cual, el sistema identifica más meta-resúmenes y meta-resúmenes mucho más fiables cuando hay más resúmenes. Sin embargo, la calidad es un factor fundamental. El sistema encontró mejores meta-resúmenes en el curso 2019-2020 con 82 resúmenes que en el 2018-2019 con 89 resúmenes. Esto parece indicar que ambos factores: el número de resúmenes y la calidad de los resúmenes son necesarios para crear meta-resúmenes fiables.

Aunque los resultados expuestos hasta el momento sirvan para detectar las diferencias principales entre alumnos de distintas etapas escolares, no se pueden utilizar estos datos para hacer una evaluación individual de los estudiantes y una evaluación colectiva de la clase. Las dos limitaciones que se exponen a continuación hacen que la evaluación deje a un lado algunos resúmenes (cobertura) y es

necesario que la precisión sea la mayor posible, para que la evaluación sirva en el proceso de aprendizaje en la escuela.

- Baja representación de los estudiantes: los meta-resúmenes tienen que tener un valor mínimo de 0,6 en kappa (valor establecido por los autores). Los resúmenes que no logren ese valor quedan fuera, y cabe la posibilidad de que un resumen no se incluya en ningún meta-resumen. Por lo tanto, si se quiere evaluar la clase en su conjunto con los meta-resúmenes, se deberá solventar esta limitación.

- El acuerdo entre la calidad del resumen y el meta-resúmenes: los meta-resúmenes fueron creados teniendo en cuenta primero el acuerdo de los segmentos del texto con kappa (medición cuantitativa) y después se calculó la puntuación de coherencia del meta-resumen (medición cualitativa). Por tanto, puede surgir el problema de que dos resúmenes distintos acaben estando en el mismo meta-resumen siendo cualitativamente muy diferentes entre sí (pudiendo ser que la única diferencia sea que uno contenga la idea más importante y el otro no). En ese caso, al menos uno de los dos resúmenes no está debidamente incluido en ese meta-resumen, en cuanto a la medida de coherencia. En consecuencia, es indispensable descartar los resúmenes que no se ajustan al meta-resumen.

4.2 Evaluación de la clase

Para la evaluación de toda una clase, el primer paso es cerciorarse de resolver las limitaciones mencionadas previamente. Para saber si un resumen está correctamente representado en un meta-resumen e incluir los resúmenes que habían quedado fuera de los meta-resúmenes, se utilizan la puntuación de coherencia (C) y la distancia de Hamming (H) para armonizar los meta-resúmenes: método C+H (criterios de coherencia y Hamming). Para incluir o mantener cualquier resumen en el meta-resumen, la distancia de coherencia (C) entre resumen y meta-resumen debe estar entre -0,1 y 0,1 (20 % de la puntuación) y la distancia de Hamming (H) debe ser inferior a 0,2 (20 % de la puntuación).

En la Figura 3 se presentan los meta-resúmenes y los resúmenes del Texto-1. El

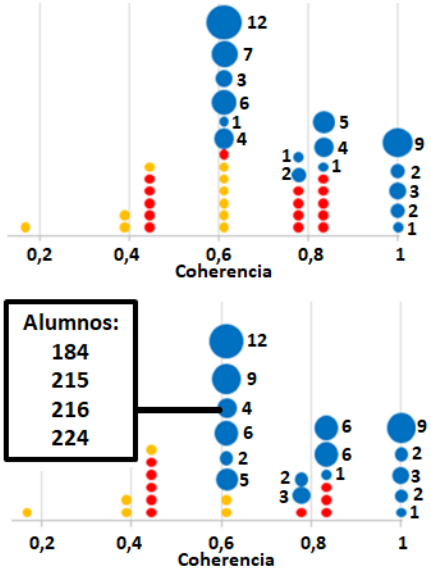

Figura 3: Evaluación del aula de primaria con el método de Coherencia y Hamming (C+H). Las burbujas azules muestran los meta-resúmenes creados por el algoritmo y la cantidad de resúmenes incluidos en ellos. Las burbujas amarillas representan cada resumen no incluido en ningún meta-resumen. Las burbujas rojas representan resúmenes incluidos en meta-resúmenes que no cumplen con los criterios del método C+H.

algoritmo ha creado 17 meta-resúmenes (burbujas azules), los cuales constituyen el 72 % de los resúmenes. El algoritmo no ha incluido 10 resúmenes (burbujas amarillas) para crear meta-resúmenes, es decir, el 11 % de los resúmenes había quedado fuera. Los últimos 15 resúmenes (burbujas rojas) (17 % de los resúmenes) según el método C + H, no estaban correctamente representados en los meta-resúmenes donde habían sido incluidos por el algoritmo.

Tras aplicar el método C+H, los resultados se muestran en la Figura 3. En esta figura se muestran los resúmenes no incluidos en los meta-resúmenes (burbujas amarillas), los resúmenes que se han mantenido en los meta-resúmenes después de aplicar el método C+H (burbujas azules) y los resúmenes que se han dejado fuera de los meta-resúmenes después de aplicar el método C+H (burbujas rojas). En el gráfico inferior se han añadido en los meta-resúmenes (burbujas azules), los resúmenes que anteriormente se habían dejado fuera (burbujas rojas y amarillas), pero que cumplen los criterios del método C+H. Al aplicarlo, el 40 % (4 de 10) de las burbujas

amarillas se han incluido en los meta-resúmenes. Por otra parte, el 60 % (9 de 15) de los resúmenes excluidos (burbujas rojas) se han incluido en otros meta-resúmenes.

La mayoría de los estudiantes está en torno al 0,6 en cuanto a la coherencia. Son 8 meta-resúmenes diferentes los que se sitúan en ese punto (0,6), conteniendo un total de 41 estudiantes (46 % de los estudiantes). Estos 41 estudiantes han mostrado una capacidad similar de resumir el Texto-1. Sin embargo, el docente tiene que tener en cuenta que aun habiendo logrado la misma puntuación, el sistema creó 8 meta-resúmenes diferentes. Esto significa que puede haber estudiantes que hayan obtenido la misma nota, pero que hayan utilizado estrategias totalmente diferentes para resumir. Por tanto, el docente deberá analizar cuáles son esas estrategias y cómo trabajarlas pudiendo dar un *feedback* personalizado.

Además del ya mencionado grupo de alumnos, los resultados muestran otros tres grupos: i) algunos de estos estudiantes (17 de 88) han alcanzado la máxima puntuación C. ii) Otros (16 de 88) están alrededor de 0,8 C. Se puede decir que todos estos estudiantes han resumido fácilmente el texto y han obtenido buenos resultados. iii) Sin embargo, hay otro grupo (8 de 88) que obtuvo una C más baja, alrededor de 0,4. Estos estudiantes han tenido más dificultades para resumir el Texto-1. iv) Para concluir, un estudiante ha obtenido una puntuación C (0,2) muy baja, por lo que el docente podría preocuparse.

Este tipo de gráficos ofrecen al docente una visión general para analizar en qué grupo se encuentra cada alumno y esto puede ser muy útil para programar actividades futuras. Por ejemplo, el profesor puede saber qué alumno puede ayudar a otros y quién necesita ayuda, o quién tiene que trabajar con textos más simples o más complejos. Además, puede indicarle qué resúmenes puede utilizar con los alumnos como modelo a seguir o cuales pueden servirle para emplearlos como *feedback*.

4.3 Evaluación automática de los estudiantes

El algoritmo creó 17 meta-resúmenes, de los cuales hay que decidir cuales utilizar para la evaluación de cada estudiante. Para decidirlo, se han seguido los dos siguientes criterios: i) el número de resúmenes contenidos en el meta-resumen es al menos el 10 % total de los resúmenes y ii) el modelo alcanza al menos una puntuación de 0,5 (C). Los meta-resúmenes A, B, C, D, E y F (ver figura 4) cumplen los criterios ya mencionados, por lo cual, se emplearan para evaluar a los estudiantes de manera automática.

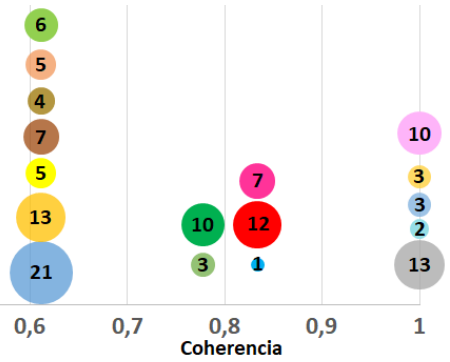

Figura 4: Meta-resúmenes utilizados para la evaluación de estudiantes de primaria. El eje horizontal muestra la puntuación de la coherencia. El tamaño de las burbujas muestra cuántos resúmenes hay en cada meta-resumen. Los números de cada burbuja muestran la cantidad de resúmenes que contiene el meta-resumen.

Cada resumen es evaluado frente a los 6 meta-resúmenes seleccionados previamente mediante kappa (para calcular el acuerdo) y Fisher (para calcular la probabilidad con la que el resumen haya sido hecho al azar). En cuanto al acuerdo, un valor Kappa superior a 0,6 se considera óptimo. Por otra parte, el resumen extractivo es una acción subjetiva, por tanto, se considerara que se ha realizado un resumen de forma aleatoria si obtiene un valor de Fisher de 0,8 o superior.

La Tabla 3 muestra los resultados obtenidos por los estudiantes 265, 206 y 214 después de evaluar automáticamente sus resúmenes con los seis meta-resúmenes seleccionados.

		Modelos					
Est.		**A**	**B**	**C**	**D**	**E**	**F**
265	K	0,26	0,33	0,10	0,46	0,84	0,72
	F	0,29	0,23	1,00	0,12	0,02	1,00
206	K	0,33	1,00	0,45	0,45	0,26	0,20
	F	0,23	0,59	0,17	0,17	0,29	0,51
214	K	-0,18	-0,08	-0,16	-0,16	-0,20	-0,21
	F	1,00	1,00	1,00	1,00	1,00	0,09

Tabla 3: Evaluación de los estudiantes 265, 206 y 214 mediante la comparación con los 6 meta-resúmenes utilizando Kappa y Fisher.

El resumen 265 ha logrado un gran acuerdo Kappa con los meta-resúmenes E (0,84) y

F (0,72). Es una buena señal, ya que el resumen 265 se ha incluido previamente en ambos meta-resúmenes. Además, si se calcula la coherencia del resumen 265, la puntuación es 1, igual que los dos meta-resúmenes (ver Figura 4). De ahí que, aunque el resumen sea diferente al de los dos meta-resúmenes, tienen la misma calidad. Sin embargo, se debe tener en cuenta la métrica estadística de Fisher, para saber si el resumen se ha hecho al azar. En el Meta-resumen-E el valor de Fisher es de 0,02 (inferior a 0,8), lo cual indica que el resumen no se hizo al azar.

El resumen 206 es idéntico al Meta-resumen-B, ya que han alcanzado la máxima puntuación en Kappa y, además, el valor de Fisher es bajo (0,59). El Meta-resumen-B tiene una puntuación de coherencia de 0,61 (consulte la Tabla 3); por lo tanto, el resumen 206 también debería alcanzar esa puntuación, ya que son iguales. Cabe señalar que el resumen 206 se ha incluido previamente en el Meta-resumen-B, pero no en el A. Los dos meta-resúmenes obtuvieron la misma puntuación de coherencia, esto indica que los meta-resúmenes A y B tienen la misma calidad pero son bastante diferentes.

Al evaluar el resumen del estudiante 214, (Tabla 3), el acuerdo alcanzado con los meta-resúmenes ha sido prácticamente nulo. Además, los valores de Fisher indican una alta probabilidad de haber hecho el resumen al azar. A pesar de todo, es importante recordar lo siguiente. Cuando un resumen logra un gran acuerdo con un meta-resumen, se deduce que el resumen es bueno. Sin embargo, cuando el acuerdo es bajo, no es posible garantizar que sea malo. En este caso, se calculó la puntuación del resumen 214 manualmente (0,16 C), para comprobar que el resumen es de baja calidad. En futuros trabajos será necesario incluir como modelos de referencia meta-resúmenes de baja calidad.

5 Conclusiones

En este artículo hemos presentado un método para la evaluación automática de resúmenes por extracción en euskera, aunque el sistema y él método podrían usarse en otras lenguas con una mínima adecuación. En el método se hace uso del crowd-sourcing para crear meta-resúmenes de referencia y se ha utilizado para realizar tres tipos de evaluación: i) diferencias entre distintas etapas escolares, ii) evaluación colectiva del aula y iii) evaluación automática individual de cada estudiante usando meta-resúmenes fiables. En los experimentos se ha trabajado con alumnos reales y utilizado sus textos escolares, es decir, textos que son utilizados para trabajar contenidos curriculares en la escuela en tiempo real. Se ha propuesto el método C+H, permitiendo evaluar el acuerdo cualitativo (coherencia) y cuantitativo (Hamming) entre resúmenes y meta-resúmenes.

Los resultados demuestran que la evaluación propuesta es viable y robusta para aplicarse a mayor escala; lo cual, nos permitirá seguir trabajando en un contexto real pero con intervenciones más largas y sistemáticas en el tiempo, además de incluir diferente tipología de textos para dar el salto a los resúmenes de abstracción.

En futuros trabajos sería interesante analizar las relaciones discursivas para después poder ponderarlas. De esta forma, se podría desarrollar un algoritmo que cree meta-resúmenes basándose en un acuerdo cuantitativo y cualitativo. También será importante trabajar con más documentos pues permitiría analizar qué tipos de textos son más fáciles de resumir para los estudiantes y qué texto es conveniente para desarrollar las habilidades que requiere el resumen. Para ello, puede resultar de gran utilidad analizar tanto la distribución de la información más importante en el texto como la distribución de las relaciones de coherencia.

Agradecimientos

El trabajo de Unai Atutxa está financiado por una beca de doctorado (PIF18/118) de la Universidad del País Vasco (UPV/EHU).

Bibliografía

Álvarez, I. A. 2004. Evaluación y calificación de resúmenes de textos expositivos en el aula de ile/ife: la guía"babar". *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, 1(8):81–99.

Atutxa, U. 2018. Ikasleen laburpen-corpusa eta laburpen-gaitasunaren ebaluazioa: oinarri metodologikoak. Master's thesis, University of the Basque Country (UPV/EHU), Donostia.

Atutxa, U., M. Iruskieta, O. Ansa, y A. Molina. 2017. Compress-eus: I (ra) kasleen laburpenak lortzeko tresna. *EU-*

DIA: Euskararen bariazioa eta bariazioaren irakaskuntza-III, páginas 87–98.

Cabrera-Diego, L. A. y J.-M. Torres-Moreno. 2018. Summtriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering*, 113:184 – 197.

CY, L. 2004. Rouge: a package for automatic evaluation of summaries. En *Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain*, páginas 56–60.

Fleiss, J. L., B. Levin, y M. C. Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Iruskieta, M., A. D. 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 *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 466–475.

Louis, A. y A. Nenkova. 2009. Automatically evaluating content selection in summarization without human models. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, páginas 306–314. Association for Computational Linguistics.

Mann, W. C. y S. A. Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Molina, A. y J.-M. Torres. 2015. El test de turing para la evaluación de resumen automático de texto. *Linguamática*, 7(2):45–55.

Molina-Villegas, A. 2013a. Compresión automática de frases: un estudio hacia la generación de resúmenes en español. *Inteligencia Artificial*, 16(51):41–62.

Molina-Villegas, A. 2013b. Sistemas web colaborativos para la recopilación de datos bajo el paradigma de ciencia ciudadana. *Komputer Sapiens*, 1(5):6–18.

Nenkova, A. y R. J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. En *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, páginas 145–152.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 311–318. Association for Computational Linguistics.

Radev, D. R., H. Jing, M. Styś, y D. Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Radev, D. R. y D. Tam. 2003. Summarization evaluation using relative utility. En *Proceedings of the twelfth international conference on Information and knowledge management*, páginas 508–511.

Saggion, H. 2008. A robust and adaptable summarization tool. *Traitement Automatique des Langues*, 49(2).

Saggion, H. y T. Poibeau. 2013. Automatic text summarization: Past, present and future. En *Multi-source, multilingual information extraction and summarization*. Springer, páginas 3–21.

Saggion, H., J.-M. Torres-Moreno, I. d. Cunha, y E. SanJuan. 2010. Multilingual summarization evaluation without human models. En *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, páginas 1059–1067. Association for Computational Linguistics.

Sanz, A. 2005. Irakurmena lantzeko jarduerak nola prestatu: Lehen hezkuntzako 3. zikloa eta dbhko 1. zikloa. *Nafarroako Gobernua*.

Zipitria, I., A. Arruarte, y J. A. Elorriaga. 2008. Lea: A summarization web environment based on human instructors' behaviour. En *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, páginas 564–568. IEEE.

Zipitria, I., P. Larrañaga, R. Armañanzas, A. Arruarte, y J. A. Elorriaga. 2008. What is behind a summary-evaluation decision? *Behavior Research Methods*, 40(2):597–612.