La contribución de los métodos de aprendizaje automático no supervisado al diseño de métodos para la clasificación textual según el grado de especialización

Sergio Rodríguez-Tapia - Universidad de Córdoba
sergio.rodriguez@uco.es

Julio Camacho-Cañamón - Universidad de Córdoba
julio.camacho@uco.es

Rebut / Received: 28-7-17
Acceptat / Accepted: 16-11-17

Resum. La contribució dels mètodes d'aprenentatge automàtic no supervisat al disseny de mètodes per a la classificació textual segons el grau d'especialització. Les teories terminològiques modernes es basen en la hipòtesi que existeix un grau d'especialització textual, que depèn de factors diversos, tant lingüístics com extralingüístics. Aquest article té per objectiu mesurar la utilitat dels algoritmes d'aprenentatge automàtic no supervisat (en concret, l'algorisme *simple k-mitjans*) per classificar textos segons el grau d'especialització. Per això, s'usa com a font una base de dades amb informació intra i extratextual i es comparen els resultats amb les etiquetes de classe assignades prèviament mitjançant un mètode numèric de classificació. Els resultats obtinguts suggereixen l'existència del grau i demostren la presència de patrons particulars que se situen en els límits entre classes, la qual cosa revela l'existència de límits difusos i problemes en el mètode plantejat.

Paraules clau: aprenentatge automàtic no supervisat, *k-mitjans*, mètode, terminologia, text, classificació.

Abstract. The contribution of unsupervised machine learning to design methods to study text classification according to specialization degree. Modern terminology theories are based on the hypothesis of the existence of a text specialization degree that depends on different elements, both linguistic and extralinguistic. This article aims to test how useful unsupervised machine learning algorithms (specifically simple k-means algorithm) are to classify texts according to its

specialization degree. To that end, a database with intra and extra textual information is used as a source tool. Results are compared with the class tags previously assigned by means of a numerical classification method. The obtained results suggest the existence of the degree and prove the presence of particular texts that are placed in limits between classes. This fact reveals the existence of vague limits and problems in the proposed method.

Keywords: unsupervised machine learning, *k-means*, method, terminology, text, classification.

1. Estado de la cuestión

La investigación en torno al grado de especialización textual reconoce la existencia de un contínuum que varía de la mayor especialización de un texto (texto especializado: TE) a la menor especialización (texto divulgativo o no especializado: TD) pasando por un estadio intermedio al que denominamos texto semiespecializado (TSE). La investigación teórica se ha centrado en la caracterización lingüística de los polos opuestos (Cabré, 2007; Cabré, Bach, Castellà y Martí, 2007) y ha reconocido la eficiencia de los métodos de las ciencias de la computación en el contraste cuantitativo de dichas características, como avalan los trabajos de Cabré, Bach, Da Cunha, Morales y Vivaldi (2010), Cabré, Da Cunha, Sanjuan, Torres-Moreno y Vivaldi (2011), Cabré, Da Cunha, Sanjuan, Torres-Moreno y Vivaldi (2014) y Da Cunha et al. (2011), o en la recomendación de textos a partir de un modelado de tópicos (Hernández, Tomás y Navarro, 2015). En cuanto a la aplicación específica de algoritmos a tareas de *clustering* de documentos, encontramos estudios como el de Nigam, McCallum, Thrun y Mitchell (2000), aplicado a la clasificación textual; trabajos sobre desambiguación semántica (Martín y Berlanga, 2012) y de nombres de persona (Delgado, Martínez, Fresno y Montalvo, 2014), así como de creación de sistemas de predicción de palabras (Cruz et al., 2011).

A pesar de estas aproximaciones al grado de especialización, el nivel intermedio, controvertido desde un punto de vista teórico y metodológico, queda silenciado en dichos análisis.

En trabajos anteriores, se han realizado propuestas para analizar el contínuum al que hacemos referencia a través de la interrelación de diversos elementos analizables en un corpus con diferentes tipos textuales. La investigación que aquí presentamos hunde sus raíces en las líneas de investigación futuras que se proyectaron en dichos trabajos (Rodríguez-Tapia, 2015, 2016a, 2016b) sobre los criterios teóricos y metodológicos para analizar el grado de especialización textual. En concreto, dos de los nuevos objetivos propuestos en el trabajo de Rodríguez-Tapia (2015, p. 29) pretendían dar respuesta a las siguientes necesidades:

a. Elaborar un corpus mucho mayor con el objetivo de analizar el mayor número de usos, casos y ocurrencias posibles, así como de tipos textuales.
b. Aplicar la lingüística computacional y la lingüística de corpus a los resultados obtenidos en la caracterización del texto semiespecializado.

2. Objetivos

En este artículo, se pretende trabajar con un corpus ligeramente más amplio y emplear los métodos de las ciencias de la computación para identificar los distintos grados de especialización textual a partir de una base de datos con información intra y extratextual. En concreto, nos centraremos en métodos de aprendizaje automático no supervisado (Bishop, 2010). Partiendo de la hipótesis de que el aprendizaje automático, supervisado y no supervisado, contribuye notablemente a interpretar de manera más eficiente los datos relativos a dicha base de datos, nos planteamos los siguientes objetivos:

a. Contribuir al análisis de las lagunas teórico-metodológicas en la identificación, extracción e interpretación de datos relativos a tipos textuales según la especialización.
b. Como objetivo secundario, y como fruto del objetivo anterior, avanzar en la construcción de un método fiable y eficiente para caracterizar los objetos textuales según su grado de especialización, que sirva de proyecto piloto para una investigación con mayores dimensiones. Este método tendrá por objetivo interrelacionar los elementos de análisis del texto para comprobar la probabilidad de que un texto se clasifique dentro de una clase u otra.

A nuestro juicio, esta tímida experimentación resulta necesaria para evaluar la utilidad de un método que emplee el aprendizaje automático. Este método, a partir de los resultados de clasificación obtenidos por los algoritmos de clasificación, permitirá poder correlacionar características lingüísticas y extralingüísticas y clases textuales, proponer un modelo teórico sobre el grado de especialización y validarlo posteriormente sobre un corpus.

Debe subrayarse que este trabajo no usa el método no supervisado para demostrar la hipótesis del contínuum sino que nuestro objetivo es comprobar si el método que propusimos en Rodríguez-Tapia (2016a) (véanse Tablas 1, 2 y 3) es descriptivo de los límites de las categorías del grado de especialización.

3. Marco teórico

La utilidad de la distinción de textos según su grado de especialización tiene relación directa con el trabajo que llevan a cabo los profesionales e investigadores en lingüística aplicada. Los objetivos finales de un proyecto que se entronque a discriminar los textos según su grado de especialización deberían corresponder con la construcción de un modelo que proporcione unos criterios teóricos sólidos y rigurosos a partir de datos empíricos. Estos criterios serían útiles, por ejemplo, en la enseñanza de lenguas para fines específicos, la traducción o la propia lingüística de corpus. En la enseñanza de lenguas podría servir para medir el grado de dificultad que a veces comportan las tradiciones discursivas de ciertas disciplinas, de forma que, aun siendo reales, los materiales proporcionados al aprendiz pueden ser seleccionados según su nivel de competencia sociocognitiva y lingüística. En relación con la traducción, estos criterios son útiles en la búsqueda de textos paralelos de

donde extraer terminología o en la elaboración de estrategias discursivas según el grado de especialización. Por último, para la lingüística de corpus estos criterios son útiles para la compilación de corpus que tengan por objetivo estudiar características muy diversas de cualquier texto siempre que se vinculen con su grado de especialización.

En relación con los enfoques sobre el contínuum, la postura a favor de este usa como objetos de referencia los dos objetos textuales opuestos. Esto permite contrastar sus características como una idealización del texto especializado y una idealización del texto no especializado. Entre ambos, es posible hallar una escala gradual de realizaciones "de tal manera que hay palabras o conceptos más o menos especializados o generales según el contexto y el uso que se les otorgue. Hay, pues, niveles de especialización y generalidad en todos los textos" (Cabré, Domènech, Morel y Rodríguez, 2001, p. 178).

La teoría comunicativa de la terminología considera que el criterio principal para distinguir textos especializados y no especializados tiene que ver con el emisor del texto. Si el emisor es especialista, el texto será clasificado como especializado. En este nivel, según el texto y la situación comunicativa, es posible hallar diferente nivel de especialización y experticidad (especialización alta, media y baja). En esta investigación divergimos ligeramente de dicha propuesta y consideramos que es posible discriminar tres grados de especialización (textos especializados, semiespecializados y no especializados). No obstante, en cada uno de estos estadios puede identificarse gradación y, a nuestro juicio, el emisor no tiene por qué ser un especialista. Nuestro modelo difiere en este sentido en que es necesario conjugar el emisor con otros criterios, como el receptor y las características lingüísticas del texto, para poder clasificar el texto dentro de alguno de los diferentes grados en los que es posible dividir los diferentes niveles del contínuum.

4. Corpus y materiales

El presente artículo constituye uno de los proyectos piloto para elaborar un modelo integrador para analizar el grado de especialización textual y pretende experimentar con una metodología fiable y eficiente para estudiar relaciones entre atributos y proponer probabilidad de clases. Este proyecto permitirá, por tanto, conocer y poner a prueba las lagunas teóricas y problemas metodológicos que suponen los valores de los atributos y las clases con las que trabajamos, así como las posibles conclusiones con respecto al uso de algoritmos de aprendizaje automático y su utilidad al relacionar atributos, clases y límites difusos. Para extraer estas conclusiones, el principal recurso de análisis es la preparación de una base de datos que sintetiza los datos extraídos del análisis manual realizado por el experto a partir de un corpus lingüístico.

4.1. Criterios de confección de la base de datos

En este trabajo utilizamos el corpus analizado en Rodríguez-Tapia (2015, 2016b) aunque ha sufrido ciertas modificaciones para que, basándonos en la posición teórica

respecto a la relevancia de los corpus lingüísticos (Biber, 2012; McEnery y Hardie, 2012), cumpla con las especificaciones teóricas que permitan diseñar un corpus más representativo para nuestro estudio. Así, los criterios seguidos para su diseño han sido los siguientes:

a. Partimos del campo temático médico, en concreto, del objeto *insuficiencia cardíaca*[1].
b. Seleccionamos 6 tipos de texto o superestructuras en español y 10 textos por cada superestructura, lo que suma un total de 60 textos, 189 000 palabras, 118 000 unidades léxicas, de forma que podamos garantizar que trabajemos con un corpus balanceado en tipos textuales y patrones.
c. Una vez analizados de forma manual los 60 textos, se ha diseñado una base de datos mediante MO Excel que contiene información relativa al número de palabras que contiene cada texto, su número de unidades léxicas, su número de términos, su índice de densidad terminológica, la relación discursiva entre interlocutores de cada texto, su función lingüística y su superestructura. De esta forma, cada patrón de la base de datos cuenta con un valor para cada atributo diferente y, como resultado de la suma de estos, con una clasificación según la clase (en TE, TSE y TD). Esta investigación no explota el corpus de forma automática, por lo que no se usan las palabras, lexemas, categorías morfosintácticas, etc. como características de los textos que son de entrada al algoritmo de *clustering*.

Debemos recordar que este estudio constituye un proyecto piloto de un proyecto más amplio y que nuestro objetivo aquí no es caracterizar lingüística y extralingüísticamente las clases textuales según el grado de especialización, sino identificar las fortalezas y debilidades de los métodos que proponemos usando el aprendizaje automático. Así, somos conscientes de que trabajamos con un corpus muy restringido y reducido en número de instancias y número de palabras. No obstante, a nuestro juicio, resulta suficiente para comprobar coincidencias entre máquina y experto, identificar problemas de clasificación derivados de los límites difusos, así como analizar problemas teóricos en la concepción del método.

4.2. Valores de los atributos y método numérico de clasificación

Según se justifica en el modelo numérico descrito en Rodríguez-Tapia (2015 y 2016b) y Rodríguez-Tapia y Camacho-Cañamón (2018), los valores de cada atributo

1. Las hipótesis que se barajan con este proyecto pretenden también saber qué papel desempeña el campo temático en el grado de especialización de los textos. Para ello se cuenta con un corpus multidisciplinar. En este trabajo solo se presentan los resultados aplicados a la medicina como simplemente una muestra del corpus. En futuros trabajos se compararán los resultados de las diferentes muestras. Además, se decidió seleccionar el objeto *insuficiencia cardíaca* debido a la atención que recibe en el sector público, profesional, investigador y divulgador actualmente, tanto en las lenguas inglesa como española.

(Tabla 1) se codifican numéricamente de mayor a menor según su relevancia en el grado de especialización de un texto[2] (Tabla 2).

Tabla 1. Valores de atributos y clases de la base de datos

Atributos	
Función	**Índice de densidad terminológica**
Representativa Representativa-comunicativa Comunicativa	Valores entre 0 y 1
Relación discursiva de los interlocutores	**Superestructura**
Especialista-especialista Especialista-lego Instruido-lego	Artículo científico Artículo divulgativo Tesis doctoral Carta al editor Guía para el paciente Manual para el especialista
Clases	
Texto especializado Texto semiespecializado Texto divulgativo	

Tabla 2. Valores porcentuales por atributos

Relación discursiva entre interlocutores (Máximo 0,3)	Especialista-especialista	0,3
	Especialista-lego	0,2
	Instruido-lego	0,1
Función principal del texto (Máximo 0,4)[3]	Representativa	0,4
	Representativa-comunicativa	0,3
	Comunicativa	0,2
Índice de densidad terminológica (Máximo 0,3)	Valor numérico del IDT multiplicado por 100	0,3

Cada texto es calificado con un valor numérico (por ejemplo, en el caso de la relación discursiva, 0,3; 0,2; 0,1) cuando se identifica en él el correspondiente valor del atributo.

2. Debe tenerse en cuenta que, desde nuestra perspectiva, la superestructura no es condicionante esencial del grado de especialización, por lo que no está presente en dicha tabla.

3. El modelo eminentemente funcional que empleamos considera que la función lingüística tiene mayor relevancia en la clasificación textual. Las investigaciones que se están llevando a cabo actualmente en el marco del proyecto pretenden analizar las implicaciones de esta decisión y la posibilidad de modificar dicha concepción funcional hacia una más sociocognitiva.

Así, la conjunción de los valores de los atributos proporciona un valor numérico total que permite clasificar los textos de forma objetiva según un intervalo de valores en tres clases diferenciadas (Tabla 3).

Tabla 3. Valores numéricos para clasificar las clases

Clase	**Valor clasificador**
Texto especializado	11 ≤ x
Texto semiespecializado	7 ≤ x < 11
Texto divulgativo	0,3 ≤ x <7

El mínimo obtenible de la suma de valores es 0,3: 0,1 del valor instruido-lego y 0,2 del valor de función comunicativa. El rango hasta 7 y hasta 11 se estableció *a priori*, considerando la hipótesis de que existe un número mayor de textos divulgativos en la realidad. Esta decisión metodológica repercute en la clasificación final, si bien nuestro interés no se halla tanto en proponer un método definitivo de clasificación, sino en hallar las fortalezas y debilidades de esta propuesta metodológica en este proyecto piloto.

Estos valores, como se defendía en Rodríguez-Tapia (2016b) silencian los límites difusos entre clases, aunque permiten aproximarse al contínuum. En otras palabras, trabajar con un modelo numérico basado en un análisis manual permite clasificar los textos de forma objetiva y teorizar sobre sus características prototípicas, con la desventaja de que los límites discretos no son totalmente descriptivos (y mucho menos explicativos) del fenómeno de gradación. La comparación de la clasificación con el método matemático y la clasificación de los algoritmos no supervisados deberían poder revelar dicha debilidad, aun siendo el porcentaje de acierto muy elevado.

4.3. Software

Uno de los programas más extendidos y utilizados hoy en día para el tratamiento de datos y la extracción de conocimiento es WEKA (Waikato Environment for Knowledge Analysis). Contiene una avanzada y puntera colección de algoritmos de aprendizaje automático y herramientas de preprocesado de datos, como pueden ser el modelado de tablas de decisión, los árboles de decisión, las reglas de decisión y clasificación, o la creación de agrupaciones de datos en *clusters* (Witten y Frank, 2005, pp. 365-366).

5. Resultados

Utilizando la base de datos con la información extraída del análisis del corpus y los diferentes algoritmos supervisados y no supervisados del programa WEKA, se obtienen resultados muy variados, cuyo estudio repercute directamente en el planteamiento metodológico de investigación del grado de especialización textual. Este apartado se

divide en: a) un primer análisis estadístico de la información de la base de datos, y b) una sección dedicada a analizar los resultados del aprendizaje automático mediante algoritmos no supervisados. Tras esto, se comparan los resultados con la clase determinada por el modelo numérico mencionado anteriormente.

5.1. Primer análisis de la base de datos

En esta primera sección, analizamos los valores de cada atributo, además de la clase, para conocer algunos datos de partida relevantes a la hora de trabajar con los diferentes algoritmos.

5.1.1 Sobre la clase

El corpus (y, por tanto, la base de datos), se encuentra desbalanceado hacia el texto especializado (25 textos de un total de 60), por lo que se trata de la muestra más representativa sobre el total de las clases (Figura 1).

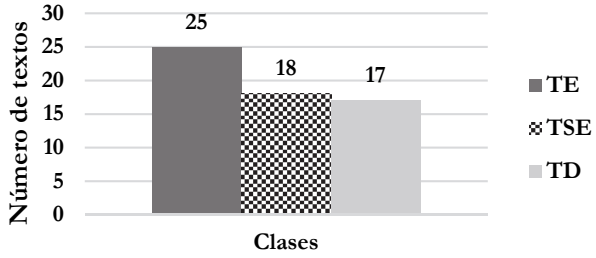

Figura 1. Número de textos por clase.

5.1.2 Sobre el índice de densidad terminológica

El índice de densidad terminológica codifica la relación que existe entre número de unidades léxicas especializadas o términos entre el número de unidades léxicas totales de un texto. En otras palabras, codifica de forma numérica el grado de conocimiento especializado que integra un texto. Los valores del índice de densidad terminológica oscilan entre 0,575 (mayor densidad) y 0,009 (menor densidad), siendo el valor promedio de 0,168.

5.1.3 Sobre la relación discursiva

Existen dos clases que se vinculan exclusivamente con un tipo de relación discursiva, no existiendo casos de diversos valores de relación discursiva para un mismo atributo. Se trata de los TE, que siempre presentan un valor de especialista-especialista, y los TD, que siempre presentan un valor de instruido-lego. El TSE presenta opciones más variadas, lo que coincide con su condición de transición entre polos opuestos.

5.1.4 Sobre la superestructura

No se comprueba relación directa entre clase y superestructura como ocurre con la relación discursiva. Aunque no sean valores exclusivos para cada clase, sí que es posible percibir varias cuestiones:

a. El artículo científico, el manual para el especialista y la tesis doctoral casi siempre son TE y, en menor medida, TSE.
b. La guía para el paciente y el artículo divulgativo casi siempre son TD y, en menor medida, TSE.
c. La carta al editor es la forma más variada y destaca la clasificación en TSE por encima de los otros dos valores.

5.2. Resultados con algoritmos automáticos no supervisados: Tres clusters

El aprendizaje automático no supervisado no emplea la etiqueta de clase para construir el modelo, sino que se nutre únicamente de los atributos facilitados. De esta forma, el algoritmo desconoce la clase que hemos asignado a cada patrón mediante nuestro modelo numérico de clasificación. En ocasiones, se puede aplicar aprendizaje automático no supervisado aun teniendo la etiqueta de clase, pero ignorándola en todo el proceso de modelado. Esta solo se emplea para la evaluación del modelo para comprobar la similitud entre los resultados del experto y el algoritmo.

Frente a los algoritmos supervisados, los métodos no supervisados tienen la desventaja de que no pueden utilizar el conocimiento que se podría extraer de la etiqueta de clase para crear el modelo. Pero, a cambio, son más robustos en tanto que son capaces de realizar las agrupaciones sin necesidad de conocer dicha etiqueta (Marsland, 2015, pp. 195-196).

A continuación, presentamos una tabla con los resultados de los algoritmos utilizados (Tabla 4). En ella, se puede observar el porcentaje de error en la clasificación textual de acuerdo con los valores de los atributos proporcionados en la base de datos y el número de grupos (*clusters*) que propone cada algoritmo (utilizando como datos de comprobación de los resultados la clase asignada a cada texto). Los tres últimos algoritmos necesitan que el experto les indique el número de *clusters* con los que desea trabajar (en nuestro caso, tres agrupaciones, puesto que nuestra hipótesis distingue tres clases).

Toda la experimentación se ha realizado siguiendo una metodología de validación cruzada con un 10-fold para estimar los parámetros de cada uno de los algoritmos sobre el conjunto de entrenamiento. Los resultados que se muestran en la Tabla 4 son los obtenidos tras aplicar los algoritmos en el conjunto de generalización. De este modo se ha garantizado la validez de la experimentación y la independencia de los conjuntos de entrenamiento y generalización.

Tabla 4. Resultados de los algoritmos no supervisados

Algoritmo	**Número de clusters propuesto**	**Textos mal clasificados**	
Simple expectation maximization (EM)	3	14	23,33 %
Farthest First	3	13	21,67 %
Cluster jerárquico	3	14	23,33 %
Simple k-means	3	12	20 %

En primer lugar, cabe destacar que el algoritmo EM trata de encontrar el número óptimo de grupos que deben crearse y, en este caso, acierta en el número de *clusters* (3), puesto que trabajamos con tres clases diferentes. No obstante, el porcentaje de error es elevado (23,33 %).

En contraste, *simple k-means* obtiene resultados muy positivos en relación con el resto de algoritmos empleados (de hecho, muestra una tasa de acierto del 80%). La contribución del algoritmo no es la creación de tres grupos, pues es lo que se le indica al algoritmo, sino que los tres grupos que crea el algoritmo coinciden en gran medida con las agrupaciones realizadas manualmente mediante el método numérico de clasificación, muy posiblemente porque los resultados con los que se comparan son fruto de un método matemático. Las asignaciones que realiza se relacionan en la Tabla 5.

Tabla 5. Clasificación de textos usando *simple k-means* (tres clases)

Asignado a →	**C2 (TE)**	**C0 (TSE)**	**C1 (TD)**
TE	20 textos correctos	5: [15,33,36,38,48]	0
TSE	2:[3,4]	13 textos correctos	3: [29,53,54]
TD	0	2: [20,11]	15 textos correctos

Puede comprobarse que el objeto TSE es el que mayor conflicto causa, siendo los polos opuestos los que no conllevan error en la clasificación. La Figura 2 muestra la matriz de confusión de forma gráfica.

A pesar de esta nula coincidencia, cabe estudiar los textos que han sido mal clasificados (Tabla 6). Para ello, recurrimos a los valores numéricos que empleamos para clasificar cada texto dentro de una clase (Tabla 3).

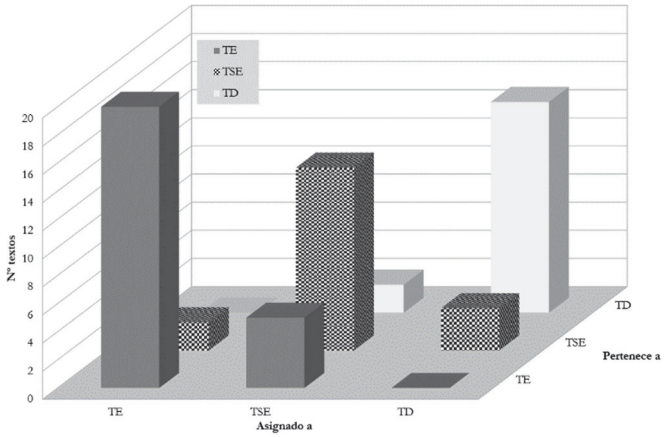

Figura 2. Representación de los *clusters* generados por el algoritmo *K-means* con las etiquetas para tres clases.

Tabla 6. Comparación de clasificaciones según nuestro modelo numérico y *simple K-means*

Textos	**Nuestra clasificación**	**Clasificación de *k-means***	**Valor numérico de clasificación**
3	TSE	TE	9,75
4	TSE	TE	9,36
11	TD	TSE	6,41
15	TE	TSE	11,26
20	TD	TSE	5,9
29	TSE	TD	9,69
33	TE	TSE	11,79
36	TE	TSE	16,37
38	TE	TSE	15,45
48	TE	TSE	11,93
53	TSE	TD	7,37
54	TSE	TD	7,09

Con estas tablas, podemos justificar a través de los valores numéricos de la clasificación los límites difusos entre categorías y, de alguna forma, las confusiones del algoritmo:

- El texto 11 (TD) tiene un valor de 6,41, valor que se aproxima estrechamente al TSE (límite numérico = 7) por debajo.
- Los textos 53 y 54 (TSE) tienen valores de 7,37 y 7,09 respectivamente, valores que se aproximan estrechamente al TD (límite numérico = 7) por encima.
- Los textos 15, 33 y 48 (TE) tienen valores de 11,26, 11,79 y 11,93 respectivamente, valores que se aproximan estrechamente al TSE (límite numérico = 11) por encima.

Asimismo, encontramos casos (textos 3, 4, 20, 29, 36, 38) que no son justificables desde el punto de vista del valor numérico del método de clasificación. De esta forma, podría dilucidarse que el algoritmo no asigna la misma relevancia que el modelo de Rodríguez-Tapia (2016a) a cada uno de los atributos (lo que podría justificar el cambio de concepción funcionalista en la asignación de relevancias en los atributos, Tabla 2). Esta hipótesis y estos textos deberían tratarse en trabajos futuros para comprobar sus condiciones especiales, quizá usando 4 o 5 *clusters*.

Esta cercanía en los valores al límite numérico de clasificación permite defender la idea de gradualidad. De hecho, podría afirmarse que, por ejemplo, los textos 15, 33 y 48 (asignados a TSE pero clasificados por el experto como TE)[4] se configuran como textos cuya tendencia es a constituirse como tipo TE más que TD. Este hecho nos hace pensar en la amplitud de las regiones de cada clase, que, por ejemplo, podrían ser más amplias para el TE y el TD y más estrechas para el TSE. O al contrario, dependiendo de la perspectiva teórica que elijamos. Estas múltiples posibilidades tienen estrecha relación con la lógica difusa (Zadeh, 1965 y 1968), que permite explicar la posición diferencial que existe entre TE y TD.

5.3. Resultados con algoritmos automáticos no supervisados: cuatro clusters y comparación

Observados los límites del método numérico propuesto, en investigaciones posteriores (Rodríguez-Tapia y Camacho-Cañamón, 2017) se decidió realizar de nuevo la misma prueba variando tanto el método de clasificación como el número de clases: a) el método sigue los fundamentos de la teoría sociocognitiva de la percepción (Caravedo, 2014), por la que el experto es quien decide las clases que corresponden a cada texto; y b) las clases pasan a ser cuatro. Dichas clases se correspondían con:

a. TE: texto especializado.
b. TE>TD: texto especializado con tendencia o características del texto divulgativo.
c. TD>TE: texto divulgativo con tendencia o características del texto especializado.
d. TD: texto divulgativo.

4. El mismo caso se aplica a los objetos 53 y 54 para el TD.

El estudio de Rodríguez-Tapia y Camacho-Cañamón (2017) volvió a realizar las pruebas con *k-means* con cuatro *clusters* en lugar de tres. La Tabla 7 muestra los patrones correcta e incorrectamente clasificados y la Figura 3 permite observar gráficamente la clasificación de textos relacionada en la Tabla 8.

Tabla 7. Comparación de clasificaciones según estudios con tres y cuatro clases

	Estudio con tres clases	**Estudio con cuatro clases**
Patrones bien clasificados	80,0 %	76,7 %
Patrones mal clasificados	20,0 %	23,3 %
Textos mal clasificados	12 textos	14 textos

Tabla 8. Clasificación de textos usando *SIMPLE K-MEANS* (cuatro clases)

Asignado a →	**TE**	**TE>TD**	**TD>TE**	**TD**
TE	25 textos correctos	4: [3, 29, 31 y 35]		
TE>TD	3: [22, 33, 45]	5 textos correctos	1: [9]	
TD>TE	1: [13]	1: [1]	6 textos correctos	2: [2, 26]
TD			2: [47, 48]	10 textos correctos

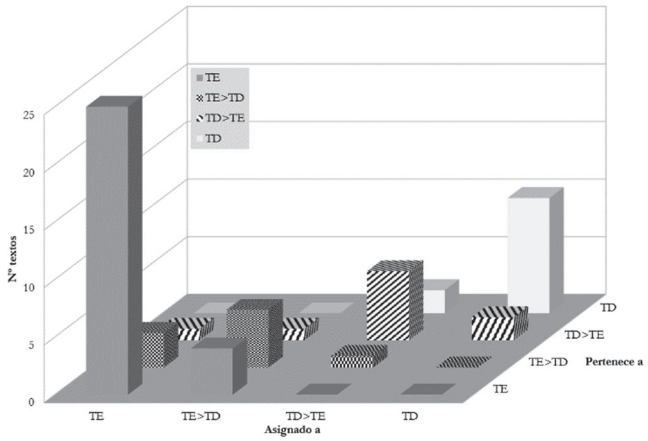

Figura 3. Representación de los *CLUSTERS* generados por el algoritmo *K-MEANS* con las etiquetas para cuatro clases.

Los resultados por ahora sugieren que en los dos *clusters* intermedios es posible hallar un elevado porcentaje de patrones muy similares, divergentes en valores como la densidad terminológica. Aunque estos resultados deben seguir siendo estudiados y contrastados con otras agrupaciones, consideramos que parte de los dos *clusters* intermedios coinciden con el *cluster* de texto semiespecializado de la prueba con tres clases.

Como es posible comprobar en la Tabla 9, parece ser que el nivel intermedio es el que genera mayores problemas al algoritmo. Con las pruebas con tres clases, los textos que siempre se confunden tienen por protagonista al TSE; con las pruebas de cuatro clases, los textos que siempre se confunden tienen por protagonistas las categorías próximas: TE con TE>TD, TE>TD con TD>TE y TD>TE con TD (a excepción del texto 13). Desde nuestro punto de vista este es un argumento claro a favor de los límites difusos que existen entre las categorías y permiten esbozar la hipótesis de la existencia del nivel semiespecializado. Si este es divisible en dos, tres o cuatro estadios intermedios (lo mismo podría argumentarse para los textos especializados o divulgativos) es una cuestión que debe resolverse en próximas investigaciones.

Tabla 9. Comparación de los textos cuya clasificación es errónea (tres y cuatro clases)

N.º	**(A) Clasificación numérica (tres clases)**	**(A) Clasificación *k-means* (tres clases)**	**(B) Clasificación perceptiva (cuatro clases)**	**(B) Clasificación *k-means* (cuatro clases)**	**¿Textos clasificados erróneamente en ambos casos?**
1	TE	TE	TD>TE	TE>TD	NO
2	TE	TE	TD>TE	TD	NO
3	TSE	TE	TE	TE>TD	**SÍ**
	IDT: 0,09 Relación: especialista-especialista		Función: representativa Superestructura: artículo científico		
4	TSE	TE	TE	TE	NO
9	TE	TE	TE>TD	TD>TE	NO
11	TD	TSE	TD>TE	TD>TE	NO
13	TSE	TSE	TD>TE	TE	NO
15	TE	TSE	TE	TE	NO
20	TD	TSE	TD>TE	TD>TE	NO
22	TD	TD	TE>TD	TE	NO

N.º	(A) Clasificación numérica (tres clases)	(A) Clasificación *k-means* (tres clases)	(B) Clasificación perceptiva (cuatro clases)	(B) Clasificación *k-means* (cuatro clases)	¿Textos clasificados erróneamente en ambos casos?
26	TD	TD	TD>TE	TD	NO
29	TSE	TD	TE	TE>TD	**SÍ**
	IDT: 0,19 Relación: especialista-lego			Función: comunicativa Superestructura: guía para el paciente	
31	TE	TE	TE	TE>TD	NO
33	TE	TSE	TE>TD	TE	**SÍ**
	IDT: 0,19 Relación: especialista-especialista			Función: representativa-comunicativa Superestructura: manual para el especialista	
35	TE	TE	TE	TE>TD	NO
36	TE	TSE	TE	TE	NO
38	TE	TSE	TE	TE	NO
45	TE	TE	TE>TD	TE	NO
48	TE	TSE	TD	TD>TE	**SÍ**
	IDT: 0,20 Relación: especialista-especialista			Función: comunicativa Superestructura: tesis doctoral	
53	TSE	TD	TD>TE	TD>TE	NO
54	TSE	TD	TD>TE	TD>TE	NO

En la Tabla 9 también identificamos los cuatro textos que han sido clasificados erróneamente tanto al usar *k-means* con tres y con cuatro clases. Dadas las similitudes que existen en el IDT y en la relación discursiva en los cuatro textos, parece ser que el algoritmo *k-means* con cuatro clases usa los valores de la función como atributo discriminador de clases. La relevancia de los atributos en las clasificaciones y las posibles justificaciones se tratarán en trabajos futuros que usen aprendizaje automático supervisado.

6. Conclusiones y futuros trabajos

La primera aproximación al análisis de clases según el grado de especialización empleando métodos de aprendizaje automático no supervisado ha permitido trabajar

con tres agrupaciones de patrones. El primer análisis de los tres *clusters* propuestos por *k-means (*Figura 2) presenta grupos lo suficientemente homogéneos como para defender la existencia de clases diferenciadas por cada *cluster.*

No obstante, existe una agrupación muy confusa: la perteneciente a la clase TSE. En su mayoría, los textos de tipo TSE han sido agrupados en un solo *cluster* independiente. Sin embargo, algunos textos de tipo TSE se han infiltrado en los grupos donde principalmente estaban agrupados los de tipo TE y TD respectivamente. Algo parecido sucede con los textos TE y TD, son agrupados en *clusters* individuales pero algunos textos se confunden y aparecen agrupados en el *cluster* donde se incluyen los de tipo TSE. A la luz de estos datos, puede concluirse que la clase TSE tiene unos límites más difusos de los que *a priori* podía parecer.

En definitiva, con este primer experimento pretendemos comprobar si las agrupaciones realizadas por *k-means* y las clases propuestas manualmente por el experto a través del método numérico coinciden (y en qué medida coinciden). Para ello, una vez agrupados los textos por el algoritmo, hemos revisado la clase predominante en cada grupo, de forma que esta pudiese representar la clase del *cluster*. De esta forma se puede analizar qué textos han sido agrupados en un *cluster* cuya clase predominante coincide o no con la clase del texto.

Así, es posible identificar patrones que implican problemas de clasificación según los tres grados que proponemos al algoritmo, debido a que se constituyen como textos con particularidades que los sitúan en los límites entre clases. Estos textos son los que demuestran: a) los problemas teóricos y metodológicos de clasificación textual según el grado de especialización debido a la inexistencia de límites claros y b) las debilidades del método de clasificación propuesto por Rodríguez-Tapia (2016a) en cuanto a su potencial para describir completamente el fenómeno del grado de especialización.

Con una segunda prueba, esta vez con cuatro clases y usando un método de clasificación manual siguiendo la teoría de la percepción, vemos que los errores de clasificación se dispersan más entre las dos categorías intermedias. Los datos obtenidos tanto con las pruebas con tres como con cuatro clases no permiten demostrar el número de estadios que existen en el contínuum pero sí refuerzan nuestra hipótesis del grado de especialización.

Este punto de partida hace posible contar con líneas de trabajo futuras que permiten evaluar la eficiencia y utilidad de los métodos de clasificación de textos según el grado de especialización, como pueden ser las siguientes:

- Aplicar métodos de aprendizaje automático supervisado a una base de datos que contenga las clases usando el método numérico de clasificación y a otra cuyas clases hayan sido establecidas usando otro tipo de clasificaciones, como las obtenidas a partir de un grupo de informantes.
- Aplicar métodos de aprendizaje automático no supervisado a bases de datos con dos o cinco clases con el objetivo de profundizar en el análisis empírico del contínuum.

- Aplicar métodos de clasificación ordinal, atendiendo a la hipótesis del contínuum.
- Aumentar el número de atributos, sobre todo de carácter lingüístico, como tipos de términos o tipos de reformulación. La preparación de los siguientes estudios tiene previsto poder incluir unos 10 o 12 atributos.
- Debido a que los métodos numéricos resultan demasiado restrictivos, proponemos abandonar dichos modelos matemáticos e incorporar la teoría sociocognitiva de la percepción como teoría vertebradora de la clasificación textual según el grado de especialización.

Referencias

Biber, D. (2012). Representativeness in corpus linguistics. En D. Biber y R. Reppen, (Eds.), *Corpus linguistics* (pp. 3-33). Londres: SAGE.

Bishop, C. M. (2010). *Pattern recognition and machine learning*. Nueva York: Springer.

Cabré, M. T. (2007). Constituir un corpus de textos de especialización: condiciones y posibilidades. En M. Ballard y C. Pineira-Tresmontant (Eds.), *Les corpus en linguistique et en traductologie* (pp. 89-106). Arras: Artois Presses Université.

Cabré, M. T., Domènech, M., Morel, J., y Rodríguez, C. (2001). Las características del conocimiento especializado y la relación con el conocimiento general. En M. T. Cabré y J. Feliu (Eds.), *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (pp. 173-186). Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cabré, M. T., Bach, C., Castellà, J. M., y Martí, J. (2007). La caracterización lingüística del discurso especializado. En R. Mairal (Coord.) *Aprendizaje de lenguas, uso del lenguaje y modelación cognitiva. Actas del XXIV Congreso Nacional de AESLA* (pp. 851-857). Madrid: UNED.

Cabré, M. T., Bach, C., Da Cunha, I., Morales, A., y Vivaldi, J. (2010). Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: El caso del discurso económico. En M. R. Caballero Rodríguez y M. J. Pinar Sanz (Eds.), *Modos y formas de la comunicación humana* (pp. 453-460). Ciudad Real: Universidad de Castilla-La Mancha.

Cabré, M. T., Da Cunha, I., Sanjuan, E., Torres-Moreno, J.-M., y Vivaldi, J. (2011). Automatic specialized vs. non-specialized texts differentiation: a first approach. En N. Talaván, E. Martín Monje y F. Palazón (Eds.), *Technological innovation in the teaching and processing of LSPs: Proceedings of TISLID'10* (pp. 301-310). Madrid: UNED.

Cabré, M. T., Da Cunha, I., Sanjuan, E., Torres-Moreno, J.-M., y Vivaldi, J. (2014). Automatic specialized vs. non-specialized text differentiation: The usability of grammatical features in a Latin multilingual context. En E. Bárcena, T. Read y J. Arús (Eds.), *Languages for specific purposes in the digital era* (pp. 223-241). Berlín: Springer.

Caravedo, R. (2014). *Percepción y variación lingüística: Enfoque sociocognitivo*. Madrid/Frankfurt am Main: Iberoamericana/Vervuert.

Cruz Cavalieri, D., Bastos Filho, T., Sarcinelli Filho, M., Palazuelos Cagigas, S. E., Macias-Guarasa, J., y Martín Sánchez, J. L. (2011). A part-of-speech tag clustering for a word prediction system in portuguese language. *Procesamiento del Lenguaje Natural*, 47, 197-205. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/984/737.

Da Cunha, I., Cabré, M. T., Sanjuan, E., Sierra, G., Torres-Moreno, J.-M., y Vivaldi, J. (2011). Automatic Specialized vs. Non-Specialized Sentence Differentiation. En A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science (LNCS)* (pp. 266-276). Berlín: Springer.

Delgado, A., Martínez, R., Fresno, V., y Montalvo Herranz, S. (2014). An unsupervised algorithm for person name disambiguation in the web. *Procesamiento del Lenguaje Natural*, 53, 51-58. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5042/2930.

Hernández, A., Tomás, D., y Navarro, B. (2015). Una aproximación a la recomendación de artículos científicos según su grado de especificidad. *Procesamiento del Lenguaje Natural*, 55, 91-98. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5220/3024.

Marsland, S. (2015). *Machine learning: An algorithmic perspective*. Boca Ratón: CRC.

Martín, T., y Berlanga, R. (2012). A clustering-based approach for unsupervised word sense disambiguation. *Procesamiento del Lenguaje Natural*, 49, 49-56. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4553/2719.

McEnery, T., y Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University.

Nigam, K., McCallum, A. K., Thrun, S., y Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2), 103-134. Recuperado de http://www.kamalnigam.com/papers/emcat-mlj99.pdf.

Rodríguez-Tapia, S. (2015). Estrategias de traducción inglés-español basadas en el análisis cuantitativo de procedimientos de reformulación formal y conceptual del texto semiespecializado. *Tonos Digital: Revista Electrónica de Estudios Filológicos*, *29*(2), 1-34. Recuperado de http://www.tonosdigital.com/ojs/index.php/tonos/article/viewFile/1330/805.

Rodríguez-Tapia, S. (2016a). Los textos especializados, semiespecializados y divulgativos: una propuesta de análisis cualitativo y de clasificación cuantitativa. *Signa: Revista de la Asociación Española de Semiótica*, 25, 987-1006. Recuperado de http://revistas.uned.es/index.php/signa/article/view/16926/14512.

Rodríguez-Tapia, S. (2016b). Clasificación cuantitativa de los textos según su grado de especialidad: Parámetros para la elaboración de los índices de densidad terminológica y de reformulación de un corpus sobre insuficiencia cardíaca. *Anuario de Estudios*

Filológicos, 39, 227-250. Recuperado de https://dialnet.unirioja.es/descarga/articulo/5854830.pdf.

Rodríguez-Tapia, S., y Camacho-Cañamón, J. (2017). *Las ciencias de la computación como apoyo a la clasificación tipológica textual en terminología*. Ponencia presentada en el II Congreso Internacional Ciencia y Traducción, Universida de Córdoba.

Rodríguez-Tapia, S., y Camacho-Cañamón, J. (2018). Los métodos de aprendizaje automático supervisado en la clasificación textual según el grado de especialización. *Tonos Digital: Revista electrónica de estudios filológicos*, 35, 1-26.

Witten, I. H., y Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Ámsterdam: Morgan Kaufmann.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338-353. Disponible en http://www.sciencedirect.com/science/article/pii/S001999586590241X/pdf.

Zadeh, L. A. (1968). Fuzzy algorithms. *Information and Control*, *12*(2), 94-102. Disponible en http://www.sciencedirect.com/science/article/pii/S0019995868902118/pdf.