Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 109-116 recibido 30-03-2018 revisado 30-04-2018 aceptado 04-05-2018

Clasificación automatizada de marcadores discursivos

Automatic categorization of discourse markers

Hernán Robledo, Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
hernan.robledo.n@mail.pucv.cl, rogelio.nazar@pucv.cl

Resumen: Presentamos un método de clasificación de marcadores del discurso. A partir de una taxonomía generada inductivamente en un trabajo anterior, desde un corpus paralelo de gran tamaño y utilizando una técnica de *clustering*, proponemos ahora un sistema que permite clasificar un marcador discursivo no incluido en esa taxonomía en alguna de las categorías emergentes. Está basado en el cálculo de la similitud estadística entre el nuevo marcador y las categorías. Destacamos la naturaleza cuantitativa del enfoque, que permite la reproducción del experimento en otras lenguas. Además, el sistema propuesto es un clasificador multicategoría, y esto es importante ya que representa un primer acercamiento al estudio de la polifuncionalidad de los marcadores del discurso desde un enfoque empírico e inductivo.

Palabras clave: marcadores del discurso, métodos cuantitativos, métodos inductivos, clasificación multicategorial

Abstract: We present a method for the categorization of discourse markers. Starting from the result of a previous research, in which we generated a taxonomy of discourse markers by inductive methods from parallel corpus, we propose now a method to classify new discourse markers in one or more of the categories discovered in our previous research. The method is based on the statistical similarity between a new marker and the emerging categories. We highlight the quantitative nature of the approach, because it will allow to replicate experiments in other languages. Furthermore, ours is a multi-label classification method, which is important because it represents a first approach to the study of the polyfunctionality of discourse markers from an empirical and inductive point of view.

Keywords: discourse markers, inductive methods, quantitative methods, multi-label categorization

1 Introducción

Los llamados marcadores del discurso (MDs) corresponden a un amplio y heterogéneo conjunto de unidades lingüísticas (por ejemplo, *sin embargo*, *no obstante*, *es decir*, *por lo tanto*, *en consecuencia*, *a todo esto*, *por una parte*, *en primer lugar*, *claramente*, entre muchas otras) de uso muy habitual en las lenguas naturales, tanto en la escritura como en la oralidad.

Los MDs han sido estudiados en una gran variedad de lenguas, como español, inglés, francés, alemán, chino, italiano, japonés, portugués, ruso y muchas otras –incluso, en lenguas de señas– y se han explorado en una diversidad de géneros discursivos y contextos de interacción, como narraciones, discursos políticos, discursos periodísticos, salas de clases, programas radiales, etc. (Maschler y Schiffrin, 2015). Así, si por un lado, los MDs muestran este uso tan extendido entre las lenguas naturales, en la teoría lingüística, en tanto, su delimitación como objeto de estudio, su denominación categorial, sus propiedades funcionales (y/o formales) y, en consecuencia, su clasificación han sido campo de mucha controversia (Fischer, 2006; Loureda y Acín, 2010; Maschler y Schiffrin, 2015).

La categoría de unidades que funcionan como MDs está constituida por elementos que provienen de distintas categorías gramaticales, como conjunciones y locuciones

ISSN 1135-5948. DOI 10.26342/2018-61-12

conjuntivas, adverbios y locuciones adverbiales, interjecciones, preposiciones, expresiones performativas, sintagmas preposicionales, entre otras. Además de esta heterogeneidad de orígenes, los MDs operan en distintos niveles: conectan oraciones, cumplen funciones en el texto y operan a nivel interpersonal (Brinton, 1996; Jucker y Ziv, 1998; Aijmer, 2002). Muchos de ellos, además, son polifuncionales, fenómeno que ha sido descrito desde distintas posturas teóricas (Schiffrin, 1987; Wierzbicka, 2003; Aijmer, Foolen, y Vandenbergen, 2006; Fischer, 2006; Fraser, 2006, por ejemplo) y que refiere al hecho de que un MD puede cumplir distintas funciones pragmáticas en el discurso (ver sección 2.3).

El análisis de estos distintos usos ha sido la motivación de la presente investigación. En una investigación previa (Robledo, Nazar, y Renau, 2017) desarrollamos un método para la inducción automática de taxonomías de MDs a partir de corpus paralelos, utilizando una técnica de *clustering* (conglomerados). Por ejemplo, uno de esos *clusters* nos presenta elementos que denominamos conectores contraargumentativos, tales como *sin embargo*, *por el contrario* o *en vez de ello* (ver sección 2.2).

Lo que presentamos en este artículo es cómo, a partir de esa taxonomía ya generada, desarrollar un sistema de clasificación de cualquier MD, naturalmente sin necesidad de que esté ya incluido en esa taxonomía. Así, un elemento no incluido en esa taxonomía, como *empero*, es detectado como similar –desde el punto de vista distribucional– al *cluster* de los conectores contraargumentativos. En otras palabras, aprovechamos el resultado de un estudio exploratorio, que nos proporcionó *clusters* de MDs que cumplen la misma función y que nosotros etiquetamos con nombres de categorías, para utilizarlo como material de entrenamiento para un sistema de clasificación automática de nuevos marcadores.

Destacamos cuatro aspectos como los más relevantes de esta investigación: 1) que las categorías de MDs no provienen de la literatura, sino que se llega a ellas como categorías emergentes a partir del *clustering* inicial; 2) que ahora podemos clasificar un MD en más de una categoría, ya que algunos de ellos son polifuncionales; 3) que más allá de las consecuencias teóricas de la investigación, encontramos diversas aplicaciones prácticas en lexicografía, *parsing* discursivo y redacción asistida por computador, entre otras; y 4) que este estudio está basado íntegramente en análisis cuantitativo, lo que facilita la reproducción de los resultados del experimento en otras lenguas.

El artículo se organiza de la siguiente forma: en la Sección 2 presentamos un breve estado de la cuestión en el estudio de los MDs. Allí (Subsección 2.2) explicamos también los resultados de nuestra propia investigación anterior y describimos la taxonomía inductiva de MDs. La Sección 3 presenta la metodología para convertir la taxonomía de MDs en un sistema de clasificación. De los resultados (Sección 4) destacamos por un lado una alta precisión en un experimento de detección de errores en la taxonomía (93 % precisión y 78 % cobertura con 145 detecciones en 805 ensayos) y, por otro lado, en la clasificación de MDs (96 % precisión y 98 % cobertura en 619 ensayos). Las conclusiones del artículo y la discusión sobre los próximos pasos a seguir se encuentran en la Sección 5. Un sitio web acompaña al artículo, con documentación sobre avances y demostradores:

http://www.tecling.com/emad

2 Marco teórico

En principio, los marcadores del discurso son palabras de índole gramatical y no léxica, puesto que se utilizan para formar las construcciones gramaticales y no para representar de manera inmediata la realidad (Cuartero, 2002).

Si bien los MDs suelen mantener los comportamientos de las clases gramaticales de las que provienen, por ejemplo, las propiedades sintácticas (Fraser, 1999; Martín Zorraquino y Portolés, 1999), ni estos rasgos ni los semánticos son siempre suficientes o necesarios, por cuanto los MDs tienen un alcance operativo a nivel de enunciado y no a nivel oracional (Fraser, 1990; Lenk, 1997; Waltereit, 2006). Aceptamos que los MDs configuran una categoría pragmática, como muchos autores señalan (Fraser, 1996; Fraser, 1999; Pons, 1998; Martín Zorraquino y Portolés, 1999), que cumplen un papel fundamental en el procesamiento de la coherencia, la cohesión, la adecuación y la eficacia del discurso (Blakemore, 1987; Fraser, 1990; Montolío, 2001; Bazzanella, 2006; Maschler y Schiffrin, 2015) y que no aportan significado léxico a las proposiciones (Fraser, 1996; Lenk, 1997; Bazzanella, 2006). Desde el punto de vista del enfoque procedimental (Sperber y Wilson, 1986; Blakemore, 1987), los MDs son concebidos como guías de las inferencias del interlocutor. En este sentido, los MDs funcionarían como señales metadiscursivas que señalan la estructura y la organización del discurso para beneficio del interlocutor.

La tarea de clasificarlos debe hacerse, entonces, según criterios funcionales y así han procedido muchos autores, por ejemplo, en el ámbito hispánico (Casado, 1993; Martín Zorraquino y Portolés, 1999; Pons, 2000; Montolío, 2001; Fuentes, 2009, entre otros). Cada uno de ellos, sin embargo, considera distintos elementos, conceptos y propiedades para categorizar los MDs.

2.1 Clasificaciones automáticas de MDs

Una clasificación automática puede suponer la introducción de un instrumento de medida objetivo como medio para superar las discusiones que conlleva la subjetividad inherente al método introspectivo, comúnmente usado en las clasificaciones manuales de MDs. Sin embargo, el principal obstáculo para una clasificación automática es la falta de consenso entre los especialistas sobre cuáles son las propiedades delimitadoras de la clase de los MDs. Alonso, Castellón, y Padró (2002) han atribuido esta falta de consenso a la preeminencia de las aproximaciones de tipo deductivo, con un sesgo importante por una teoría subyacente.

En la dédaca de 1990, se llevaron a cabo las primeras propuestas de mecanismos formales para detectar y sistematizar los MDs (Knott y Dale, 1995; Knott, 1996; Marcu, 1997). Ya en el nuevo siglo, Hutchinson (2004) utilizó métodos de aprendizaje automático supervisado para caracterizar conectores discursivos. Si bien obtuvo resultados de alta presición con respecto a un *gold standard*, el aprendizaje de los modelos dependió de instancias anotadas manualmente, lo que requiere de importante trabajo manual antes de la aplicabilidad del método y conlleva un sesgo relativo a los anotadores.

Un enfoque para solucionar este problema es el que adoptaron Alonso et al. (2002), quienes presentan la construcción de un lexicón computacional de marcadores discursivos, proponiendo la utilización de una técnica de *clustering* para agrupar instancias de uso de conectores extraídas de un gran corpus. El resultado es que los *clusters* obtenidos contienen, principalmente, instancias en las que los conectores tienen un comportamiento sintáctico similar. Si bien esta propuesta soluciona, en parte, los problemas anteriores, el hecho de que la selección de los atributos se haya hecho a partir de información sintáctica, semántica y retórica de un léxico de conectores codificado a mano, implica que las categorías aglomeradas sean apenas corroboradas con datos de corpus y no provengan de ellos de manera emergente.

Más recientemente, Muller et al. (2016) obtuvieron automáticamente *clusters* de conectores fundados empíricamente, basándose en la significación de la asociación entre conectores y pares de predicados verbales en contexto. Tal como en el caso anterior, los conectores se extrajeron de una lista codificada previamente.

2.2 La clasificación inductiva

En nuestro trabajo previo (Robledo, Nazar, y Renau, 2017), presentamos una propuesta de inducción automática de taxonomías de MDs a partir de un corpus paralelo. Propusimos un método para extraer ocurrencias parentéticas (desagregadas de la oración mediante signos de puntuación) de MDs desde un corpus paralelo español-inglés de 1.1 mil millones de tokens para inducir automáticamente categorías de MDs según la similitud entre los elementos, sin recurrir a ningún tipo de anotación previa.

El procedimiento involucró la alineación de los MDs en ambas lenguas aplicando estadísticas de coocurrencia sobre los segmentos alineados del corpus paralelo. Esto proporcionó un listado de MDs equivalentes en las dos lenguas, a partir del cual pudimos obtener los grupos de MDs con un misma función en español. Por ejemplo, si se ha visto en el corpus paralelo una asociación estadística entre las veces en que distintos traductores traducen *además* por *furthermore* y las veces en que se traduce *furthermore* por *asimismo*, suponemos que *además* y *asimismo* cumplen una misma función en el discurso.

Estas asociaciones semántico-pragmáticas fueron plasmadas en una matriz binaria (Tabla 1) que asigna un 0 o un 1 en función de si el MD aparece o no en su lista de equivalentes funcionales.

Sobre esta matriz binaria aplicamos un método de *clustering* aglomerativo con el que

a continuación	**1**	0	0	0	0	0	0	0	0
a su vez	-	**1**	0	0	0	0	0	0	0
a veces	-	-	**1**	0	0	0	0	0	0
actualmente	-	-	-	**1**	0	0	0	0	0
además	-	-	-	-	**1**	0	0	0	**1**
ahora	-	-	-	-	-	**1**	0	0	0
ahora bien	-	-	-	-	-	-	**1**	0	0
al menos	-	-	-	-	-	-	-	**1**	0
asimismo	-	-	-	-	-	-	-	-	**1**
...	-	-	-	-	-	-	-	-	...

Tabla 1: Fragmento de muestra de la matriz binaria en la que se basa el experimento. Es una matriz simétrica, por tanto, las columnas corresponden a los mismos marcadores de las filas y los datos debajo de la diagonal principal son redundantes

obtuvimos un total de 100 *clusters* de MDs en español según la similitud entre los elementos. A modo de ilustración, la Figura 1 muestra un ejemplo de resultado de la inducción de taxonomías de MDs en forma de dendrograma. Esta imagen fue realizada con una pequeña muestra aleatoria de 62 MDs para favorecer la legibilidad. El resultado total es un dendrograma mucho más complejo.

2.3 Polifuncionalidad de los MDs

A pesar del interés que puede presentar una clasificación inductiva por medio del método de *clustering*, encontramos una limitación importante que es que solo permite clasificar cada elemento en una sola categoría, de manera que este método de clasificación no permite dar cuenta de la polifuncionalidad de los MDs.

La polifuncionalidad es un fenómeno que se observa en muchos MDs y refiere a que un MD puede expresar distintos significados pragmáticos (Wierzbicka, 2003). De ahí que se investigue la posibilidad de identificar un significado nuclear, básico y estable, de un MD, del cual procedan matices eventuales y significados más contingentes, específicos del contexto de emisión (Aijmer, Foolen, y Vandenbergen, 2006). Así, por ejemplo, un MD como *es decir* puede funcionar como un reformulador explicativo (Martín Zorraquino y Portolés, 1999; Fuentes, 2009, por ejemplo), cumpliendo una función similar a la de *o sea* o *esto es*, pero, también puede funcionar, en ciertos contextos, como un conector consecutivo, cumpliendo una función similar a la de *por lo tanto* o *en consecuencia*.

Con la presente investigación es posible abordar la limitación anterior, es decir, dar cuenta, en alguna medida, de la polifuncionalidad de los MDs.

3 *Metodología*

Tal como explicamos en la introducción, en este artículo planeamos utilizar una taxonomía ya generada para utilizarla luego como un sistema de clasificación de MDs. Esto lo conseguimos utilizando la matriz M cuyo fragmento se expuso en la Tabla 1. Destacamos que en esta etapa de clasificación utilizamos las mismas métricas que para la generación del *clustering*, esto es, mismas características para generación de los vectores y misma distancia.

3.1 Primer paso: elaboración de una matriz M de asociación entre MDs

Esta matriz traduce cada MD como un vector binario, lo que permite computar cálculos de similitud que reflejan el grado de asociación distribucional entre dos MDs. Dado un input i, que sería un determinado MD que se debe clasificar, procedemos a convertirlo en un vector binario utilizando los mismos pasos que en el caso de los marcadores que ya están en la matriz. El algoritmo 1 expone el pseudocódigo del proceso de vectorización.

Algorithm 1 Vectorización de MDs.

```
Require: un marcador castellano $i$
Buscar $i$ en el corpus paralelo
Generar un conjunto $E$ de equivalentes de $i$ en inglés con el corpus paralelo
for each $j \in E$ do
    Agregar a conjunto $S$ los equivalentes en castellano de $j$
end for
Generar vector binario $\vec{i}$ para $i$ registrando con valor 1 los elementos de $S$
```

El marcador discursivo se puede decir que es el elemento ideal para ser investigado en el corpus paralelo porque es independiente del contenido de los textos, por lo que se los encuentra en abundancia, y calcular la coocurrencia de los marcadores en los segmentos alineados es una tarea que no presenta dificultad. La ecuación 1 muestra el cálculo de coocurrencia, para un marcador en castellano i y un candidato a equivalente en inglés j.

$$cooc(i,j) = \frac{f(i,j)}{\sqrt{f(i)}.\sqrt{f(j)}} \qquad (1)$$

En términos simples, lo que esta medida hace es oponer las veces en que aparecen juntos con las veces en que aparecen separados, y esto da la pauta del grado de asociación de los elementos. Normalmente, un MD en castellano tiene múltiples equivalentes en inglés

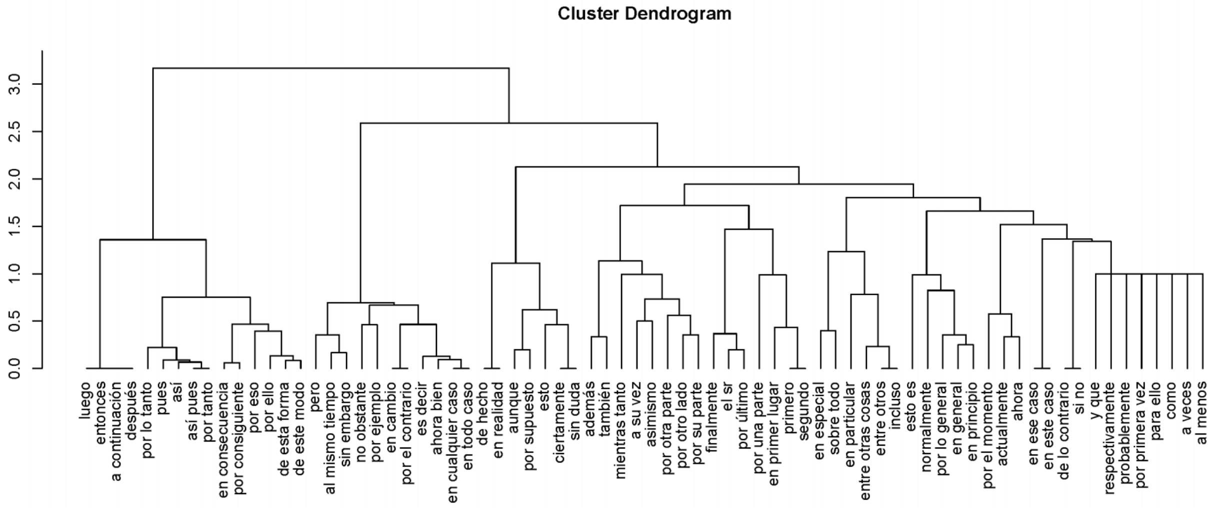

Figura 1: Ejemplo de resultado de inducción automática con una pequeña muestra aleatoria de 62 candidatos a MDs. El dendrograma total contiene 758 MDs, pero la taxonomía final que quedó después de una revisión manual alberga 619 de ellos. Esta figura no tuvo revisión manual y por eso muestra errores, como “el sr”

(o en cualquier otra lengua). Por tanto, de manera general seleccionamos los primeros n equivalentes más probables ($n \approx 3$). Con los equivalentes en inglés (línea 4) hacemos exactamente lo mismo pero al revés, para obtener los equivalentes de nuevo en castellano. La Tabla 2 muestra un fragmento del resultado de esta operación para el caso del marcador contraargumentativo *sin embargo*. Se encuentran equivalentes, y esos equivalentes a su vez arrojan sus propios equivalentes, generando una asociación de los elementos en castellano, que son los que figuran en la tercera columna.

Castellano	Inglés	Castellano
sin embargo	however	sin embargo no obstante ahora bien con todo pero
	nevertheless	no obstante sin embargo con todo a pesar de ello

Tabla 2: Algunos equivalentes en primer y segundo grado (inglés/castellano) de la entrada *sin embargo*

3.2 Segundo paso: utilización de la matriz M para la clasificación de marcadores

Teniendo los elementos descritos en 3.1, procedimos a diseñar un sistema de clasificación, que resumimos en forma de pseudocódigo en el algoritmo 2.

Algorithm 2 Clasificación de MDs.

```
Require: (marcador castellano $i$; taxonomía de marcadores $T$; matriz binaria $M$; umbral de similitud $u \approx 0{,}4$; hash table $S$)
1: if $i \notin M$ then
2:     $\vec{i}$ = vectorizar $i$
3: end if
4: for each $\vec{j} \in M$ do
5:     if $i \neq j \wedge (Jaccard(\vec{i}, \vec{j}) > u)$ then
6:         $S[T(j)]++$
7:     end if
8: end for
9: $T(i) = \arg\max(S)$
```

Esto es, además del elemento a clasificar (i), necesitamos la matriz M y la taxonomía T, que contiene un total de 619 MDs en castellano clasificados en 19 categorías y que hemos revisado manualmente antes de comenzar este proceso para descartar errores. Las categorías de los *clusters* también fueron asignadas de manera manual con nombres que consideramos descriptivos del contenido, tales como “Refuerzo argumentativo” (ej.: *básicamente, en el fondo, en esencia, fundamentalmente, ...*); “Conclusivos” (ej.: *en definitiva, finalmente, para acabar, ...*), etc. Así, si x = *en definitiva*, entonces $T(x)$ = “Conclusivo”.

El algoritmo 2, entonces, compara el vector de i con cada vector j de la matriz M (línea 4 del pseudocódigo). Esa comparación se realiza utilizando el coeficiente de *Jaccard*

(2), una medida apropiada para la comparación de vectores binarios.

$$Jaccard(\vec{i},\vec{j}) = \frac{|\vec{i} \cap \vec{j}|}{|\vec{i} \cup \vec{j}|} \qquad (2)$$

Si la comparación arroja un resultado superior a un umbral arbitrario u (línea 5 del pseudocódigo), entonces se obtiene la categoría de j en la taxonomía T y sumamos 1 a ese valor de la estructura de datos (*hash table*) S (línea 6). Finalmente, seleccionamos la categoría de S con el valor más alto.

3.3 Tercer paso: detección de polifuncionalidad

Para la detección de polifuncionalidad de los MDs hemos procedido de manera similar a la clasificación monocategorial, pero con una clasificación de más de una categoría. Es decir que se procede a seleccionar las n categorías de S con el valor más alto. Para ello, en la ecuación 3 se establece una función $P(x)$ con el criterio para el filtrado de una categoría c.

$$P(c) = \begin{cases} 1 & S[c] > w \wedge \frac{S[c]}{S[T(i)]} > z \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

Los umbrales w y z son arbitrarios. Así, por ejemplo, si $w = 2 \wedge z = 0{,}1$, entonces solo aceptamos una categoría con un valor superior a 2, y si ese valor representa una proporción del 10 % del valor de la categoría ganadora.

4 Resultados

Evaluamos primero la capacidad del algoritmo para distinguir entre lo que es y lo que no es un MD (Sección 4.1). Esto es porque si el algoritmo no puede proporcionar una categoría para un elemento i, se interpreta entonces que i en realidad no es un MD, por tanto es una operación de depuración automática de la taxonomía. Medimos esto a través de la introducción deliberada de errores en la taxonomía. En tanto, para la evaluación de los resultados de la clasificación de MDs (Sección 4.2), procedimos a reclasificar cada uno de los 619 marcadores que están en la taxonomía T y comparar automáticamente si el algoritmo de clasificación asigna a cada marcador la misma categoría que tiene en la taxonomía revisada manualmente, lo que arroja la precisión y la cobertura de la clasificación de MDs. Finalmente, evaluamos de manera cualitativa qué precisión tienen las clasificaciones que hace el algoritmo en segunda instancia (Sección 4.3).

	Detec. errores	**Clasif. marcadores**
Tp	145	587
Fp	10	22
Fn	41	10
Pre	93 %	96 %
Rec	78 %	98 %
F1	85	97

Tabla 3: Resultados de la evaluación: la columna 1 es la detección de “falsos marcadores”, (elementos que no son MDs introducidos de forma deliberada en el listado.), y la columna 2 la clasificación de los MDs asignando una única categoría

4.1 Resultado detección de errores

Para la evaluación del desempeño del algoritmo en la operación de detección de errores procedemos con la taxonomía de MDs T descrita en la Sección 3.

Para evaluar el desempeño de la detección de errores, hemos agregado un total de 186 falsos marcadores, para determinar cuántos de esos errores eran detectados frente a cuántos pasaban desapercibidos para el sistema. En la taxonomía hay entonces 619 MDs correctos y 186 errores, 805 en total.

La Tabla 3 muestra que el resultado arrojó una alta precisión en la detección de errores (93 % precisión, con 10 falsos positivos después de 805 ensayos). De esta manera hemos conseguido también depurar los resultados de nuestra taxonomía anterior, lo que nos permite entrar en un círculo de reproducción de los experimentos entrenando al clasificador nuevamente con una taxonomía más depurada.

4.2 Resultado clasificación de MDs

Para evaluar el desempeño en la clasificación, realizamos un total de 619 ensayos de clasificación tomando cada vez uno de los MDs y determinando si es posible clasificarlo en función de los 618 MDs ya clasificados, estrategia conocida como *leave-one-out cross validation* (Mitchell, 1997, p. 235). Con este procedimiento resulta sencillo evaluar los resultados de la monoclasificación, aquella en la que cada MD recibe una sola categoría, ya que se procede automáticamente mediante el contraste con el listado inicial.

En la Tabla 3 se muestran los resultados de esta clasificación. Como se puede ver, los resultados presentan una alta precisión en la clasificación de MDs asignando una sola categoría (96 % precisión en 619 ensayos).

4.3 Resultado detección de polifuncionalidad

En esta tercera parte de la evaluación medimos el desempeño del algoritmo en el descubrimiento de segundas categorías propuestas por el clasificador, en tanto podrían ser indicativas de polifuncionalidad. La evaluación de la policlasificación, donde un mismo MD es asignado a distintas categorías, resulta más compleja. Aquí solo podemos proceder mediante el análisis cualitativo, aceptando o descartando las propuestas de clasificación del algoritmo según nuestro conocimiento de la materia.

Debido a que estos experimentos se hicieron en función de dos parámetros w y z, en la Tabla 4 se exponen los resultados obtenidos según se modifiquen esos valores. La Ecuación 3 define w como la intensidad de la asociación entre los elementos y z como la proporción que existe entre la primera y la segunda categoría.

Precisión en detección de polifuncionalidad						
%	w=1	w=2	w=3	w=4	w=5	w=6
z=.1	28	28	31	28	25	50
z=.2	35	28	31	29	29	50
z=.3	36	28	33	30	25	50
z=.4	32	29	30	30	27	**75**
z=.5	30	30	30	30	27	75
z=.6	31	31	31	31	27	75
z=.7	33	33	33	33	28	66
z=.8	66	66	66	66	66	0

Tabla 4: Resultados de la evaluación del proceso de policlasificación de los MDs, es decir, de aquellos casos en los que el clasificador indica que hay otra categoría, además de la principal

En el mejor de los casos, cuando el clasificador asigna una segunda categoría a un MD, muestra una precisión de 75 % en 4 ensayos. El número bajo de ensayos se explica porque a mayor z y w, menor cantidad de ensayos posibles. Por ejemplo, en el caso de $z = 0{,}1$ y $w = 1$ tenemos 99 ensayos y en $z = 0{,}4$ y $w = 4$, tenemos 23. Como es natural, a medida que es más restrictivo, menos arriesga.

No evaluamos cobertura por no disponer de un marco de referencia, que podría ser el relevamiento manual previo de los elementos polifuncionales que existen actualmente en la taxonomía T, tarea que dejamos para el futuro.

5 *Conclusiones*

En este artículo hemos presentado una propuesta para clasificar MDs en categorías funcionales generadas inductivamente en un estudio exploratorio previo. Utilizamos los resultados de ese estudio como datos de entrenamiento con el fin de generar un sistema que permita asignar un MD no incluido en la taxonomía inicial a una o más categorías emergentes. Con esto, hemos superado la limitación inicial de asignar un elemento a solo una categoría, con lo cual hacemos una primera aproximación al estudio de la polifuncionalidad de los MDs desde un enfoque inductivo.

A diferencia de trabajos anteriores que han propuesto clasificaciones automáticas de MDs (Alonso, Castellón, y Padró, 2002; Alonso et al., 2002; Hutchinson, 2004; Muller et al., 2016), no hemos partido de ninguna lista de marcadores codificada previamente y abarcamos no solo elementos de conexión, sino una variedad más amplia de MDs. Hemos destacado ya, además, la independencia de la lengua debido a la naturaleza cuantitativa del método. Con esto, creemos que este estudio complementa las clasificaciones previas con una aproximación derivada naturalmente de datos de la lengua en uso.

Además de consecuencias teóricas, este trabajo tiene varias aplicaciones prácticas tales como la segmentación discursiva, la extracción de información o la traducción automática, debido a que los MDs son importantes señales de la estructura del discurso (Popescu-Belis y Zufferey, 2006). Una vía de trabajo futuro será mejorar los resultados de clasificación múltiple de MDs con métodos de aprendizaje automático.

Agradecimientos

Este trabajo ha sido posible gracias a una Beca Doctoral Conicyt otorgada por el Gobierno de Chile al primer autor. Agradecemos también a los revisores por sus comentarios.

Bibliografía

Aijmer, K. 2002. *English discourse particles: Evidence from a corpus.* John Benjamins.

Aijmer, K., A. Foolen, y A.-M. Vandenbergen. 2006. Pragmatic markers in translation: a methodological proposal. En K. Fischer, editor, *Approaches to discourse particles.* Elsevier, páginas 101–114.

Alonso, L., I. Castellón, K. Gibert, y L. Padró. 2002. An empirical approach to discourse markers by clustering. En *Proceedings of the 5th Catalonian Conference on AI: Topics in Artificial Intelligence*, páginas 173–183.

Alonso, L., I. Castellón, y L. Padró. 2002. Lexicón computacional de marcadores del discurso. *Procesamiento del lenguaje natural*, 29:239–246.

Bazzanella, C. 2006. Discourse markers in italian: towards a 'compositional' meaning. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 449–464.

Blakemore, D. 1987. *Semantic constraints on relevance*. Blackwell.

Brinton, L. J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Mouton de Gruyter.

Casado, M. 1993. *Introducción a la gramática del texto del español*. Arco/Libros.

Cuartero, J. 2002. *Conectores y conexión aditiva. Los signos incluso, también y además en español actual*. Gredos.

Fischer, K. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 1–20.

Fraser, B. 1990. An approach to discourse markers. *Journal of pragmatics*, 14(3):383–398.

Fraser, B. 1996. Pragmatic markers. *Pragmatics*, 6(2):167–190.

Fraser, B. 1999. What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Fraser, B. 2006. Towards a theory of discourse markers. En K. Fischer, editor, *Approaches to discourse particles*. Elsevier, páginas 189–204.

Fuentes, C. 2009. *Diccionario de conectores y operadores del español*. Arco/Libros.

Hutchinson, B. 2004. Acquiring the meaning of discourse markers. En *Proceedings of the 42nd Annual Meeting on ACL*, página 684.

Jucker, A. y Y. Ziv. 1998. *Discourse Marker: Description and Theory*. John Benjamins.

Knott, A. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. tesis, University of Edinburgh.

Knott, A. y R. Dale. 1995. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, 18(1):35–62.

Lenk, U. 1997. Discourse markers. En J. Verschueren, editor, *Handbook of pragmatics. Installment*. John Benjamins, páginas 1–17.

Loureda, Ó. y E. Acín. 2010. Cuestiones candentes en torno a los marcadores del discurso en español. En Ó. Loureda y E. Acín, editores, *Los estudios sobre marcadores del discurso en español, hoy*. Arco/Libros, páginas 7–59.

Marcu, D. 1997. From discourse structures to text summaries. En *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, páginas 82–88.

Martín Zorraquino, M. A. y J. Portolés. 1999. Los marcadores del discurso. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española, Vol. 3*. Espasa-Calpe, páginas 4051–4213.

Maschler, Y. y D. Schiffrin. 2015. Discourse markers: Language, meaning, and context. En D. Tannen H. E. Hamilton, y D. Schiffrin, editores, *The handbook of discourse analysis*. John Wiley & Sons, páginas 189–221.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Montolío, E. 2001. *Conectores de la lengua escrita: contraargumentativos, consecutivos, aditivos y organizadores de la información*. Ariel.

Muller, P., J. Conrath, S. Afantenos, y N. Asher. 2016. Data-driven discourse markers representation and classification. En *TextLink–Structuring Discourse in Multilingual Europe. Károli Gáspár University of the Reformed Church in Hungary, Budapest*, página 93.

Pons, S. 1998. *Conexión y conectores: estudio de su relación en el registro informal de la lengua. Anejo XXVII de la revista Cuadernos de filología*. Universitat de València.

Pons, S. 2000. Los conectores. En A. Briz y Val.Es.Co, editores, *¿Cómo se comenta un texto coloquial?* Ariel, páginas 193–220.

Popescu-Belis, A. y S. Zufferey. 2006. Contrasting the automatic identification of two discourse markers in multiparty dialogues. *ISSCO Working Paper 65*.

Robledo, H., R. Nazar, y I. Renau. 2017. Un enfoque inductivo y de corpus para la categorización de los marcadores del discurso en español. En *Proceedings of the 5th International Conference "Discourse Markers in Romance Languages: Boundaries and Interfaces"*, páginas 91–93. Université Catholique de Louvain, Belgium.

Schiffrin, D. 1987. *Discourse markers*. Cambridge University Press.

Sperber, D. y D. Wilson. 1986. *Relevance: communication and cognition*. Harvard University Press.

Waltereit, R. 2006. The rise of discourse markers in italian: a specific type of language change. En K. Fischer, editor, *Approaches to discourse particle*. Elsevier, páginas 61–76.

Wierzbicka, A. 2003. *Cross-Cultural Pragmatics: The Semantics of Human Interaction*. Mouton/de Gruyter.