

Hacia una clasificación verbal automática para el español: estudio sobre la relevancia de los diferentes tipos y configuraciones de información sintáctico-semántica*

Towards an automatic verb classification for Spanish: study of the relevance of different types and configurations of syntactico-semantic information

Lara Gil-Vallejo
Universitat Oberta de Catalunya
lgilva@uoc.edu

Irene Castellón
Universitat de Barcelona
icastellon@ub.edu

Marta Coll-Florit
Universitat Oberta de Catalunya
mcollfl1@uoc.edu

Jordi Turmo
Universitat Politècnica de Catalunya
turmo@lsi.upc.edu

Resumen

En este trabajo nos centramos en la adquisición de clasificaciones verbales automáticas para el español. Para ello realizamos una serie de experimentos con 20 sentidos verbales del corpus Sensem. Empleamos diferentes tipos de atributos que abarcan información lingüística diversa y un método de clustering jerárquico aglomerativo para generar varias clasificaciones. Comparamos cada una de estas clasificaciones automáticas con un gold standard creado semi-automáticamente teniendo en cuenta construcciones lingüísticas propuestas desde la lingüística teórica. Esta comparación nos permite saber qué atributos son más adecuados para crear de forma automática una clasificación coherente con la teoría sobre construcciones y cuales son las similitudes y diferencias entre la clasificación verbal automática y la que se basa en la teoría sobre construcciones lingüísticas.

Palabras clave

Clasificación verbal, clustering, construcciones

Abstract

In this work we focus on the automatic acquisition of verbal classifications for Spanish. To do so, we perform a series of experiments with 20 verbal senses that belong to the Sensem corpus. We use different kinds of features that include diverse linguistic information and an agglomerative hierarchical clustering method to generate a number of classifications. We compare each of these automatic classifications with a semi-automatically created gold standard, which is

*Este trabajo ha sido realizado en el marco del proyecto Adquisición de escenarios de conocimiento a través de la lectura de textos (SKATeR, TIN2012-38584-C06-01) y gracias a una beca de investigación de la escuela de doctorado de la Universitat Oberta de Catalunya.

built on the basis of linguistic constructions proposed by theoretical linguistics. This comparison allows us to investigate which features are adequate to build a verb classification coherent with linguistic constructions theory and which are the similarities and differences between an automatic verbal classification and a verb classification based on the theory of linguistic constructions.

Keywords

Verb classification, clustering, constructions

1 Introducción

Los lexicones computacionales tienen un gran valor dentro del área del Procesamiento del Lenguaje Natural. VerbNet (Schuler, 2005) ha sido empleado en múltiples tareas, como por ejemplo etiquetación de papeles semánticos (Giuglea y Moschitti, 2006), elaboración de sistemas de diálogo automático (Swift, 2005) o desambiguación de sentidos verbales (Brown, Dligach, y Palmer, 2014). El modelo de lexicon de VerbNet presenta además la ventaja de estar organizado por clases. Las clases verbales estructuran información relativa al verbo y a sus argumentos, lo que permite eliminar información redundante y elaborar generalizaciones (Schulte im Walde, 2006). Por ejemplo, la clase *appear-48.1.1* de VerbNet contiene 41 verbos que comparten esquemas sintáctico-semánticos, lo que permite usar los atributos asociados a la clase en tareas de Procesamiento del Lenguaje Natural, generalizando la información que aporta cada verbo individualmente.

Sin embargo, la elaboración manual de un lexicon es costosa y requiere bastante tiempo y re-

cursos que en ocasiones no están disponibles. Por ello, en los últimos años se han realizado varios experimentos y trabajos con el objetivo de adquirir un lexicón verbal de forma automática o semi-automática que pueda aplicarse satisfactoriamente a diversas tareas. En concreto, se ha utilizado texto anotado a diferentes niveles o lexicones de subcategorización como VALEX (Korhonen, Krymolowski, y Briscoe, 2006) para crear clasificaciones verbales automáticas asociadas a información sintáctico-semántica.

El objetivo de este trabajo es averiguar qué atributos lingüísticos son más adecuados para una clasificación sintáctico-semántica automática de verbos para el español usando técnicas de clustering, con el fin de hacer una selección de los mismos y aplicarlos posteriormente en una clasificación más amplia de unidades verbales. Para este fin, hemos realizado diversos experimentos con varias clasificaciones verbales. Para obtener estas clasificaciones verbales hemos escogido un conjunto controlado de verbos que presentan diferentes iniciadores, campos semánticos y esquemas sintácticos y hemos empleado varios tipos de atributos y un algoritmo de clustering para crear la clasificación. Los atributos contienen información lingüística sintáctico-semántica (funciones sintácticas, roles semánticos, preferencias selectivas, entre otros). Además de utilizar diferente tipo de información, hemos experimentado con diversas configuraciones de los rasgos lingüísticos y diferentes tipos de valor de los atributos. En cuanto al algoritmo, elegimos el clustering jerárquico aglomerativo, ya que es coherente con las clasificaciones verbales manuales, que son taxonómicas y que recogen la idea de la existencia de diferentes grados de similitud entre los miembros de las clases. Por otro lado, consideramos que es interesante poder observar la distribución de los sentidos verbales por clases en función del nivel de la jerarquía escogido.

2 Trabajos previos

Las clasificaciones verbales automáticas se elaboran generalmente a partir de la aplicación de un algoritmo supervisado o no supervisado a datos extraídos de un corpus. Presentan una serie de ventajas e inconvenientes sobre las manuales. Como desventaja podemos apuntar el hecho de que, al ser generadas automáticamente a partir de información de corpus, pueden contener ruido o clases no del todo coherentes, frente a la precisión que podemos encontrar en una clasificación manual. Por otro lado, las clasificaciones verbales automáticas pueden alcanzar una gran cobertu-

ra con un coste mínimo. El número de propuestas de clasificaciones verbales automáticas creció considerablemente a partir del trabajo teórico de clasificación verbal de Levin (1993), en el que se basa VerbNet, uno de los lexicones verbales más empleados en Procesamiento del Lenguaje Natural. La hipótesis de Levin es que el significado de un verbo determina su comportamiento en cuanto a la expresión e interpretación de sus argumentos. Esta hipótesis ha sido la base para muchas de las propuestas de clasificación verbal automática. Por lo tanto, la mayor parte del trabajo realizado en el área de clasificación verbal automática tiene por objetivo crear unas clases verbales similares a las que propone Levin. Para ello exploran diferentes características lingüísticas y algoritmos de clustering. A continuación ofrecemos un panorama general del trabajo realizado en esta área, tanto el que está basado en las clasificaciones verbales de Levin, como aquellas propuestas que tienen como objetivo adquirir otro tipo de clasificación verbal.

Con relación a aquellos trabajos cuyo objetivo es adquirir una clasificación similar a la de Levin (y que, por tanto, usan adaptaciones o traducciones de la clasificación de Levin como gold standard) podemos diferenciar entre aquellos que usan un enfoque supervisado y los que usan un enfoque no supervisado (clustering). Ambos tipos modelan los verbos basándose en un conjunto de características lingüísticas orientadas a capturar las alternancias de diátesis en las que Levin basa su clasificación. Sin embargo, en el caso de los enfoques no supervisados, la clase a la que pertenece un verbo no es conocida a priori.

En cuanto a los enfoques no supervisados, que será nuestra perspectiva, generalmente emplean patrones de subcategorización en combinación con diferentes algoritmos, como por ejemplo Joanis, Stevenson, y James (2008) y Li y Brew (2008). Los patrones de subcategorización enriquecidos con preferencias selectivas han demostrado dar lugar a una mayor precisión a la hora de inducir las clases de Levin como vemos en Sun y Korhonen (2009) y Vlachos, Korhonen, y Ghahramani (2009)

Este método para realizar clasificaciones verbales también se ha empleado para otras lenguas diferentes del inglés. Para evaluar estas clasificaciones se han empleado diferentes métodos: Brew y Schulte im Walde (2002) y Schulte im Walde (2006) crean un gold standard manual para el alemán, mientras que Falk, Gardent, y Lamirel (2012) construyen automáticamente una base de datos para el francés con criterios similares a la de VerbNet. Otra alternativa común es la de tra-

ducir las clases de Levin, lo que permite una comparación entre los resultados en ambos idiomas. Sun y Korhonen (2009) obtienen para el francés una medida-F de 54.6 (la medida-F para el equivalente inglés es de 80.4). Scarton et al. (2014) obtienen una medida-F de 42.77 para el portugués brasileño. En ambos casos los atributos que mejor funcionan son los patrones de subcategorización enriquecidos con preferencias selectivas y preposicionales. Para el español, Ferrer (2004) aplica un clustering jerárquico a 514 verbos y los evalúa con la clasificación manual de Vázquez, Fernández, y Martí (2000). Usa probabilidades de diferentes tipos de patrones de subcategorización, obteniendo una medida Rand de 0.07 para 15 clusters.

Una aproximación diferente es el trabajo de Sun, McCarthy, y Korhonen (2013), que no emplea patrones de subcategorización, sino que propone un método alternativo para capturar las alternancias de diátesis de los verbos, basándose en la idea de que una alternancia de diátesis puede aproximarse calculando la probabilidad conjunta de dos patrones de subcategorización.

Entre aquellos trabajos que se apartan del objetivo de adquirir una clasificación verbal similar a la de Levin, podemos mencionar la propuesta de Merlo y Stevenson (2001), que utiliza un enfoque supervisado para clasificar verbos en tres grupos: inacusativos, inergativos y de objeto nulo. Finalmente, cabe mencionar también el trabajo de Lenci (2014), cuyo objetivo es descubrir clases verbales. Para ello usa patrones de subcategorización y preferencias selectivas en un corpus del italiano, empleando uno de estos patrones de subcategorización como semilla para después hacer particiones según rasgos más específicos entre los verbos que lo contienen.

En general los trabajos mencionados asignan los lemas verbales a una sola clase, lo que no permite dar cuenta de la polisemia verbal. Este factor puede ser muy importante, ya que la mayoría de los verbos tienen al menos dos sentidos. Al modelar un verbo sin tener en cuenta sus sentidos puede obtenerse un modelo poco preciso, ya que en realidad la mayor parte de la información se obtiene del sentido más frecuente, mientras que aquellos sentidos menos frecuentes quedan sin modelar o distorsionan el modelo (Korhonen, Krymolowski, y Marx, 2003).

3 Metodología

A continuación explicamos la metodología que hemos seguido en este trabajo. En primer lugar, detallamos los criterios para seleccionar los sen-

tidos verbales para los experimentos (3.1). Seguidamente explicamos el proceso de creación de un gold standard (3.2), tomando construcciones lingüísticas propuestas desde la lingüística teórica. El gold standard es una referencia con la que se pueden comparar las clasificaciones verbales automáticas para comprobar si se obtienen clases equivalentes. A continuación (3.3), explicamos el proceso de extracción de información lingüística del corpus para generar los datos que sirven de base para los experimentos. Además, en este mismo apartado explicamos el tipo de algoritmo de clustering que empleamos para elaborar las diferentes clasificaciones verbales automáticas.

3.1 Selección de sentidos verbales

En nuestro trabajo hemos optado por realizar experimentos con sentidos verbales, en vez de lemas, para obtener modelos más precisos. En concreto, trabajamos con un único sentido por verbo, esto es, no incluimos pares polisémicos para poder modelizar sin ambigüedad. No obstante, reconocemos que el fenómeno de la polisemia verbal es algo que se ha de tener en cuenta y tratar en cualquier aplicación computacional.

Se escogen 20 sentidos verbales del corpus Sensesem (Fernández-Montraveta y Vázquez, 2014) que aparecen con una frecuencia mayor de 10 frases en el corpus para asegurar la representatividad de las diferentes propiedades sintáctico-semánticas asociadas con los sentidos. Estos 20 sentidos presentan diferentes esquemas sintácticos, pertenecen a diferentes campos semánticos, correspondientes a los supersenses de Wordnet asociados a los synsets del Multilingual Central Repository (Gonzalez-Agirre y Rigau, 2013) y poseen diferentes tipos de iniciadores del evento: causativos, agentivos y experimentadores. Estas tres características permiten que el conjunto escogido sea representativo, pese al limitado número de sentidos verbales. A continuación mostramos la clasificación de los sentidos seleccionados según el campo semántico al que pertenecen:¹

- estado: parecer_1, valer_1, estar_14.
- comunicación: valorar_2, explicar_1.
- cognición: gustar_1, pensar_2.
- movimiento: perseguir_1, viajar_1, volver_1, montar_2.
- cambio: abrir_18, cerrar_19, crecer_1, morir_1.

¹Para una definición de los sentidos verbales y número de ocurrencias de cada uno en el corpus se puede consultar el anexo A

- percepción: ver_1, escuchar_1.
- actividad (social y corporal): trabajar_1, dormir_1, gestionar_1.

3.2 Creación del gold standard

A continuación detallamos el proceso de creación del gold standard, que es una clasificación verbal basada en propuestas teóricas sobre construcciones. Definimos la noción de *construcción* como un signo lingüístico, con forma y significado, que comprende estructura sintáctica y roles semánticos. Esta definición es coincidente con la noción de construcción de Goldberg (1994) y la de diátesis de Levin. Esta clasificación servirá para evaluar las clasificaciones creadas automáticamente, lo que permitirá escoger los atributos adecuados para crear una clasificación automática similar a una clasificación manual, mucho más costosa de realizar.

El primer paso en la creación del gold standard es seleccionar los atributos lingüísticos que configurarán las clases verbales. En nuestro caso hemos utilizado estructuras sintácticas básicas descritas en múltiples gramáticas como Barreto y Bosque (1999). Además, hemos empleado construcciones adaptadas de Levin, teniendo en cuenta los trabajos de Cifuentes Honrubia (2006) y Vázquez, Fernández, y Martí (2000). Tomamos las construcciones aisladas, es decir, cada uno de los pares en una alternancia de diátesis, lo que no impone restricciones sobre el tipo de alternancia en el que participan los verbos. Dado que la cantidad de sentidos escogidos es limitada para controlar el efecto de los diferentes atributos, se han preferido aquellas construcciones que tienen un carácter más general sobre aquellas específicas para determinados verbos. A continuación listamos y explicamos brevemente estas estructuras y construcciones. Empleamos como atributos cinco estructuras sintácticas básicas: transitiva, intransitiva, ditransitiva, predicativa y atributiva; además, contamos con trece construcciones:

1. Causativa prototípica: Construcción en la que se explicita la causa de un evento por medio de un sujeto. El sujeto puede ser un agente (volitivo) o una causa (no volitiva). El objeto está afectado por el evento en diversos grados. Ej. *La falta de lluvias secó el río*
2. Anticausativa prototípica (con “se”): Es una construcción intransitiva donde la entidad afectada ocupa la posición de sujeto. Ej. *El río se secó*

3. Causativa de perífrasis: Es una causativa en la que el predicado aparece en infinitivo junto con el auxiliar “hacer”. Ej. *Los fuertes vientos han hecho bajar las temperaturas*
4. Anticausativa sin “se”: el constituyente que expresa la causa se elide. Una entidad no afectada ocupa la posición de sujeto. Ej. *Las temperaturas han bajado*
5. Voz media: Expresa un estado o propiedad del sujeto sin combinarse con un verbo atributivo. Generalmente van con un complemento adverbial que refuerza la lectura estativa, a diferencia de la anticausativa prototípica, que tiene una interpretación dinámica. Ej. *La pintura se esparce con facilidad*.
6. Impersonal pronominal: El verbo aparece en tercera persona, no tienen sujeto gramatical explícito ni recuperable por el contexto. Ej. *Se aconseja el uso obligatorio del cinturón*
7. Sujeto oblicuo: El iniciador del evento aparece en una posición encabezada por una preposición. Se suele subdividir en varios tipos, pero dado que nuestro número de ejemplos es pequeño, no hemos tenido en cuenta estas subdivisiones. Ejs. *La gente se beneficia de las nuevas medidas*
8. Reflexiva: La acción expresada por el sujeto recae sobre sí mismo. Ej. *María se peina*.
9. Recíproca: El sujeto de estas construcciones es plural. Cada uno de los componentes del sujeto ejerce una acción sobre los otros, a la vez que la recibe de los demás. Ej. *Juan y Pedro se desafiaron*.
10. Pasiva perifrástica: El objeto ocupa una posición topicalizada y el verbo se construye con un auxiliar. Generalmente el agente se puede expresar mediante un sintagma preposicional. Ej. *Los bizcochos fueron comidos por los niños*
11. Pasiva refleja: Se construye con la partícula “se”. El sujeto se pospone a la partícula. El iniciador de la acción no se explicita pero suele ser agente. Ej. *Se pasaron los trabajos a ordenador*
12. Objeto cognado: El objeto que mantiene una relación etimológica con el verbo, por ello las frases con esta construcción tienen un sentido tautológico. Ej. *Cantamos una canción*
13. Resultativa con “estar”: Detalla el estado resultado de la acción expresada por el verbo. Ej. *El pan está cortado*

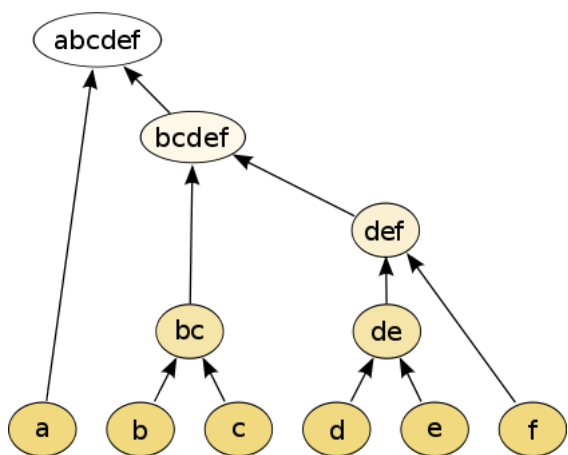


Figura 1: Modelización del clustering jerárquico aglomerativo (adaptado de la entrada de agrupamiento jerárquico de Wikipedia)

Para crear las clases verbales del gold standard aplicamos un clustering jerárquico aglomerativo junto con estos atributos y los sentidos verbales descritos. En el clustering jerárquico aglomerativo cada elemento (en nuestro caso sentidos verbales) pertenece inicialmente a un grupo. En cada paso se van fusionando los dos grupos con menor distancia (ver figura 1). La distancia entre dos grupos se calcula aplicando una función de distancia entre algunos de sus elementos (por ejemplo, distancia euclídea, distancia del coseno, etc.). La selección de dichos elementos se puede realizar de formas diferentes, que se definen como tipos de enlace. En nuestros experimentos hemos empleado cuatro tipos de enlace distintos (simple, completo, promedio y promedio ponderado) para comprobar el efecto que tienen en las agrupaciones de los verbos, con lo cual obtenemos un gold standard para cada tipo de enlace.

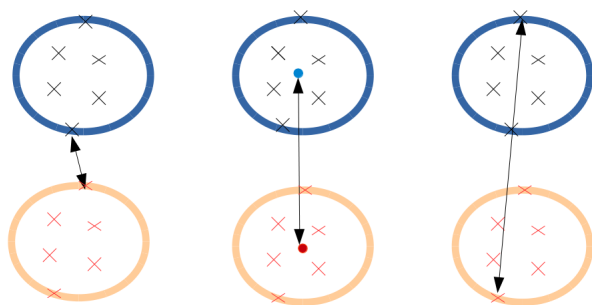


Figura 2: Enlace simple, promedio y completo

En la figura 2 podemos ver cómo se calcula la distancia entre grupos en los tres tipos de enlace: de izquierda a derecha mostramos el enlace simple, el enlace promedio (el promedio ponderado es una variante de este) y el enlace completo. En el enlace simple la distancia entre dos grupos viene dada por la mínima distancia entre los elementos

de ambos grupos. En el enlace promedio la distancia entre dos grupos se calcula como promedio de la distancia entre todos los pares de elementos de ambos grupos. En el enlace promedio ponderado la distancia entre dos grupos se define de la misma manera que en el caso del enlace promedio, pero se tienen en cuenta los grupos previos que pasaron a formar parte de los grupos actuales. Finalmente, en el enlace completo la distancia entre dos grupos se calcula teniendo en cuenta los elementos más dispares de ambos grupos.

El resultado del clustering jerárquico es una jerarquía de posibles agrupaciones, cada una de ellas definida por cada nivel de la jerarquía. Una vez obtenido el resultado del clustering, se debe decidir el nivel de agrupación más apropiado. Para ello, tres lingüistas evaluaron las distintas agrupaciones que contenían de 4 a 10 clases y finalmente, después de varias reuniones de discusión, se llegó al acuerdo de que el modelo de 6 clases era el más adecuado, ya que en él hay una serie de agrupaciones de los sentidos verbales comunes para los cuatro tipos de enlace que son coherentes con la teoría lingüística. Aparecen siempre en la misma clase los sentidos de carácter estativo *estar_14* y *parecer_1*. En otra clase aparecen juntos *abrir_18*, *cerrar_19*, *crecer_1* y *morrir_1*, que son verbos que expresan cambio (junto con ellos aparece también *dormir_1*, que se trata de una actividad). También juntos en una clase se agrupan *escuchar_1*, *explicar_1*, *gestionar_1*, *perseguir_1*, *ver_1* y *valorar_2*, que generalmente tienen iniciadores humanos u organizaciones. *Trabajar_1* y *volver_1*, intransitivos agentivos, también permanecen juntos en todos los tipos de enlace y en ocasiones se agrupan con otros verbos. *Valer_1* y *gustar_1* siempre son miembros únicos de su grupo. Los demás alternan entre los grupos ya mencionados. Las clases resultantes pueden consultarse en la columna izquierda del anexo B.

3.3 Experimentación

En los experimentos se emplean diferentes atributos lingüísticos extraídos del corpus Sensem:

- atributos semánticos de los argumentos:
 - Roles semánticos obtenidos a partir de un mapping jerárquico realizado entre los roles de Sensem y la propuesta de Lyrics (Bonial et al., 2011):
 - roles semánticos finos (40 roles),
 - roles semánticos abstractos (16 roles);

- Supersenses de Wordnet (Miller, 1995) (45 supersenses);
- Ontología de SUMO (Niles y Pease, 2003) (1000 términos). Los supersenses y los términos de la ontología de SUMO se obtienen a partir del núcleo de los argumentos verbales, que en Sensem están anotados con synsets.

- atributos morfosintácticos: función sintáctica; categoría morfológica; construcción, que recoge aspectos como la topicalización o destopicalización del sujeto lógico, la reflexividad o la impersonalidad.
- aspecto oracional (estado, evento, proceso).

Mediante la selección de esta información configuramos diferentes espacios de atributos para los experimentos. Con el fin de obtener una representación lo más completa posible de los predicados, cada atributo semántico se combina con uno sintáctico. Por otro lado, para explorar el rol del aspecto, que no se ha tenido en cuenta generalmente a la hora de elaborar clasificaciones verbales automáticas, realizamos una versión de estos atributos combinada con el aspecto de las frases. Finalmente, para valorar el potencial de los roles semánticos a la hora de definir una clasificación verbal, añadimos otro atributo que consiste en roles semánticos sin combinarlos con información sintáctica.

Como resultado tenemos 27 tipos de atributos según el tipo de información lingüística que recogen (por ejemplo, *sintaxis+supersenses*, *sintaxis+roles de sensem*, *categoría morfosintáctica+ontología SUMO+aspecto*, etc). A su vez, estos atributos admiten tres configuraciones diferentes de información: rasgos aislados, constituyentes y patrones. En la figura 3 presentamos un ejemplo de anotación de la frase en Sensem *Remedios abrió su bolso*.

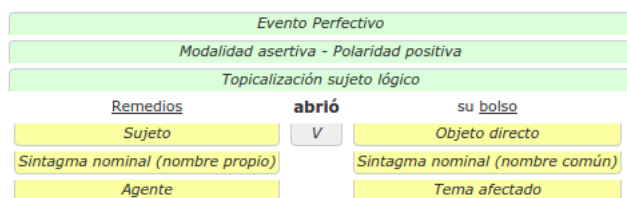


Figura 3: Anotación de una frase en el corpus Sensem

Para esta frase, con información lingüística relativa a roles y funciones sintácticas, obtendríamos las siguientes configuraciones:

- rasgos aislados (4 atributos): sujeto, agente, objeto directo, tema afectado

- constituyentes (2 atributos): sujeto-agente, objeto directo-tema afectado
- patrones (1 atributo): sujeto-agente+objeto directo-tema afectado

Por lo tanto, los 20 sentidos tomados de Sensem quedan caracterizados por los atributos sintáctico-semánticos de las frases en las que participan. En cuanto al valor del atributo, experimentamos con dos tipos diferentes: binarios (0/1) y probabilidades. Los atributos binarios toman valores 0 o 1 dependiendo de si para un sentido y un atributo dados (por ejemplo *abrir_18* y *sujeto-agente+objeto-tema*) hay al menos una frase que recoja ambos (1) o no la hay (0). Por otro lado, las probabilidades expresan, para un sentido y un atributo dados (por ejemplo *abrir_18* y *sujeto-agente+objeto-tema*), la proporción de frases en las que aparecen ambos en relación a las frases en las que participa el sentido verbal.

Para cada set de datos con un tipo de atributo obtenido mediante las combinaciones que acabamos de explicar, realizamos clustering jerárquico aglomerativo con cada uno de los cuatro enlaces posibles. Las funciones de distancia empleadas son dos: una basada en el coeficiente Dice (Dice, 1945), ya que es adecuada y ampliamente empleada para atributos binarios y otra basada en el Coseno para los probabilísticos, también muy utilizada en estos casos. Estas medidas se emplean para calcular la distancia entre dos elementos en función de los valores asociados a los atributos que los caracterizan. El número de clases deseadas en el resultado del clustering que compararemos con el gold standard se establece en un rango entre 4 y 10.

4 Evaluación y resultados

Para evaluar los resultados, comparamos cada gold standard correspondiente a un tipo de enlace con los resultados de los experimentos para este enlace. En las tablas 1, 2, 3 y 4 mostramos los resultados para cada tipo de enlace. Para cada variación de atributo-valor (rasgos aislados, constituyentes y patrones con valores probabilísticos y binarios) se muestra el número de clases y el tipo de información lingüística que conforman la clasificación automática más similar al respectivo gold standard. La similitud entre el gold standard y el resultado de cada experimento se mide empleando la información mutua ajustada, que da cuenta de la similitud entre dos etiquetados diferentes para los mismos datos. En nuestro caso, los dos etiquetados son las clases verbales definidas en el gold standard y las clases obtenidas

automáticamente con datos de Sensem. La medida de información mutua ajustada oscila entre 0 si las clases verbales son independientes y 1 si las clases verbales son idénticas. Hay una gran variedad de medidas de evaluación externa para algoritmos de clustering. Hemos elegido esta medida porque no presenta sesgos en cuanto al número de clases, al contrario que otras medidas muy utilizadas como la de pureza (Manning et al., 2008) y está ajustada, es decir, en el caso de una agrupación aleatoria de sentidos verbales, el valor de la medida de información mutua ajustada es 0 (Strehl, 2002).

5 Análisis de resultados

Si observamos globalmente los resultados correspondientes a todos los tipos de enlace, vemos que la información lingüística que en más ocasiones contribuye a generar una clasificación similar a la del gold standard es la combinación de supersenses y funciones sintácticas, con una información mutua ajustada media de 0.530 (este tipo de información obtiene mejores resultados en cuatro ocasiones para el enlace completo, dos para el simple y una para el de tipo promedio ponderado). Sin embargo, la combinación de roles abstractos más funciones sintácticas, que es la segunda información lingüística que más frecuentemente aparece en las tablas (dos veces para el enlace promedio, una para el completo, una para el simple y dos para el promedio ponderado) tiene una información mutua ajustada media ligeramente mayor: 0.542. En general observamos que las funciones sintácticas aparecen en muchos de los atributos que mejores resultados obtienen.

Si nos centramos en el tipo de valor, vemos que las probabilidades dan lugar a una información mutua ajustada media mayor que los atributos binarios: 0.55 frente a 0.49. En cuanto a la configuración de los atributos, las configuraciones que generalmente dan lugar a una clasificación más similar a la del gold standard son las de patrones y constituyentes, ambas con una información mutua ajustada media de 0.54. Los rasgos aislados obtienen peores resultados, con un 0.49.

En conjunto, la configuración que mejores resultados arroja es la que contiene información acerca de los supersenses y la función sintáctica organizada en patrones y con valores probabilísticos. Este tipo de atributos y valores en el enlace simple obtiene una medida de información mutua de 0.647. También cabe destacar que la combinación *roles abstractos+aspecto+función sintáctica* obtiene una de las mejores medidas, 0.627, lo que pone de relieve la importancia del aspecto como

información relevante a la hora de crear una clasificación verbal automática.

Como hemos visto en el apartado de trabajos previos, hay una clasificación verbal automática para el español realizada por Ferrer (2004), que consigue una medida Rand ajustada de 0.07 clasificando 514 verbos en 15 grupos. Para tener una referencia, calculamos la medida Rand ajustada de la clasificación verbal generada por la configuración que obtiene una mayor información mutua ajustada. La medida Rand de esta clasificación es de 0.619. Pese a que se trata de un valor notablemente más alto que el que alcanza Ferrer (2004), hay que tener en cuenta que el tipo de gold standard es diferente y la cantidad de verbos es menor en nuestro caso, lo que limita el posible ruido que se generaría con un número mayor de sentidos. Aunque ambas clasificaciones no son directamente comparables, consideramos que los resultados que hemos obtenido son prometedores y nos animan a seguir trabajando en esta línea.

En lo relativo a las clases que se obtienen haciendo clustering con los datos de Sensem, vemos como en las cuatro mejores agrupaciones, una por enlace², hay unos rasgos comunes: de forma similar a lo que ocurre en el gold standard, estar y parecer se mantienen en una misma clase que tampoco contiene ningún otro miembro. Por el contrario, el grupo de verbos que expresaban cambio junto con *dormir_1* no se mantiene. En concreto, *abrir_18* y *cerrar_19* generalmente aparecen en un grupo separado de *crecer_1* y *dormir_1*. En este punto coinciden con la distinción hecha por Levin y Hovav (1995) entre verbos de cambio de estado que expresan un evento de causa externa y aquellos que expresan un evento de causa interna. En Levin y Hovav (1995) se definen los eventos de causa interna como aquellos en los que el argumento que acompaña al verbo posee una propiedad que es responsable del evento denotado (por ejemplo ‘la planta creció’) y los eventos de causa externa como aquellos en los que hay una causa externa que tiene el control del evento (por ejemplo ‘la puerta se abrió’), que además puede ser hecha explícita en una construcción transitiva (por ejemplo ‘el viento abrió la puerta’). *Explicar_1*, *escuchar_1*, *gestionar_1*, *perseguir_1*, *valorar_2* y *ver_1*, que aparecían siempre en el mismo grupo en el gold standard, independientemente del tipo de enlace, se mantienen juntos también en todos los enlaces de las clases obtenidas con datos de corpus. *Valer_1* aparece como único miembro de su grupo en todos los casos, tanto en el gold

²En negrita en las tablas 1-4, los verbos que componen estas clases están en la columna izquierda de las tablas del anexo B

Tabla 1: Enlace promedio

Configuración de los atributos	Valor de los atributos	Información lingüística de los atributos	Número de grupos	Información Mutua Ajustada
rasgos aislados	binario	SUMO aspecto funciones sintácticas	7	0.425
rasgos aislados	probabilidades	roles abstractos construcciones	6	0.598
constituyentes	binario	roles abstractos funciones sintácticas	6	0.591
constituyentes	probabilidades	roles abstractos aspecto funciones sintácticas	6	0.627
patrones	binario	roles abstractos morfología	6	0.598
patrones	probabilidades	roles abstractos funciones sintácticas	7	0.609

Tabla 2: Enlace completo

Configuración de los atributos	Valor de los atributos	Información lingüística de los atributos	Número de grupos	Información Mutua Ajustada
rasgos aislados	binario	SUMO aspecto morfología	7	0.389
rasgos aislados	probabilidades	supersenses funciones sintácticas	8	0.488
constituyentes	binario	supersenses funciones sintácticas	5	0.519
constituyentes	probabilidades	supersenses funciones sintácticas	7	0.479
patrones	binario	roles abstractos funciones sintácticas	6	0.422
patrones	probabilidades	supersenses funciones sintácticas	8	0.551

standard como en los grupos creados a partir de corpus.

Respecto a las diferencias entre el gold standard y las clases obtenidas, vemos que *viajar_1* y *trabajar_1* aparecen siempre juntos en estas clases, mientras que en el gold standard *trabajar_1* aparecía siempre junto con *volver_1*. *Gustar_1*, que en las clases del gold standard aparecían como único miembro de su grupo, aparece en una ocasión en el mismo grupo que *crecer_1*. El resto de los verbos alternan entre dos grupos principales en las clasificaciones hechas con datos de Sensem: *pensar_2* alterna entre el grupo de *escuchar_1* y aislado, *montar_2* alterna entre el grupo de *volver_1* y aislado, *volver_1* alterna entre el grupo de *montar_2* y aislado. Finalmente, *morrir_1* alterna entre el grupo de *abrir_18* y el de

crecer_1, lo que no es consecuente con el criterio de causa externa e interna, ya que de mantenerse este criterio en la clasificación automática debería permanecer con *crecer_1*.

6 Conclusiones

En este trabajo hemos analizado parámetros relevantes a la hora de hacer clasificaciones verbales automáticas empleando clustering jerárquico aglomerativo. Para ello hemos creado un gold standard para cada tipo de enlace de forma semi-automática, utilizando atributos motivados en la teoría lingüística. Posteriormente hemos realizado varios experimentos empleando diferentes tipos de parámetros y hemos analizado los resultados.

Tabla 3: Enlace simple

Configuración de los atributos	Valor de los atributos	Información lingüística de los atributos	Número de grupos	Información Mutua Ajustada
rasgos aislados	binario	SUMO aspecto funciones sintácticas	6	0.567
rasgos aislados	probabilidades	roles abstractos aspecto funciones sintácticas	5	0.590
constituyentes	binario	roles abstractos funciones sintácticas	6	0.561
constituyentes	probabilidades	supersenses funciones sintácticas	6	0.561
patrones	binario	SUMO aspecto funciones sintácticas	6	0.561
patrones	probabilidades	supersenses funciones sintácticas	6	0.647

Tabla 4: Enlace promedio ponderado

Configuración de los atributos	Valor de los atributos	Información lingüística de los atributos	Número de grupos	Información Mutua Ajustada
rasgos aislados	binario	SUMO aspecto funciones sintácticas	6	0.372
rasgos aislados	probabilidades	roles abstractos	7	0.479
constituyentes	binario	supersenses funciones sintácticas	9	0.468
constituyentes	probabilidades	roles abstractos funciones sintácticas	6	0.532
patrones	binario	roles abstractos morfología	5	0.503
patrones	probabilidades	roles abstractos funciones sintácticas	7	0.539

En concreto, para el clustering jerárquico aglomerativo, comprobamos que los diferentes tipos de enlace tienen un efecto en la configuración de las clases. En cuanto al diseño de los atributos, hemos visto como la configuración en patrones y constituyentes ofrece unos resultados mejores que los rasgos aislados. Si tenemos en cuenta la mejor clasificación por enlace, son los patrones los que mejor funcionan, algo que va en la línea de los trabajos previos. En relación con esto, hemos observado que el tipo de valor que recoge de forma más efectiva la información proporcionada por los datos son las probabilidades de co-ocurrencia de verbo y atributo.

En cuanto a la información lingüística, hemos comprobado que las funciones sintácticas tienen un papel fundamental, y que ofrecen buenos re-

sultados combinadas con roles semánticos abstractos o los supersenses de Wordnet. Además hemos demostrado que el aspecto, que generalmente no se ha tenido en cuenta en los trabajos previos, es un rasgo útil. Una inspección manual de las clases nos ha permitido observar la existencia de similitudes básicas globales entre el gold standard y las clases elaboradas con datos de SenseM.

En definitiva, en este trabajo hemos evaluado qué tipo de información sintáctico-semántica es más relevante para una clasificación automática verbal del español, así como el tipo de valor y configuración de los atributos más adecuados, empleando un conjunto acotado y controlado de sentidos verbales. Esto nos ha permitido hacer un estudio de los cambios en la configuración de

las clases según el empleo de diferentes parámetros. A partir de estos resultados, el próximo paso será aplicar los parámetros obtenidos a una clasificación más amplia de unidades verbales del español.

transporte. (111)

volver_1: Dirigirse hacia el lugar donde ya se ha estado.(84)

A Definición de los sentidos verbales

Entre paréntesis se indica el número de ocurrencias en el corpus.

abrir_18: Descorrer el pestillo o cerrojo, desechar la llave, levantar la aldaba o desencajar cualquier otra pieza o instrumento semejante con que se cierra algo. (15)

cerrar_19: Asegurar con cerradura, pasador, pestillo, tranca u otro instrumento, una puerta, ventana, tapa, etc., para impedir que se abra. (14)

crecer_1: Incrementar la cantidad o la importancia de algo, desarrollarse. (116)

dormir_1: Permanecer en un estado en el cual todos los movimientos voluntarios son suspendidos, generalmente para descansar. (18)

escuchar_1: Poner atención a lo que se oye. (107)

estar_14: Encontrarse alguien o algo en un estado determinado.(101)

explicar_1: Aclarar algo, dar información sobre un asunto. (106)

gestionar_1: Realizar un trámite para la consecución de una cuestión. (36)

gustar_1: Encontrar atractivo o agradable alguna cosa o a alguien.(117)

montar_2: Subirse alguien en un animal o un vehículo. (26)

morir_1: Fallecer, dejar de existir algo o alguien. (115)

parecer_1: Aparentar algo, sin serlo necesariamente.(51)

pensar_2: Usar la mente alguien para examinar una idea, razonar. (25)

perseguir_1: Ir detrás de alguien o algo para alcanzarle. (53)

trabajar_1: Emplearse en cualquier ejercicio, obra, trabajo o ministerio.(80)

valorar_2: Admitir la importancia de un hecho, cosa o acción. (70)

valer_1: Tener algo un determinado valor. (45)

ver_1: Recibir una imagen a través de la vista. (86)

viajar_1: Ir de un lugar a otro que suele estar distante, generalmente mediante algún medio de

B Clases verbales (gold standard y clase más similar)

Tabla 5: Enlace promedio

Gold standard	Clasificación más similar
1: estar_14 parecer_1 2: escuchar_1 explicar_1 gestionar_1 pensar_2 perseguir_1 valorar_2 ver_1 viajar_1 3: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 4: trabajar_1 volver_1 5: gustar_1 6: valer_1	1: estar_14 parecer_1 2: valer_1 3: montar_2 volver_1 4: abrir_18 cerrar_19 morir_1 5: crecer_1 dormir_1 gustar_1 6: escuchar_1 explicar_1 gestionar_1 pensar_2 perseguir_1 trabajar_1 valorar_2 ver_1 viajar_1

Tabla 6: Enlace completo

Gold standard	Clasificación más similar
1: estar_14 parecer_1 2: trabajar_1 volver_1 3: gustar_1 4: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 5: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 pensar_2 viajar_1 6: valer_1	1: abrir_18 cerrar_19 2: estar_14 parecer_1 3: crecer_1 dormir_1 morir_1 trabajar_1 viajar_1 volver_1 4: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 5: valer_1 6: pensar_2 7: montar_2 8: gustar_1

Tabla 7: Enlace simple

Gold standard	Clasificación más similar
1: estar_14 parecer_1 2: trabajar_1 volver_1 3: abrir_18 cerrar_19 crecer_1 dormir_1 escuchar_1 explicar_1 gestionar_1 morir_1 pensar_2 perseguir_1 valorar_2 ver_1 viajar_1 4: montar_2 5: gustar_1 6: valer_1	1: estar_14 parecer_1 2: abrir_18 cerrar_19 crecer_1 dormir_1 escuchar_1 explicar_1 gestionar_1 morir_1 perseguir_1 trabajar_1 valorar_2 ver_1 viajar_1 volver_1 3: montar_2 4: valer_1 5: pensar_2 6: gustar_1

Tabla 8: Enlace promedio ponderado

Gold standard	Clasificación más similar
1: estar_14 parecer_1 2: escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 3: abrir_18 cerrar_19 crecer_1 dormir_1 montar_2 morir_1 pensar_2 viajar_1 4: trabajar_1 volver_1 5: gustar_1 6: valer_1	1: estar_14 parecer_1 2: valer_1 3: crecer_1 dormir_1 morir_1 trabajar_1 viajar_1 volver_1 4: abrir_18 cerrar_19 escuchar_1 explicar_1 gestionar_1 perseguir_1 valorar_2 ver_1 5: montar_2 6: gustar_1 7: pensar_2

Bibliografía

- Barreto, Violeta Demonte y Ignacio Bosque. 1999. *Gramática descriptiva de la lengua española*. Espasa Calpe.
- Bonial, Claire, William Corvey, Martha Palmer, Volha V Petukhova, y Harry Bunt. 2011. A hierarchical unification of lyrics and verbnet semantic roles. En *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, páginas 483–489. IEEE.
- Brew, Chris y Sabine Schulte im Walde. 2002. Spectral clustering for german verbs. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, páginas 117–124. Association for Computational Linguistics.
- Brown, Susan Windisch, Dmitriy Dligach, y Martha Palmer. 2014. Verbnet class assignment as a wsd task. En *Computing Meaning*. Springer, páginas 203–216.
- Cifuentes Honrubia, JL. 2006. Alternancias verbales en español. *Revista Portuguesa de Humanidades*, 10:107–132.
- Dice, Lee R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.
- Falk, Ingrid, Claire Gardent, y Jean-Charles Lamière. 2012. Classifying french verbs using french and english lexical resources. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, páginas 854–863. Association for Computational Linguistics.
- Fernández-Montraveta, Ana y Gloria Vázquez. 2014. The sense corpus: an annotated corpus for spanish and catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2):273–288.
- Ferrer, Eva Esteve. 2004. Towards a semantic classification of spanish verbs based on subcategorisation information. En *Proceedings of the ACL 2004 workshop on Student research*, página 13. Association for Computational Linguistics.
- Giuglea, Ana-Maria y Alessandro Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. En *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, páginas 929–936. Association for Computational Linguistics.
- Goldberg, A. 1994. *Constructions, A Construction Grammar Approach to Argument Structure*. Chicago University Press, Chicago, IL.
- Gonzalez-Agirre, Aitor y German Rigau. 2013. Construcción de una base de conocimiento léxico multilíngüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.
- Joanis, Eric, Suzanne Stevenson, y David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.
- Korhonen, Anna, Yuval Krymolowski, y Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. En *Proceedings of LREC*, volumen 6.
- Korhonen, Anna, Yuval Krymolowski, y Zvika Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. En *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, páginas 64–71. Association for Computational Linguistics.
- Lenci, Alessandro. 2014. Carving verb classes from corpora. *Word Classes: Nature, typology and representations*, 332:17.
- Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.
- Levin, Beth y Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*, volumen 26. MIT press.
- Li, Jianguo y Chris Brew. 2008. Which are the best features for automatic verb classification. En *ACL*, páginas 434–442.
- Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze, y others. 2008. *Introduction to information retrieval*, volumen 1. Cambridge university press Cambridge.
- Merlo, Paola y Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.
- Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Niles, Ian y Adam Pease. 2003. Mapping wordnet to the sumo ontology. En *Proceedings of the iee international knowledge engineering conference*, páginas 23–26.

- Scarton, Carolina, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer, y Anna Korhonen. 2014. Verb clustering for brazilian portuguese. En *Computational Linguistics and Intelligent Text Processing*. Springer, páginas 25–39.
- Schuler, Karin Kipper. 2005. Verbnets: A broad-coverage, comprehensive verb lexicon.
- Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.
- Strehl, Alexander. 2002. Relationship-based clustering and cluster ensembles for high-dimensional data mining.
- Sun, Lin y Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, páginas 638–647. Association for Computational Linguistics.
- Sun, Lin, Diana McCarthy, y Anna Korhonen. 2013. Diathesis alternation approximation for verb clustering. En *ACL (2)*, páginas 736–741.
- Swift, Mary. 2005. Towards automatic verb acquisition from verbnets for spoken dialog processing. En *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, páginas 115–120.
- Vázquez, Gloria, Ana Fernández, y M. Antònia Martí. 2000. Clasificación verbal. *Alternancias de diátesis. Quaderns de Sintagma*, 3.
- Vlachos, Andreas, Anna Korhonen, y Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. En *Proceedings of the workshop on geometrical models of natural language semantics*, páginas 74–82. Association for Computational Linguistics.