HACIA UNA DEFINICIÓN DE LA SIMILITUD VERBAL PARA LA EXTRACCIÓN DE EVENTOS.

LARA GIL-VALLEJO
Universitat Oberta de Catalunya
IRENE CASTELLÓN
Universitat de Barcelona
MARTA COLL-FLORIT
Universitat Oberta de Catalunya

RESUMEN

La extracción de eventos consiste en la obtención de conocimiento factual a partir de textos, lo que requiere delimitar el evento ofreciendo una interpretación precisa de las entidades y relaciones presentes en él. Abordamos esta tarea caracterizando estos dos elementos a partir de la información morfológica, sintáctica y semántica de los constituyentes que los realizan. Para ello introducimos la aplicación de kybots, heurísticas que permiten explorar el contexto verbal y determinar las relaciones semánticas entre los elementos de la oración. Con el objetivo de simplificar su desarrollo, realizamos un experimento consistente en aplicar kybots diseñados para verbos concretos, que tomamos como modelos, a verbos similares en cuanto a su comportamiento sintáctico-semántico. Los resultados muestran que esta estrategia es viable, aunque es necesario seguir trabajando en la determinación de parámetros para establecer los grupos de verbos similares.

Palabras clave: extracción de eventos, similitud verbal, kybot.

ABSTRACT

Event extraction consists in obtaining factual knowledge from texts and requires an accurate interpretation of the contribution of linguistic constituents that are used to define both the participants in the event and the

relationships established between them. In this work we use kybots (extraction profiles) to carry out this task. Kybots are heuristic tools that explore the verbal context and determine the semantic relationships established among the elements in the sentence. To simplify its development we investigate through an experiment the application of kybots developed for specific verbs to similar verbs. Results show that this strategy is suitable, but it is necessary to refine the parameters used to define groups of similar verbs.

Keywords: event extraction, verb similarity, kybot.

1. INTRODUCCIÓN[1]

La extracción de eventos se enmarca dentro del ámbito más general de la extracción de información y se define como la identificación de eventos en texto, generalmente caracterizados mediante verbos o sustantivos deverbales junto con sus participantes.

Uno de los principales retos es la gran variabilidad lingüística presente en la expresión de un evento (Vossen et al., 2013). Veamos este ejemplo tomado de un diario digital:

"*Este informe claramente muestra que la crisis de 2008 condujo a una significativa aceleración del declive industrial europeo (...)*"[2]

En esta oración se describe un evento principal y sus correspondientes participantes y circunstancias. Sin embargo, para expresar este mismo evento se podrían utilizar múltiples variantes léxicas (*aumento, caída, recesión,* etc.) y sintácticas (*el declive industrial se acentuó con la crisis, la crisis aceleró la caída de la industria europea*, etc). La vía que seguimos para abordar esta complejidad explora la explotación de las estructuras argumentales (Surdeanu, 2003).

En este trabajo presentamos la aplicación de kybots a la extracción de eventos. Los kybots son perfiles de extracción de información desarrollados en el marco del proyecto Kyoto[3] (Vossen et al., 2008), que permiten identificar y etiquetar con roles semánticos aquellos constituyentes que poseen información relevante para el evento. Para una aplicación de kybots al campo de los textos biomédicos se puede consultar Casillas et al. (2011). Detallamos dos experimentos preliminares que consisten en el desarrollo de kybots para verbos específicos (verbos modelo) y su aplicación a verbos que muestran un comportamiento sintáctico-semántico similar. Para ello se ha realizado una primera aproximación a la definición de similitud verbal teniendo en cuenta la estructura argumental de las oraciones que expresan los eventos.

2. KYBOTS: APLICACIÓN Y EXTENSIÓN

Los kybots son reglas heurísticas que exploran documentos enriquecidos con información lingüística en formato KAF (Agirre et al. 2009). Cada uno de estos perfiles representa un patrón de extracción de información relevante para un cierto evento.

El formato KAF (Kyoto Annotation Framework) es un formato de anotación no supervisada (Bosma et al., 2009) que emplea Freeling 3.1. (Padró et al. 2012) y comprende el texto segmentado y tokenizado, y los correspondientes lemas, categorías morfológicas junto con un análisis sintáctico de dependencias de los constituyentes (LLoberes et al. 2010). Además permite agregar información de referencias externas: bases de conocimiento y ontologías, como Spanish Wordnet 3.0[4] (Fernández-Montraveta et al., 2008; Oliver y Climent 2012) y la ontología DOLCE Lite Plus[5] (Masolo et al., 2003).

```
<!-- abrir_suj_objeto -->
<!-- S-V-O -->

<Kybot id="ABR4">
<variables>
  <var name="S" type="term" synfunc="subj"/>
  <var name="V" type="term" pos="V*" lemma="abrir"/>
  <var name="O" type="term" synfunc="dobj" pos="N*"/>
  <var name="V2" type="term" pos="V*"/>
</variables>
<relations>
  <root span="V"/>
  <rel span="S" pivot="V" direction="preceding"/>
  <rel span="O" pivot="V" direction="following" notInBetween="V2"/>
</relations>
<events>
  <event target="$V/@tid" lemma="$V/@lemma" pos="$V/@pos"/>
  <role target="$S/@tid" rtype="ACTOR" lemma="$S/@lemma" pos="$S/@pos"/>
  <role target="$O/@tid" rtype="UNDERGOER" lemma="$O/@lemma" pos="$O/@pos"/>
</events>
```

Figura 1. Ejemplo de kybot

En la figura 1 se muestra la estructura de un kybot, que consta de tres partes: variables, relaciones y eventos. Las variables expresan las restricciones de tipo morfosintáctico, semántico o léxico sobre la naturaleza de los constituyentes del evento. Las relaciones especifican el orden lineal de los constituyentes. En la parte de los eventos se declara el tipo de evento y el rol de los participantes. Los roles semánticos que hemos utilizado para etiquetar los constituyentes se han tomado de la propuesta de Bonial et al. (2011), ya que están estructurados de forma jerárquica, lo que permite flexibilidad a la hora de definir la granularidad semántica con la que se quiere trabajar y además constituyen una base de estandarización de roles semánticos.

El desarrollo de estos perfiles es una tarea costosa, ya que requiere una interpretación precisa de la contribución sintáctico-semántica de los elementos de una lengua. La estrategia que exploramos en nuestra propuesta opta por la extensión de kybots elaborados para verbos específicos a conjuntos de verbos similares argumentalmente. Para ello es necesario determinar los parámetros que permiten definir el grado de similitud sintáctico-semántica de las unidades verbales de manera óptima para la extracción de eventos.

3. METODOLOGÍA DEL EXPERIMENTO

Se han realizado dos experimentos preliminares que nos informan sobre el comportamiento de los kybots y aportan información acerca de aspectos no previstos y posibles parámetros para establecer una definición de similitud adecuada para la tarea.

3.1 Experimento 1

Se seleccionaron inicialmente cuatro verbos modelo (*abrir, aparecer, elaborar* y *crecer*) que muestran heterogeneidad en cuanto al número de sentidos verbales y estructura argumental y que están presentes en el corpus Sensem[6] (Alonso et al., 2007).

Posteriormente se diseñaron un total de 83 kybots utilizando las oraciones correspondientes a cada verbo como corpus de desarrollo. Finalmente se evaluaron estos kybots en otras 30 oraciones correspondientes a los mismos verbos tomadas del Corpus del Español Actual[7] (Subirats et al., 2012) en dos formatos: uno con las oraciones reales tomadas del corpus y otro con oraciones simplificadas en las que sólo aparecen los constituyentes directamente dependientes del verbo. Los resultados de la evaluación se muestran en la tabla 1. En la tabla 2 se muestra la mejora en las oraciones simplificadas. Los valores con símbolos negativos significan que estos son mayores en las oraciones reales.

	Oraciones Simplificadas			Oraciones Reales		
	Prec.	Rec.	F1	Prec.	Rec.	F1

Abrir	0,80	0,80	0,80	0,87	0,69	0,77
Aparecer	0,89	0,80	0,84	0,77	0,55	0,64
Crecer	0,93	0,84	0,88	0,90	0,68	0,77
Elaborar	0,74	0,80	0,77	0,75	0,58	0,65
Global	0,84	0,81	0,82	0,82	0,63	0,71

Tabla 1. Resultados de la evaluación del experimento 1

	Dif. Cob.	Dif. Prec.	Dif. F1
Abrir	-0.07	0,11	0,03
Aparecer	0,12	0,25	0,20
Crecer	0,03	0,16	0,11
Elaborar	-0,01	0,22	0,12

Tabla 2. Diferencia entre las oraciones simplificadas y las reales del experimento 1

3.2 Experimento 2

En la segunda fase se aplicaron los kybots desarrollados para los verbos modelo a grupos de verbos similares. Para construir estos grupos se procedió a generar patrones sintáctico-semánticos para todos los verbos presentes en el corpus Sensem. Los patrones comprenden información relativa a la subcategorización del verbo, al orden de sus argumentos y al rol semántico de los mismos.

Ejemplo para *crecer*:

- pre-SN-Que (SN preverbal objeto semántico): *Los precios crecieron*.
- pre-SN-Que post-SP-Circ (SN preverbal objeto semántico + SP postverbal circunstancial): *Rocío Jurado se creció ante la adversidad*.

Respecto a las etiquetas utilizadas, *pre* y *post* remiten a la posición del argumento con respecto al verbo (preverbal o postverbal). También se especifica el tipo de constituyente: *SN* (sintagma nominal), *SP* (sintagma preposicional), *pronombre*, etc. En cuanto al contenido semántico, se expresa con las etiquetas *Que*, *Quien*, *Circunstancial*, *Destinatario*, *Donde*, *Como* y *Cuando*. Estas etiquetas representan una compactación realizada sobre los roles empleados en el corpus Sensem para evitar que los patrones quedaran dispersos y además coherente con los roles empleados en los kybots.

Para cada verbo modelo *(abrir, aparecer, crecer* y *elaborar*) se listaron todos los patrones existentes en el corpus. Se calculó la similitud de cada verbo de Sensem con los verbos modelo teniendo en cuenta un criterio simple que pudiera servir de *baseline* para posteriores experimentos: el porcentaje de patrones del verbo de Sensem que son compartidos con el verbo modelo. Presentamos una muestra en la tabla 3. Las ocurrencias de patrones comunes están resaltadas.

	post-SN-Que	post-SN-Que pre-SN-Quien	post-O-Que	post-SP-Que	Total	Sim.
abrir	33	22	0	1	56	1
acabar	**15**	0	1	0	16	0,94
acceder	0	0	1	**17**	18	0,94
aceptar	**26**	**33**	10	0	69	0,86

Tabla 3. Ejemplo para el cálculo de similitud verbal.

La tabla 4 muestra los diez verbos más similares a cada uno de los verbos modelo teniendo en cuenta este criterio.

elaborar		crecer		aparecer		abrir	
efectuar	0.97	morir	0.91	desaparecer	0.96	renovar	0.97

celebrar	0.95	poseer	0.90	dormir	0.95	celebrar	0.94
realizar	0.93	contener	0.86	existir	0.95	realizar	0.92
gestionar	0.91	constituir	0.83	proceder	0.87	suspender	0.92
crear	0.91	tener	0.81	morir	0.81	crear	0.92
adquirir	0.90	consistir	0.79	partir	0.75	cerrar	0.91
practicar	0.89	formar	0.76	desarrollar	0.71	efectuar	0.90
producir	0.89	desaparecer	0.74	residir	0.70	morir	0.90
provocar	0.88	incluir	0.74	producir	0.70	controlar	0.90
renovar	0.86	producir	0.74	venir	0.69	producir	0.89

Tabla 4. Los 10 verbos más similares para cada verbo modelo.

De nuevo se tomaron para cada verbo similar las oraciones correspondientes del corpus Sensem en dos formatos: las oraciones reales y una versión simplificada. Se procedió a la evaluación de los kybots, cuyos resultados se muestran en la tabla 5, donde aparecen los valores medios para los diez verbos dentro de cada grupo de verbos similares. Entre paréntesis se especifica la desviación estándar.

	Oraciones simplificadas	Oraciones reales

modelo	Cob.	Prec.	F1	Cob.	Prec.	F1
abrir	0,59 (0,16)	0,85 (0,07)	0,68 (0,13)	0,63 (0,13)	0,60 (0,11)	0,62 (0,11)
aparecer	0,45 (0,13)	0,59 (0,29)	0,51 (0,17)	0,34 (0,15)	0,43 (0,20)	0,37 (0,17)
crecer	0,82 (0,09)	0,87 (0,06)	0,84 (0,06)	0,78 (0,09)	0,78 (0,10)	0,77 (0,08)
elaborar	0,61 (0,09)	0,86 (0.06)	0,70 (0,06)	0,67 (0,07)	0,69 (0,05)	0,68 (0,05)
global	0,62	0,79	0,68	0,60	0,63	0,61

Tabla 5. Resultados de la evaluación del experimento 2.

En la tabla 6 aparece la mejora de las oraciones simplificadas respecto de las reales. (Los números con el signo negativo significan que las oraciones reales obtienen mejores resultados)

	Dif. Cob.	Dif. Prec.	Dif. F1
Abrir	-0,05	0,24	0,06
Aparecer	0,11	0,16	0,13
Crecer	0,04	0,09	0,07
Elaborar	-0,06	0,16	0,03

Tabla 6. Diferencia entre las oraciones simplificadas y las reales del experimento 2.

En la tabla 7 mostramos la diferencia entre los resultados del experimento 1 y el experimento 2, con los valores en positivo cuando son más altos en el experimento 1, y en negativo cuando lo son en el experimento 2.

	Oraciones simplificadas Modelo inic. – modelo grupos			Oraciones reales Modelo inic. – modelo grupos		
	Cob.	Prec.	F1	Cob.	Prec.	F1
Abrir	0,21	-0,05	0,12	0,24	0,09	0,15
Aparecer	0,44	0,21	0,33	0,43	0,12	0,27
Crecer	0,11	-0,03	0,10	0,12	-0,10	0,00
Elaborar	0,13	-0,06	0,07	0,08	-0,11	-0,03

Tabla 7. Diferencia entre los resultados del experimento 1 y el experimento 2.

4. ANÁLISIS

4.1. Resultados generales de la evaluación de los kybots

En general vemos como los resultados para oraciones reales oscilan en torno al 0,70 de F1 para los verbos modelo y 0,60 para los verbos similares. Estos resultados son prometedores si se tiene en cuenta el hecho de que no se utilizaron todos los recursos posibles en el desarrollo de los kybots, en especial los semánticos. Por otro lado, los criterios de similitud verbal empleados representan un baseline y permiten una mayor elaboración teniendo en cuenta las conclusiones obtenidas a partir del análisis de resultados que realizamos en los siguientes apartados.

Respecto al experimento 2, en la tabla 5 puede observarse que existe una diferencia notable entre los resultados de los diferentes modelos, con algunos con buenos resultados, como *crecer* y otros que presentan problemas como *aparecer*. Además la alta desviación estándar de la media de cobertura y precisión que presentan los resultados de los verbos pertenecientes a los modelos de *abrir* y *aparecer* nos informa de un comportamiento heterogéneo: al aplicar

los kybots a verbos del mismo modelo obtenemos resultados dispares. Esto apunta a una necesidad de redefinición de la similitud verbal para crear clases más homogéneas explorando sus dos vertientes: los atributos empleados para definirla y la medida de similitud empleada.

4.2. Comparación de resultados de oraciones simplificadas y oraciones reales

Las diferencias en F1, precisión y cobertura entre oraciones reales y simplificadas por verbo son las que podemos observar en las tablas 2 (experimento 1) y 6 (experimento 2). En el caso del experimento 1, los resultados para la precisión de etiquetado son claramente mejores en las oraciones simplificadas, con una mejora entre 11 y 25 puntos. Este es el resultado esperado, ya que en el proceso de simplificación se eliminan elementos no argumentales que podrían ser etiquetados erróneamente. En cambio, en cuanto a la cobertura, hay dos casos en los que las oraciones reales obtienen mejores resultados que las simplificadas: para *abrir* y *elaborar* la cobertura mejora en 7 y 1 puntos respectivamente. Este comportamiento podría atribuirse a los kybots desarrollados para estos dos verbos, que presentan restricciones que no se cumplen sin el contexto existente las oraciones reales, aunque es necesario profundizar en las causas de este hecho. Estos resultados se repiten en el caso del experimento 2: la precisión disminuye de forma general en el caso de las oraciones reales y la cobertura aumenta para las oraciones reales en el caso de *abrir* y *elaborar*.

4.3. Comparación de modelos

En la tabla 7 podemos ver la comparativa entre los resultados del experimento 1 y 2. Las cifras positivas indican cuanto mejoran la precisión, cobertura y F1 si empleamos kybots desarrollados para verbos específicos en esos mismos verbos. La cobertura es mejor en este caso, hasta 44 puntos para las oraciones reales de los verbos similares a *aparecer*. Por lo tanto, uno de los principales retos a la hora de aplicar kybots a verbos similares a los modelos para los que se

desarrollaron es mejorar la cobertura de los mismos. En cambio, en cuanto a la precisión, esta es mejor para los grupos de verbos similares en los casos de verbos del grupo *crecer* y *elaborar*, mejorando de 3 a 11 puntos, así como para *abrir* en el caso de las oraciones simplificadas, mejorando en 5 puntos.

5. CONCLUSIONES Y TAREAS FUTURAS

Los resultados de la evaluación apuntan a que se trata de un camino válido, aunque es necesario refinar varios aspectos: la elección de los verbos modelo, la definición de similitud verbal y la elaboración de los propios kybots.

La elección de los verbos modelo es crucial y ha de basarse en criterios objetivos. Las tareas futuras incluyen el uso de clusters automáticos para determinar las clases verbales y los predicados centrales de éstas (modelos).

Por otro lado, aunque podemos valorar positivamente la contribución de los parámetros escogidos en esta primera definición de similitud verbal, se pone de manifiesto la necesidad de analizar la interacción de estos parámetros con otros elementos que no se tuvieron en cuenta, como la presencia y frecuencia de los adjuntos, los diferentes sentidos que tiene un verbo y el cálculo de similitud basado en patrones. Respecto a este último punto, creemos que es importante utilizar una medida que sea sensible a las diferencias en las frecuencias de los patrones dentro de cada verbo, y que sea capaz de agrupar verbos que compartan esas variaciones.

Para definir las clases verbales de forma óptima para la tarea de extracción de eventos exploraremos propiedades de diferente naturaleza, como las ocurrencias de constituyentes de los patrones de manera aislada y su solapamiento, los roles semánticos que se distribuyen, las preferencias selectivas y los sentidos verbales que se distinguen en una forma verbal. Pensamos que ello nos ayudará a formar clases más homogéneas que posibilitaran una mejora en la etiquetación de participantes en eventos.

Finalmente exploraremos la generación semi-automática o automática de kybots, de tal manera que nos permita aprovechar al máximo la información presente en el corpus.

NOTAS

[1] Esta investigación se ha llevado a cabo dentro del proyecto SKATER: Adquisición de escenarios de conocimiento a través de la lectura de textos: Lingüística y cognición (SKATER) Financiado por el Ministerio de Economía y competitividad TIN2012-38584-C06-06.
[2] www.europapress.es [17/2/2014]
[3] http://kyotoproject.eu/xmlgroup.iit.cnr.it/kyoto/index.html
[4] Disponible en el *Multilingual central repository* :
(http://adimen.si.ehu.es/web/MCR)
[5] http://www.loa.istc.cnr.it/old/DOLCE.html
[6] http://grial.uab.es/sensem/corpus
[7] http://sfn.uab.es:8080/SFN/tools/cea/english

REFERENCIAS BIBLIOGRÁFICAS

Agirre E., Artola X., Díaz de Ilarraza A., Rigau G., Soroa A. y Bosma W. 2009. *Kaf: Kyoto annotation framework. Informe técnico TR 1-2009*, Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad del País Vasco.

Alonso L., Capilla J. A., Castellón I., Fernández-Montraveta A., & Vázquez G. 2007. "The sensem project: Syntactico-semantic annotation of sentences in Spanish". En N.Nikolov, K. Bontcheva, G.Angelova and R. Mitkov. (eds.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005. Current Issues in Linguistic Theory*, 292 Amsterdam: John Benjamins Publishing, pp. 89 – 98.

Bonial C., Corvey W., Palmer M., Petukhova V. y Bunt H. 2011. "A Hierarchical Unification of LIRICS and VerbNet Semantic Roles". En *Proceedings IEEE-ICSC 2011 Workshop on*

Semantic Annotation for Computational Linguistic Resources. Stanford, CA.

Bosma W., Vossen P., Soroa A., Rigau G., Tesconi M., Marchetti A., & Aliprandi C. 2009. "KAF: a generic semantic annotation format". En *Proceedings of the GL2009 Workshop on Semantic Annotation*. Pisa, Italia.

Casillas A., Díaz de Ilarraza A., Gojenola K., Oronoz Maite., Rigau G. 2011. "Using Kybots for extracting events in biomedical texts". En *Proceedings of the BioNLP Shared Task 2011 Workshop (BioNLP Shared Task '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 138-142.

Fernández-Montraveta A., Vázquez G., & Fellbaum C. 2008. *The Spanish Version of WordNet 3.0. Text Resources and Lexical Knowledge*. Mouton de Gruyter, 175-182.

Lloberes, M., I. Castellón, L. Padró (2010). "Spanish FreeLing Dependency Grammar"". En Nicoletta Calzolari et al. (ed.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 693-699.

Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A. 2003. Wonderweb deliverable D18 ontology library (final), Diciembre.

Oliver A. y S. Climent (2012). "Using Wikipedia to develop language resources: WordNet 3.0 in Catalan and Spanish", *Digithum*, 14.

Padró Ll., Stanilovsky E. 2012. "FreeLing 3.0: Towards Wider Multilinguality", *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. Estanbul, Turquía, 2012.

Reese S., Boleda G., Cuadros M., Padró Ll., Rigau G. 2010. "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". En Proceedings of 7th Language Resources and Evaluation Conference. Valletta, Malta.

Subirats C. y Ortega M. 2012. Corpus del Español Actual (<http://sfncorpora.uab.es/CQPweb/cea/>)

Surdeanu M., Harabagiu S., Williams J., Aarseth P. 2003. "Using predicate-argument structures for information extraction." En *Proceedings of the 41st Annual Meeting on Association for*

Computational Linguistics, Volumen 1. Association for Computational Linguistics, pp. 8-15.

Vossen P., Agirre E., Calzolari N., Fellbaum C., Hsieh S.K., Huang C.R., VanGent J. 2008. "KYOTO: a System for Mining, Structuring and Distributing Knowledge across Languages and Cultures". En *Proceedings of the 6th Language Resources and Evaluation Conference* (Marrakesh).

Vossen, P., Agirre E., Rigau G., Soroa A. 2013. "KYOTO: A Knowledge-Rich Approach to the Interoperable Mining of Events from Text." *New Trends of Research in Ontologies and Lexical Resources* pp. 65-90.