Clasificación de roles semánticos usando características sintácticas, semánticas y contextuales

José A. Reyes[1], Azucena Montes[1,2], Juan G. González[1] y David E. Pinto[3]

[1]Centro Nacional de Investigación y Desarrollo Tecnológico, México

[2]Universidad Nacional Autónoma de México, México

[3]Benemérita Universidad Autónoma de Puebla, México

{alexreyes06c, amr, gabriel}@cenidet.edu.mx, amontesr@iingen.unam.mx, dpinto@cs.buap.mx

Resumen. Este artículo presenta una clasificación de roles semánticos basada en características sintácticas, semánticas y contextuales. El objetivo de este artículo es identificar mediante la tarea de clasificación, el tipo de rol semántico existente entre un evento y sus actantes; por ello se presenta un análisis de características para seleccionar un subconjunto que mejore el desempeño de la tarea. Adicionalmente, se presenta una comparativa de cuatro algoritmos de clasificación: máquinas de soporte vectorial, los *k*-vecinos más cercanos, clasificador de *Bayes* y el clasificador basado en arboles de decisión C4.5, esto con la finalidad de analizar su desempeño con todas las características y con las relevantes en cada categoría de rol semántico. Con base en la experimentación, se obtiene que la selección de atributos mejora el desempeño de la tarea de clasificación, ya que con el grupo de características relevantes, se obtiene el mejor desempeño de 84.6% con el algoritmo basado en arboles de decisión C4.5. El resultado del etiquetado de roles puede ser utilizado para una representación de conocimiento o se puede utilizar para apoyar en la tarea de aprendizaje ontológico.

Palabras clave. Clasificación de roles semánticos, adquisición de conocimiento, procesamiento del lenguaje natural, aprendizaje máquina.

Classifying Case Relations using Syntactic, Semantic and Contextual Features

Abstract. This paper presents a classification of semantic roles using syntactic, semantic and contextual features. The aim of our work is to identify types of semantic roles involving events and their actors; therefore, we fulfill a feature analysis in order to select the best feature subset which improves the fulfillment of the task. In addition, we compare four classification algorithms: Support Vector Machine (SVM), k-nearest neighbor (k-NN), Bayes classifier and decision tree classifier C4.5. This comparison was made in order to analyze the performance of these algorithms with all features against relevant features for each semantic role category. In our experimentation, we obtain that feature selection improved the performance of algorithms in our classification task, since with relevant features we obtained the best performance of 84.6% with decision tree classifier C4.5. The results for the labeling task can be used for knowledge representation or ontology learning.

Keywords. Semantic roles classification, knowledge acquisition, natural language processing, machine learning.

1 Introducción

La identificación automática de componentes en los textos es una tarea que involucra el área de Procesamiento del Lenguaje Natural (PLN). Actualmente todos los esfuerzos se dirigen en resolver aspecto semánticos presentes en las diferentes lenguas. Con el crecimiento de la información en medios electrónicos se requieren herramientas de búsqueda más precisas, que puedan responder a preguntas como: ¿qué

ocurrió? ¿quién lo hizo? ¿cómo lo hizo? ¿dónde ocurrió?, ¿cuándo pasó?, entre otras. Los eventos, como acaecimientos, están presentes en la mayoría de los documentos y de manera más explícita en documentos históricos y periodísticos. Los eventos están relacionados con entidades que juegan un rol semántico en las oraciones y que son fundamentales para la comprensión de los textos.

En el presente trabajo estamos interesados en determinar qué rol semántico lleva a cabo cada entidad relacionada a un evento en documentos periodísticos en español. Para tal fin, nos situamos en la teoría de L. Tesnière [1] y M. A. K. Halliday [2] quienes consideran que los elementos fundamentales de la oración son: los actores (actantes), la acción (verbo) y el decorado (circunstantes). El verbo es considerado como el centro de toda oración y sobre él giran los demás elementos. Los actantes juegan un tipo de rol semántico y son clasificados [1] como: a) el AGENTE es el primer argumento de una oración y es el encargado de realizar la acción. Éste tiene el rasgo de ser animado y adicionalmente puede corresponder con el sujeto de la oración; b) el OBJETO o segundo argumento es el elemento que complementa el significado de la oración, adicionalmente puede corresponder con el paciente; c) el BENEFICIARIO o tercer argumento es quien recibe los beneficios o perjuicios de la acción; éste puede identificarse con el complemento indirecto de una oración. Existe una tercera entidad, el circunstante, que no es considerada como un actor, sin embargo, se define como elemento opcional que amplía el significado de una oración y que su ausencia no afecta su significado. El elemento circunstante puede estar determinado por el instrumento u objeto con que se produce la acción, la fuerza de la acción, el tiempo o el aspecto locativo de un evento.

Así, considerando la clasificación de los actantes de [1] y la posible presencia del circunstante, nosotros consideramos siguientes categorías de roles semánticos para este artículo:

- EVENTO:AGENTE (EA). El evento es iniciado o arrancado por un actante que se le conoce como sujeto.
- EVENTO:OBJETO (EO). El evento tiene un complemento que perfecciona el significado de la acción.
- EVENTO:BENEFICIARIO (EB). El beneficiario corresponde con el complemento indirecto y es quien recibe los efectos de la acción.
- EVENTO:CIRCUNSTANTE (EC). Los eventos pueden tener información adicional que extiende su significado, en este caso se consideran únicamente como circunstantes, la fuerza que indica el génesis del evento y el instrumento con qué se produce la acción.
- EVENTO:OTRO (EX). Los eventos tienes información relacionada a ellos que de alguna manera no caen bajo alguna de las cuatro categorías principales. En este caso se considera aspectos temporales, locativos, causales, entre otros.

Tabla 1. Ejemplos de las categorías de roles

La Junta General Ejecutiva del IFE enviará paquete electoral a mexicanos en el extranjero mediante mensajería el próximo 17 de junio.

Tipo de rol	Evento	Tipo de entidad
Evento: Agente	enviará	<agente>La Junta General Ejecutiva del IFE </agente>
Evento: Objeto	enviará	<objeto> paquete electoral </objeto>
Evento: Beneficiario	enviará	<beneficiario>mexicanos en el extranjero</beneficiario>
Evento: Circunstante	enviará	<circunstante>mensajería </circunstante>
Evento: Otro	enviará	<otro:temp>el próximo 17 de junio</otro:temp>

En la Tabla 1 se muestran los ejemplos de las cinco categorías de roles que se exponen en este artículo.

El objetivo de este artículo es realizar una clasificación de las entidades que rodean a los eventos en una de las cinco categorías de roles semánticos. Para ello se hace uso de características sintácticas, semánticas y contextuales. Se utilizan cuatro algoritmos de clasificación (K-NN, SVM, NaiveBayes y C4.5)

con la finalidad de mostrar su desempeño y comparar el resultado en un conjunto de pares de pruebas.

El resto del artículo se organiza de la siguiente manera: en la sección 2 se presentan los trabajos relacionados con la tarea de clasificación de relaciones y roles semánticos; en la sección 3 se describe el corpus de noticias utilizado, la descripción de las características y su extracción; la sección 4 expone el proceso de selección de características para obtener un subconjunto relevante que mejora la eficiencia de la tarea de clasificación; en la sección 5 se describe la tarea de clasificación y los algoritmos analizados para dicha tarea; en la sección 6 se detalla los experimentos realizados con el conjunto total de características y con las características relevantes, además, se analiza el desempeño de cada algoritmo de clasificación para las cinco categorías de roles semánticos; finalmente, en la sección 7 se presentan las conclusiones y los trabajos futuros.

2 Trabajos relacionados

En el área de clasificación de relaciones semánticas entre sustantivos se han propuesto varios enfoques basados en características sintácticas y semánticas [3, 4, 5, 6, 7, 8]. Trabajos como [9, 10] consideran características contextuales como modelo de representación de las entidades. La clasificación de otras relaciones también ha sido abordada en trabajos como [11] donde clasifican de manera automática relaciones temporales entre eventos utilizando atributos como el tiempo de los verbos y aspectos gramaticales. En [12] se presenta una ontología para la clasificación de relaciones temporales y locativas entre eventos. Los trabajos mencionados anteriormente realizan la identificación de relaciones semánticas entre sustantivo; algunos extraen relaciones temporales o locativas para eventos en el idioma Inglés. En [13] se realiza la extracción de información temporal y espacial para eventos o sucesos a partir de documentos periodísticos en el idioma Español. Sin embargo, los trabajos mencionados no se preocupan por la identificación o clasificación de roles entre eventos y sus actantes.

Existen trabajos interesados en relaciones semánticas como [14] que clasifica, entre varios tipos de relaciones, el rol de una entidad utilizando un enfoque basado en máquinas de soporte vectorial con características léxicas y sintácticas; en [15] se presenta un sistema para la identificación automática de roles semánticos tales como agente, paciente, y roles específicos de un dominio como tema, ponente y mensaje; y [16] que presenta un enfoque basado en características sintácticas y morfológicas para la clasificación automática de 19 tipos de roles semánticos, entre los que destacan los argumentos de los verbos, aspectos temporales y locativos. Los trabajos [14, 15, 16] realizan una clasificación de roles, sin embargo estos se enfocan en el idioma Inglés y no consideran características semánticas ni contextuales para la identificación de roles semánticos.

La tarea de identificación de roles semánticos para el español desde un enfoque semántico-contextual involucra un reta en el área del Procesamiento del Lenguaje Natural. Por lo tanto, este artículo se basa en la clasificación de cinco tipos roles semánticos entre eventos y las entidades que los rodean para la lengua Española desde un enfoque semántico-contextual. Para ello, se consideran características sintácticas, semánticas y contextuales haciendo una selección de atributos relevantes y un análisis del impacto de dichas características en diferentes algoritmos de clasificación.

3 Datos y el conjunto de características

En este artículo, la clasificación de roles semánticos se basa en pares del tipo EVENTO:ENTIDAD y la tarea consiste en determinar qué tipo de rol semántico expresa la ENTIDAD.

Un corpus ha sido creado con sentencias de noticias que provienen de cinco periódicos mexicanos electrónicos (El universal, Excélsior, La jornada, Milenio y El occidental). Los textos de las noticias datan del 15 de marzo de 2011 al 15

<entidad>La Junta General Ejecutiva del IFE</entidad>
<evento>enviará</evento> paquete electoral a mexicanos en el extranjero.

Características= (1,0,1,3,3,1,1,1,1,0,1,6,0,0,4,0,5,4,1,0,0)

Fig. 1. Ejemplo de vector de características

de noviembre de 2011. Dicho corpus es utilizado para la extracción de las características de los pares.

Las características extraídas para cada par EVENTO:ENTIDAD se representan utilizando el modelo espacio vectorial [17], donde las características se representan numéricamente, estos valores se obtienen de las sentencias del corpus. Para cada par se propone extraer 21 características, las cuales se dividen en 11 sintácticas, 4 semánticas y 6 contextuales, éstas se describen a continuación y un ejemplo es mostrado en la Figura 1.

3.1 Características sintácticas

Las características sintácticas describen los pares en el aspecto de su estructura, posición e información morfológica con la ayuda de la herramienta de análisis de textos *FreeLing 2.1* [18].

Estas características se describen como sigue:

1. Posición de la entidad. La entidad puede estar a la izquierda o derecha con respecto al evento.
2. Distancia de la entidad. El número de palabras existentes entre la entidad y el evento.
3. Información morfológica del evento. El modo, tiempo, persona y número del evento.
4. Longitud del evento. La secuencia de *n* elementos de los *n*-gramas del evento.
5. Información morfológica del núcleo de la entidad. El número y grado de la entidad.
6. Entidad definida. Una entidad se considera definida si su artículo es definido.
7. Longitud de la entidad. La secuencia de *n* elementos de los *n-gramas* de la entidad.

3.2 Características semánticas

Las características semánticas describen los pares en el aspecto de su significado y sentido. Estas características son:

1. Entidad Nombrada. Las entidades nombradas son detectadas con el módulo de Reconocimiento de Entidades Nombradas de la herramienta FreeLing 2.1 [18].
2. Tipo de preposición. Si la entidad pertenece a una frase preposicional, se determina el tipo de preposición.
3. Tipo de entidad. Las entidades se categorizan con respecto a su núcleo nominal, nombre propio o nombre común.
4. Hiperónimo de la entidad. La relación de hiperonimia ayuda a determinar el tipo de entidad en función de su rasgo semántico, se obtienen los hiperónimos de la entidad en 3 niveles y se verifica si es un tipo de agente animado u objeto.

3.3 Características contextuales

Las características contextuales describen los pares considerando las palabras que ocurren en el contexto de la entidad, con un tamaño de ventana determinado (*n*=3). Las características contextuales se describen a continuación:

1. Contexto izquierdo. Las tres palabras a la izquierda representan su contexto izquierdo y se extrae su categoría gramatical (adjetivo, adverbio, determinante, sustantivo, verbo, pronombre, conjunción, preposición o signo de puntuación) para cada una.
2. Contexto derecho. Las tres palabras a la derecha representan su contexto derecho y se extrae su categoría gramatical (adjetivo, adverbio, determinante, sustantivo, verbo, pronombre, conjunción, preposición o signo de puntuación) para cada una.

La Tabla 2 muestra las 21 características que representan cada par EVENTO:ENTIDAD y los posibles valores nominales y numérico que éstas pueden tomar con base en la sentencia donde ocurren.

4 Selección de características

Además de experimentar con las 21 características descritas en la sección anterior, realizamos un proceso de selección de características para observar su impacto en la clasificación. El objetivo de la selección es descartar características irrelevantes y obtener el mejor subconjunto que mejore la precisión de la tarea de clasificación. Para este proceso, se aplican varios algoritmos de filtrado y selección de atributos. El algoritmo basado en evaluar un subconjunto de atributos teniendo en cuenta la capacidad individual de predicción de cada característica, junto con el grado de redundancia entre ellos [19]; El evaluador que utiliza el nivel de correlación de los atributos [20]; El algoritmo que evalúa los subconjuntos de atributos usando un esquema de aprendizaje basado arboles de decisión [21]; y el algoritmo que evalúa los valores de atributo por sus repeticiones en las instancias y teniendo en cuenta los valores de atributos de la instancia más cercana de la misma clase [22].

Un análisis de la correlación de atributos y una esquema de envoltura basado en arboles de decisión fueron seleccionados como el algoritmo de selección de atributos, tal como se demuestra en diversos trabajos [3, 23, 24, 25]. Este análisis tiene la finalidad de obtener el mejor subconjunto de características, los siguientes atributos fueron obtenidos como relevantes: posición de la entidad, distancia de la entidad, modo del evento, tipo de preposición, hiperónimo de la entidad, primera palabra del contexto izquierdo ($n_k=1$) y primera palabra del contexto derecho ($n_j=1$).

Con esta tarea se disminuye la dimensión del espacio característico de la representación de los pares, se obtienen 7 características como significativas (ver Figura 2) y se rechazan 14 como irrelevantes.

5 Clasificación

En este artículo, la clasificación de roles semánticos consiste en determinar el tipo de rol existente (agente, objeto, beneficiario, circunstante u otro) en pares EVENTO:ENTIDAD.

Tabla 2 Conjunto de características

Descripción	Valores posibles
Posición de la entidad	Valor nominal {izquierda=1; derecha =2}
Distancia de la entidad	Valor numérico que indica el número de palabras entre entidad y evento.
Modo del evento	Valor nominal {indicativo=1; subjuntivo=2; imperativo=3; infinitivo=4; gerundio=5; participio=6}
Tiempo del evento	Valor nominal {presente=1; imperfecto=2; futuro=3; pasado=4; condicional=5}
Persona del evento	Valor nominal {primera persona=1; segunda=2; tercera=3}
Número del evento	Valor nominal {singular=1; plural=2}
Longitud del evento	Valor numérico que indica la longitud del evento expresada en número de palabras
Número de la entidad	Valor nominal {singular=1; plural=2; invariable=3}
Grado de la entidad	Valor nominal {aumentativo=1; diminutivo=2; sin información=0}
Entidad definida	Valor booleano que indica sí o no la entidad es definida {si=1; no=0}
Longitud de la entidad	Valor numérico que indica la longitud de la entidad en número de palabras
Entidad nombrada	Valor booleano que indica sí o no es una entidad nombrada {si=1; no=0}
Tipo de preposición	Valor nominal {a=1; ante=2; bajo=3; cabe=4; con=5; contra=6; de=7; desde=8; durante=9; en=10; entre=11; hacia=12; hasta=13; mediante=14; para=15; por=16; según=17; sin=18; sobre=19, tras=20, sin preposición=0}
Tipo de entidad	Valor nominal {nombre propio=1; nombre común=2}
Hiperónimo de la entidad	Valor nominal {agente animado=1; objeto=2; otro=0}
Categoría del contexto izquierdo y derecho ($n_k=1,2,3$)	Valor nominal {adjetivo=1; adverbio=2; determinante=3; sustantivo=4; verbo=5; pronombre=6; conjunción=7; preposición=8; signo=9}

<entidad>La Junta General Ejecutiva del IFE</entidad>
<evento>enviará</evento> paquete electoral a mexicanos en el extranjero.

Características= (1,0,1,0,5,0,0)

Fig. 2. Vector de características relevantes

Tabla 3. Resultados de la clasificación con las características relevantes

Clases	Algoritmo			
	K-NN	SVM	NaïveBayes	C4.5
EA	84.4	86.9	86.4	87.1
EO	83.7	87.3	77.7	87.9
EB	80.9	82.6	79.5	85.2
EC	82.2	81.8	77.7	83.7
EX	77.2	78.6	70.5	79.3
Promedio	81.6	83.4	78.3	84.6

Por lo tanto, surge la necesidad de construir un clasificador semántico de roles, el cual se base en el conjunto de pares de entrenamiento con las características relevantes, para predecir la clase desconocida para los pares de pruebas.

La tarea de la clasificación de roles se lleva a cabo mediante la construcción y comparación de diversos clasificadores: el método del *k*-vecino más cercano (*K-Nearest Neighbors*) [26] que estima la función de densidad de los pares a predecir por cada clase basándose en el conjunto de entrenamientos y prototipos; las máquinas de soporte vectorial (*Support Vector Machine*) [27] que construyen un conjunto de hiperplanos en un espacio *n*-dimensional de los pares de entrenamiento, estos hiperplanos son utilizados para predecir la clase de los nuevos pares; El clasificador bayesiano ingenuo (*Naïve-Bayes*) que se basa en el teorema de Bayes y su función es encontrar la hipótesis más probable que describa los pares de prueba dado sus valores de atributos y con esto obtener la probabilidad de que conocidos los valores que describen a un par, éste pertenezca a una clase dada [28]; C4.5 es un algoritmo que realiza la inducción a partir de ejemplos preclasificados generando un árbol de decisión con los datos mediante particiones realizadas recursivamente [29].

6 Experimento

A partir del corpus de sentencias de noticias se identificaron manualmente 3032 pares, los cuales los cuales se dividen en 2274 pares para la fase de entrenamiento y 758 pares para las pruebas. Los clasificadores fueron construidos con los 2274 pares de entrenamiento para generar las reglas, el modelo estadístico o matemático según sea el caso de cada algoritmo con la finalidad de predecir la clase para los pares a probar.

Todos los experimentos se realizan con los 758 pares de prueba, a los cuales se predice su tipo de rol semántico, los pares son caracterizados a partir de los textos de noticias. Se realiza una comparativa entre las características relevantes y el conjunto total de ellas con los cuatro

		Clase predicha	
		Negativos	Positivos
Clase actual	Negativos	a	b
	Positivos	c	d

$$Recuerdo\ (R) = \frac{d}{c+d}$$

$$Precisión\ (P) = \frac{d}{b+d}$$

$$Medida\ F = 2 * \frac{P * R}{P + R}$$

Las entradas de la matriz de confusión tienen el siguiente significado:

- a es el número de predicciones incorrectas que una instancia sea negativa.
- b es el número de predicciones correctas que una instancia sea negativa.
- c es el número de predicciones incorrectas que una instancia sea positiva.
- d es el número de predicciones correctas que una instancia sea positiva.

Fig. 3. Obtención de la medida F, para la cual se utiliza la matriz de confusión de salida del clasificador, la precisión y el recuerdo

clasificadores descritos anteriormente (K-NN, SVM, NaïveBayes y C4.5).

La entrada es el conjunto de 758 pares (EVENTO:ENTIDAD) caracterizados; la tarea consiste en clasificar cada uno de acuerdo a una de las cinco clases de roles semánticos.

Para obtener la eficiencia de los clasificadores en cada clase por cada grupo de características se utiliza la matriz de confusión [30] y las medidas de *Precisión* y *Recuerdo,* éstas son usadas para obtener la media armónica entre ellas, Medida-F (ver Figura 3), la cual se refleja en las Tablas 3 y 4.

Los resultados para las cuatro clases con los clasificadores considerando el total de características (21) son mostrados en la Tabla 4 y en la Figura 5 se muestra esta clasificación de manera gráfica.

En la Figura 4 se muestra la clasificación de roles utilizando las características relevantes (7), expuestas en la Tabla 3, la cual demuestra que el algoritmo C4.5 proporciona el mejor desempeño en todas las categorías, por ejemplo en la categoría EA la eficiencia es de 87.1% contra el 86.4%, 86.9% y 84.4% de los algoritmos NaïveBayes, SVM y K-NN respectivamente.

Nosotros podemos concluir, con base en la Figura 4 y 5, que para la clasificación de roles semánticos, el mejor algoritmo es el basado en reglas de decisión C4.5, ya que presenta en mejor desempeño general en ambos casos, con todas y con el grupo de características relevantes.

Tabla 4. Resultados de la clasificación con todas las características

Clases	Algoritmo			
	K-NN	SVM	NaïveBayes	C4.5
EA	83.4	80.8	83.7	85.2
EO	73.7	82.8	79.1	83.8
EB	65.2	76.9	77.2	81.4
EC	63.7	66.3	75.1	80.1
EX	67.2	75.3	68.1	78.8
Promedio	70.7	76.4	76.6	81.8

Además, claramente se puede observar que la tarea de selección de atributos que reduce el espacio característico de 21 a 7 rasgos, mejora el desempeño de los clasificadores obteniendo mejores resultados en la predicción del tipo de rol semántico que representa cada par. Haciendo uso de las características relevantes se mejora el desempeño de todos los clasificadores, por ejemplo para el algoritmo *C4.5* la eficiencia general pasa de 81.8% a 84.6%. En la Figura 6

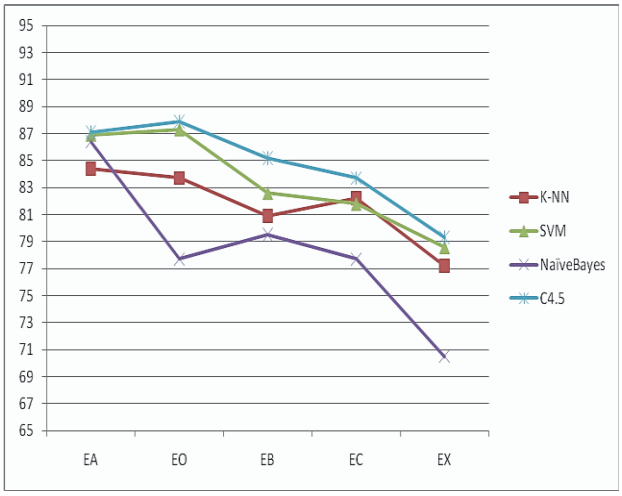

Fig. 4. Gráfica de la eficiencia de los cuatro algoritmos de clasificación por cada categoría de role semántico considerando las características relevantes

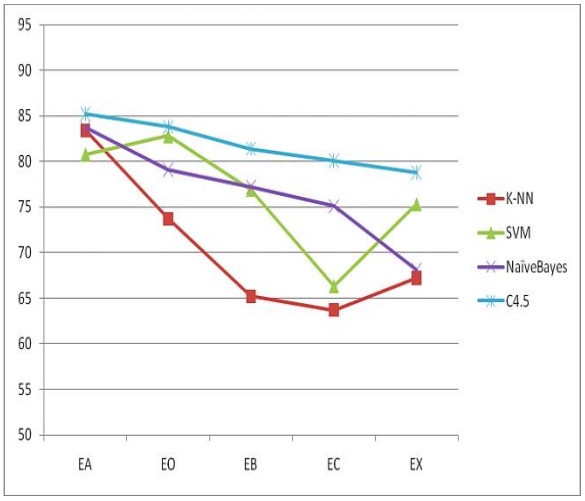

Fig. 5. Gráfica de la eficiencia de los cuatro algoritmos de clasificación por cada categoría de role semántico considerando todas las características

se muestra el desempeño del mejor algoritmo, árboles de decisión C4.5, para cada categoría de rol semántico, considerando los dos escenarios de pruebas: el total de características y el subgrupo de características relevantes.

7 Conclusiones y trabajos futuros

En este artículo se ha presentado una clasificación de roles semánticos entre eventos y sus actantes posicionándose en la teoría de Tesnière [1]. Esta clasificación está basada en características sintácticas, semánticas y contextuales con la finalidad de analizar su fusión; también se realiza una selección de características que da como resultado un subconjunto relevante. La tarea de clasificación de roles semánticos ha sido probada con los dos grupos de características (todas y relevantes) para cuatro algoritmos de clasificación: máquinas de soporte vectorial, algoritmo de los k-vecinos más cercanos, un clasificador de Bayes y el algoritmo basado en arboles de decisión C4.5.

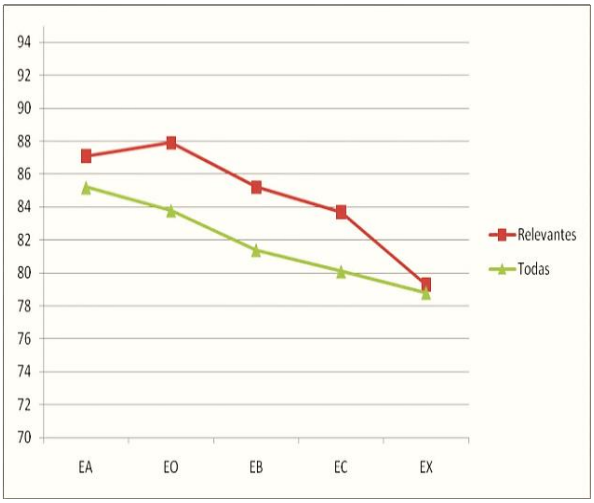

Fig. 6. Gráfica del desempeño del algoritmo C4.5 con todas las características y el grupo de relevantes

La selección de características ha proporcionado el siguiente subconjunto de siete características como relevantes: posición de la entidad, distancia de la entidad, modo del evento, tipo de preposición, hiperónimo de la entidad, primera palabra del contexto izquierdo (n_k=1) y primera palabra del contexto derecho (n_j=1). Los experimentos han demostrado que este subconjunto de características relevantes mejoran el desempeño de los clasificadores en todas las categorías de roles semánticos. Estas sietes características pertenecen a los tres tipos: sintácticas, semánticas y contextuales. Se obtiene una eficiencia total del 81.6 % para el algoritmo de los k-vecinos más cercanos, un 83.4 % para las máquinas de soporte vectorial, un 78.3 % para el clasificador de Bayes y un 84.6 % para el clasificador basado en arboles de decisión C4.5.

Por lo tanto, es notable que para la clasificación de cuatro categorías de roles semánticos, AGENTE, OBJETO, BENEFICIARIO, CIRCUNSTANTE y una quinta categoría llamada OTROS, el mejor clasificador resulta ser los arboles de decisión C4.5 utilizando además, el subconjunto de características relevantes que pertenecen a información sintáctica, semántica y contextual de los pares EVENTO:ENTIDAD.

La clasificación de roles semánticos que se realiza en este artículo otorga un tipo de rol semántico para un par EVENTO:ENTIDAD donde se desconoce dicho rol. Esta etiqueta del rol y sus componentes pueden ser utilizados en tareas de representación del conocimiento como información ontológica, es decir como tripletas del tipo Evento-Rol-Actante.

Referencias

1. **Tesnière, L. (1976).** *Éléments de syntaxe structurale (2e ed.).* Paris: Klincksieck.
2. **Halliday, M.A.K. (1994).** *An introduction to functional grammar* (2nd ed.). London: Routledge.
3. **Celli, F. (2010).** UNITN: Part-Of-Speech counting in relation extraction. *5th International Workshop on Semantic Evaluation (ACL 2010),* Uppsala, Sweden, 198–201.
4. **Tratz, S. & Hovy, E. (2010).** ISI: automatic classification of relations between nominals using a maximum entropy classifier. *5th International Workshop on Semantic Evaluation (ACL 2010),* Uppsala, Sweden, 222–225.
5. **Szarvas, G. & Gurevych, I. (2010).** TUD: semantic relatedness for relation classification. *5th*

International Workshop on Semantic Evaluation (ACL 2010), Uppsala, Sweden, 210–213.

6. **Pal, S., Pakray, P., Das, D., & Bandyopadhyay, S. (2010).** JU: a supervised approach to identify semantic relations from paired nominals. *5th International Workshop on Semantic Evaluation (ACL 2010),* Uppsala, Sweden, 206–209.
7. **Chen, Y., Lan, M., Su, J., Zhou, Z.M., & Xu, Y. (2010).** ECNU: effective semantic relations classification without complicated features or multiple external corpora. *5th International Workshop on Semantic Evaluation (ACL 2010)*, Uppsala, Sweden, 226–229.
8. **Rosario, B. & Hearst, M.A. (2004).** Classifying semantic relations in bioscience texts. *42nd Annual Meeting of the Association for Computational Linguistics (ACL'04),* Barcelona, Spain, 430–437.
9. **Rink, B. & Harabagiu, S. (2010).** UTD: classifying semantic relations by combining lexical and semantic resources. *5th International Workshop on Semantic Evaluation (ACL 2010),* Uppsala, Sweden, 256–259.
10. **Negri, M. & Kouylekov, M. (2010).** FBK NK: a wordNet-based system for multi-way classification of semantic relations. *5th International Workshop on Semantic Evaluation (ACL 2010),* Uppsala, Sweden, 202–205.
11. **Chambers, N., Wang, S., & Jurafsky, D. (2007).** Classifying Temporal Relations between Events. *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07),* Prague, Czech Republic, 173–176.
12. **Kaneiwa, K., Iwazume, M., & Fukuda, K. (2007).** An upper ontology for event classifications and relations. *20th Australian joint conference on Advances in artificial intelligence*, Gold Coast, Australia, 394–403.
13. **Téllez, A. (2005).** *Extracción de Información con Algoritmos de Clasificación*. Tesis de maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, México.
14. **Zhang, Z. (2004).** Weakly-supervised relation classification for information extraction. *Thirteenth ACM international conference on Information and knowledge management (CIKM'04),* Washington, DC., 581–588.
15. **Gildea, D. & Jurafsky, D. (2002).** Automatic labeling of semantic roles. *Computational Linguistics*, 28(3), 245–288.
16. **Xue, N. & Palmer, M. (2004).** Calibrating features for semantic role labeling. *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004),* Barcelona, Spain, 88–94.
17. **Salton, G. Wong, A., & Yang, C.S. (1975).** A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
18. **Padró, L., Collado, M., Reese, S., Lloberes, M., & Castellón, I. (2010).** FreeLing 2.1: Five Years of Open-Source Language Processing Tools. *7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 931–936.
19. **Hall, M.A. (1999).** *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, The University of Waikato, Hamilton, New Zealand.
20. **Liu, H. & Setiono, R. (1996).** A probabilistic approach to feature selection - A filter solution. *13th International Conference on Machine Learning (ICML'96),* Bari, Italy, 319–327.
21. **Kohavi, R. & John, G.H. (1997).** *Wrappers for feature subset selection*. Artificial Intelligence, 97(1-2), 273–324.
22. **Kira, K. & Rendell, L.A. (1992).** A practical approach to feature selection. *Ninth International Workshop on Machine Learning (ML92),* Aberdeen, Scotland, 249–256.
23. **Tovar, M., Reyes, J.A., Montes, A., Vilariño, D., Pinto, D., & León, S. (2012).** BUAP: A first approximation to relational similarity measuring. *First Joint Conference on Lexical and Computational Semantics,* Montreal, Canada, 502–505.
24. **Polaka, I. (2011).** Feature selection approaches in antibody display data analysis. *8th International Scientific and Practical Conference*, vol. II, Rezekne, Latviapp. 16–23.
25. **Alibeigi, M., Hashemi, S., & Hamzeh, A. (2011).** Unsupervised feature selection based on the distribution of features attributed to imbalanced data sets. *International Journal of Artificial Intelligence and Expert Systems*, 2(1), 14–22.
26. **Aha, D.W., Kibler, D., & Albert, M.K. (1991).** Instance-based learning algorithms. *Machine Learning*, 6(1), 37–66.
27. **Chang, Ch. & Lin, Ch. (2001).** LIBSVM - A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST),* 2(3), 27.
28. **John, G.H. & Langley, P. (1995).** Estimating continuous distributions in Bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence (UAI'95),* Montreal, Canada, 338–345.

29. **Quinlan, J.R. (1993).** *C4.5: Programs for Machine Learning*, San Mateo, Calif.: Morgan Kaufmann Publishers.
30. **Kohavi, R. & Provost, F. (1998).** Glossary of Terms, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning*, 30(2-3).

José A. Reyes recibió el grado de M.C. en Ciencias de la Computación en el Centro Nacional de Investigación y Desarrollo Tecnológico en 2008. Actualmente, se encuentra estudiando en el programa de Doctorado en Ciencias de la Computación en el Centro Nacional de Investigación y Desarrollo Tecnológico desde el 2009 a la fecha.

Azucena Montes Doctor en Ciencias por la *Université Paris-Sorbonne*, Francia en 2002. Profesor-Investigador de tiempo completo en el Centro Nacional Investigación y Desarrollo Tecnológico de 2002 a 2012. Actualmente, se encuentra en la Universidad Nacional Autónoma de México en el grupo de Ingeniería Lingüística. Sus áreas de interés en la investigación son: Semántica Cognitiva, Representación del conocimiento y Lingüística Computacional.

Juan G. González. Doctor en Ciencias Computacionales por el Centro de Investigación en Computación en el Instituto Politécnico Nacional, México, D.F., en el 2006. Actualmente, es profesor-investigador de tiempo completo del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Sus áreas de interés son: Modelado Semántico, Computación consciente del contexto, Servicios de recomendación sensibles al contexto, Cómputo ubicuo y HCI.

David E. Pinto. Doctor en Informática por la Universidad Politécnica de Valencia, España, en 2008. Actualmente es profesor-investigador de tiempo completo en la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla, sus áreas de interés en la investigación se relacionan con el Procesamiento del Lenguaje Natural, Ontologías y Recuperación de Información.

Articulo recibido el 16/10/2012; aceptado el 03/04/2013