Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 67-74 recibido 20-01-10 revisado 19-02-10 aceptado 04-03-10

Selección de características para la clasificación de preguntas multilingüe

Feature selection for multilingual question classification

David Tomás y José L. Vicedo
Depto. de Lenguajes y Sistemas Informáticos - Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
{dtomas,vicedo}@dlsi.ua.es

Resumen: Este trabajo presenta un estudio sobre el rendimiento de diferentes métodos de selección de características aplicados a la tarea de clasificación de preguntas en diferentes idiomas. El estudio se ha realizado sobre un corpus paralelo de preguntas en cuatro idiomas: inglés, español, italiano y catalán.
Palabras clave: Clasificación de preguntas, selección de características, búsqueda de respuestas, aprendizaje supervisado, multilingüe

Abstract: This paper presents a study on the performance of different features selection methods applied to question classification in different languages. This study was carried out on a parallel corpus of questions in four languages: English, Spanish, Italian, and Catalan.
Keywords: Question classification, feature selection, question answering, supervised learning, multilingual

1. Introducción

Los sistemas de *búsqueda de respuestas* o *question answering* (QA) tienen como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuario mediante lenguaje natural. En los sistemas de QA, un primer paso para poder devolver la respuesta solicitada por el usuario es analizar la pregunta y comprenderla, saber por qué se nos está preguntando.

La *clasificación de preguntas* se ha desmarcado como una tarea en sí misma dentro del mundo de procesamiento del lenguaje natural y del QA. Su objetivo es identificar de forma automática qué se nos está preguntando, categorizando las preguntas en diferentes clases semánticas en función del tipo de respuesta esperada. Así, ante preguntas como "¿Quién inventó el teléfono?" o "¿Dónde está el Big Ben?", un sistema de clasificación de preguntas detectaría que se está preguntando por una *persona* o un *lugar* respectivamente.

En la actualidad la gran mayoría de sistemas de clasificación de preguntas se basan en el empleo de técnicas de aprendizaje automático (Li y Roth, 2002). Uno de los principales problemas que afrontan este tipo de sistemas es el tamaño del espacio de aprendizaje, compuesto habitualmente por miles de características. Algunas de estas características, quizás la inmensa mayoría, son claramente irrelevantes o redundantes para el aprendizaje. En la práctica, el funcionamiento de los algoritmos de aprendizaje puede ser mejorado frecuentemente mediante un proceso de preselección de características o *feature selection* (Cardie, 1996). Numerosos experimentos han demostrado que añadir características irrelevantes o redundantes provoca un deterioro en el funcionamiento de los algoritmos de aprendizaje (Wang y He, 2004).

El objetivo de este trabajo es la evaluación y comparación de distintos métodos de selección automática de características para la tarea de clasificación de preguntas dentro de un contexto multilingüe. Los objetivos buscados son varios: determinar la capacidad de estos métodos para mejorar la precisión del clasificador, detectar qué parte del vocabulario del problema se puede eliminar sin afectar a la precisión del mismo y, por último, ver cómo varía su rendimiento en cuatro idiomas diferentes (inglés, español, italiano y catalán). Mientras que en clasificación de textos son diversos los estudios realizados sobre la evaluación de métodos de selección de características (Yang y Pedersen, 1997; Forman, 2003), no existen estudios previos de su aplicación a la tarea de clasificación de preguntas ni de

ISSN 1135-5948

su funcionamiento en diferentes idiomas.

En el resto del artículo, la sección 2 describe las técnicas de selección de características estudiadas; la sección 3 muestra el sistema de clasificación empleado; la sección 4 resume los experimentos y resultados obtenidos; finalmente, la sección 5 muestra las conclusiones de este trabajo.

2. Selección de características

En muchas situaciones prácticas que se dan al abordar tareas de clasificación automática hay demasiadas características de aprendizaje que manejar por parte de los algoritmos. Algunas de estas características, quizás la inmensa mayoría, son claramente irrelevantes o redundantes para el aprendizaje. Realizar una selección previa de características aporta diversos beneficios al proceso posterior de aprendizaje. Uno de ellos es la reducción de dimensionalidad que se produce al descartar características irrelevantes. Hay muchos algoritmos que no son capaces de manejar espacios de alta dimensionalidad y trabajar con vectores de características de gran tamaño. Esta reducción de tamaño puede hacer que el problema sea asequible para algoritmos que de otra forma no podrían afrontarlo, además de mejorar de forma general la velocidad de cómputo a la hora de entrenar y evaluar el clasificador. Otra ventaja de la selección de características es la eliminación de *ruido* que se produce al descartar características que realmente no contribuyen al proceso de clasificación. Esta reducción permite minimizar el problema del sobreajuste (*overfitting*).

Frente a las aproximaciones manuales de selección de características, existen técnicas que permiten determinar estadísticamente cuáles son las características que aportan más información al proceso de aprendizaje. Para llevar a cabo la selección se emplea una función que mide la importancia de cada característica en la tarea de clasificación, manteniendo un subconjunto de las características consideradas más relevantes por dicha función. De entre estas funciones podemos destacar el *umbral de frecuencia* (*frequency thresholding*), *information gain* (IG), *mutual information* (MI) y χ^2. En este trabajo vamos a centrarnos en el estudio del *umbral de frecuencia*, IG y χ^2, ya que resultados experimentales previos han demostrado que MI funciona significativamente peor a la hora de seleccionar características en tareas de clasificación (Yang y Pedersen, 1997).

El método del *umbral de frecuencia* se basa en mantener aquellas características que se dan al menos un número determinado de veces en el corpus. Este umbral puede referirse tanto al número de preguntas en las que aparece como al número total de veces que aparece en la colección. La asunción detrás de esta técnica es que las características que ocurren de forma escasa en el corpus no son relevantes para la predicción de las clases o no resultan relevantes para el rendimiento global del sistema. La eliminación de estas características poco frecuentes reduce la dimensionalidad del espacio de aprendizaje, pudiendo mejorar el funcionamiento del clasificador si los términos eliminados introducían ruido en la construcción del modelo.

La segunda técnica estudiada es χ^2, empleada en estadística para evaluar la independencia de dos eventos. En selección de características, los dos eventos son la ocurrencia de una característica y la de una clase. Proporciona una medida de cuánto se desvía la frecuencia esperada de la frecuencia observada. Un valor grande de χ^2 indica que la hipótesis de independencia, que implica que los valores observados y esperados son similares, es incorrecta. La selección se lleva a cabo ordenando las características en función del valor de χ^2 y escogiendo las m mejor valoradas.

Por último, IG mide el número de bits de información obtenidos para la predicción de clases conociendo la presencia o la ausencia de una característica en una instancia. Para la selección, al igual que sucedía con χ^2, se computa IG sobre cada una de las características, ordenando la lista resultante y eligiendo las m características mejor valoradas.

3. Configuración del sistema

En este apartado se describe el sistema de clasificación de preguntas sobre el que se aplicaran las distintas técnicas de selección de características. Como en todo sistema de aprendizaje basado en corpus, se deben definir el algoritmo de aprendizaje, las características de aprendizaje y el corpus de trabajo.

De entre los algoritmos utilizados habitualmente en clasificación de preguntas, SVM ha destacado por su buen rendimiento en esta tarea (Zhang y Lee, 2003; Bisbal et al., 2005), haciendo que nos decantemos por él para configurar el sistema de clasificación. Para los experimentos se ha utilizado la implementa-

ción de SVM proporcionada por el conjunto de herramientas de aprendizaje Weka (Witten y Frank, 2005), empleando el *kernel lineal* y el parámetro de penalización $C = 1$.

Para mantener la independencia del sistema con respecto a otras herramientas o recursos lingüísticos y poder adaptarlo de forma inmediata a diferentes idiomas, se han empleado como únicas características de aprendizaje los n-gramas obtenidos del propio corpus de preguntas: unigramas (1-gramas), bigramas (2-gramas), trigramas (3-gramas) y las combinaciones de unigramas y bigramas (1+2-gramas), unigramas y trigramas (1+3-gramas), bigramas y trigramas (2+3-gramas) y unigramas, bigramas y trigramas (1+2+3-gramas). El único preproceso llevado a cabo sobre el corpus consistió en la transformación a minúsculas y la eliminación de los signos de puntuación.

Por lo que respecta al corpus de trabajo, ante la inexistencia de conjuntos adecuados de preguntas para la tarea de clasificación multilingüe hemos desarrollado nuestro propio corpus en cuatro idiomas diferentes: inglés, español, italiano y catalán. El corpus en inglés se obtuvo de las preguntas de evaluación definidas para la tarea de QA de las conferencias TREC, desde 1999 (TREC-8) hasta 2003 (TREC-12).[1] Una vez recopiladas las preguntas en inglés, se procedió a la traducción manual de las mismas a los otros tres idiomas, obteniendo finalmente un corpus paralelo de 2.393 preguntas. Este corpus fue etiquetado con 15 clases diferentes, basándonos en el primer nivel de la jerarquía extendida de entidades nombradas de Sekine (Sekine, Sudo, y Nobata, 2002), sobre la que se añadieron las clases *definición* y *acrónimo*. Éstas no existían originalmente en la jerarquía de Sekine y fueron incluidas para aumentar la cobertura de la taxonomía. El acuerdo en el etiquetado de preguntas (*kappa agreement*) entre revisores fue de 0,87.

4. Evaluación

Se han llevado a cabo cuatro experimentos diferentes. En primer lugar se ha evaluado el rendimiento del clasificador SVM con cada uno de los conjuntos de características para los cuatro idiomas del corpus. Esta evaluación servirá para comparar seguidamente el rendimiento de las tres técnicas de selección de características en cada uno de los idiomas.

[1] `http://trec.nist.gov/data/qa.html`.

Característica	Ing	Esp	Ita	Cat
1-gramas	3764	4164	4315	4190
2-gramas	8465	8578	8644	8625
3-gramas	10015	10358	9842	10391
1+2-gramas	12229	12742	12959	12815
1+3-gramas	13779	14522	14157	14581
2+3-gramas	18480	18936	18486	19016
1+2+3-gramas	22244	23100	22801	23206

Tabla 1: Número de elementos del vector de aprendizaje para cada una de las características en los cuatro idiomas del corpus.

Característica	Ing	Esp	Ita	Cat
1-gramas	81,64	80,97	**79,62**	80,54
2-gramas	76,78	75,33	72,68	76,00
3-gramas	54,55	58,10	56,76	62,89
1+2-gramas	**81,70**	**81,17**	79,45	**81,03**
1+3-gramas	80,07	79,66	78,49	79,94
2+3-gramas	75,18	72,70	70,47	74,34
1+2+3-gramas	80,36	80,07	78,39	80,43

Tabla 2: Precisión obtenida con cada una de las características. Los mejores valores para cada idioma se muestran en negrita.

Para evitar dividir el corpus de preguntas en un conjunto de evaluación y otro de entrenamiento, en los experimentos se ha realizado una validación cruzada equilibrada en 10 particiones (*stratified 10-fold cross validation*). Para certificar la validez de los resultados se ha empleado el test estadístico *t-test*, que permite minimizar la posibilidad de que la diferencia de precisión obtenida entre dos sistemas o configuraciones pueda ser fortuita a causa del conjunto de datos empleado durante el proceso de experimentación. Para valorar los resultados en este tipo de test, se aporta un grado de confianza representado por el valor p. El grado de confianza que se considera aceptable en experimentación suele ser de $p < 0,05$ o $p < 0,01$, indicando que la diferencia obtenida entre los sistemas no se debe al azar con una seguridad del 95 % o del 99 % respectivamente.

4.1. Configuración original

En este apartado se muestran los resultados obtenidos con el sistema definido en la sección 3, sin llevar a cabo ningún tipo de selección de características. La tabla 1 muestra el tamaño del vector de aprendizaje para cada uno de los idiomas. La precisión obtenida por el clasificador se muestra en la tabla 2.

Los resultados obtenidos para cada uno de los idiomas son similares. En el caso de los 1-

gramas, los resultados para inglés (81,64 %) son ligeramente superiores al resto de idiomas, siendo el italiano el que peor resultados obtiene (79,62 %). Esta tendencia se repite de nuevo para los 2-gramas. En los experimentos con 3-gramas, sin embargo, se obtienen los peores resultados para el inglés (54,55 %), siendo los mejores resultados los de catalán (62,89 %). Las combinaciones de características, de forma similar a los experimentos individuales, obtienen mejores resultados para inglés que para el resto, siendo el italiano el idioma con peores resultados por parte del sistema. La única excepción es la combinación de 1+2+3-gramas, donde los experimentos en catalán son los que mejores resultados ofrecen. Una explicación para estas ligeras diferencias de rendimiento es el grado de flexión verbal y nominal de cada uno de los idiomas tratados. Este grado de flexión hace que el número de n-gramas diferentes que nos podemos encontrar sea mayor para idiomas como el español, el italiano y el catalán que para el inglés, lo cual dificulta la tarea de clasificación. Para el inglés, por ejemplo, existen un total de 3764 términos diferentes en el corpus de preguntas, mientras que para el italiano este número es de 4315, lo que supone casi un 15 % de incremento con respecto al anterior.

4.2. Umbral de frecuencia

La cantidad de texto de la que disponemos en la tarea de clasificación de preguntas es escasa, por lo que la frecuencia de los n-gramas en el corpus es habitualmente baja. Por esta razón no es conveniente establecer un umbral elevado en la eliminación de términos en función de su frecuencia, ya que la reducción podría ser demasiado drástica. En la figura 1 se muestra la evolución del número de 1-gramas, 2-gramas y 3-gramas en el corpus en inglés dependiendo del umbral de frecuencia. Se puede observar que, simplemente eliminando aquellos n-gramas que aparecen una única vez en el corpus, obtenemos una reducción de dimensionalidad notable. Vamos a centrarnos por ello en estudiar cómo afecta al sistema la eliminación de aquellos n-gramas que aparecen una única vez en el corpus, conocidos como *hapax legomena*.

La tabla 3 muestra cómo varía el número de 1-gramas, 2-gramas y 3-gramas al eliminar los *hapax legomena* del corpus. La tabla muestra cómo la reducción de características

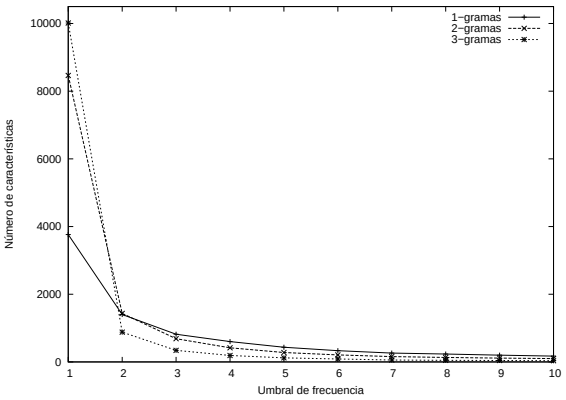

Figura 1: Número de 1-gramas, 2-gramas y 3-gramas en el conrpus en inglés dependiendo del umbral de frecuencia de corte.

Característica	Ing	Esp	Ita	Cat
1-gramas	62,81	66,11	68,16	66,18
2-gramas	83,05	80,40	82,64	81,19
3-gramas	91,22	89,08	90,20	88,88

Tabla 3: Porcentaje de reducción de características sobre el espacio original tras la eliminación de los *hapax legomena*.

para 1-gramas en los diferentes idiomas se sitúa entre el 62 % y el 68 %, para 2-gramas entre el 80 % y el 83 %, y para 3-gramas entre el 88 % y el 91 %.

Los resultados obtenidos por el sistema con este tipo de selección se muestran en la tabla 4. La comparación de estos resultados con respecto a los obtenidos con el sistema original (tabla 2) se puede ver en la tabla 5. Los símbolos "$\gg$" y "$>$" indican que los resultados tras la selección son significativamente mejor que los originales con un grado de confianza $p < 0,01$ y $p < 0,05$ respectivamente. De forma equivalente, "$\ll$" y "$<$" indican que los resultados tras la selección son significativamente peor que los originales. El símbolo "$=$" indica que no hay una diferencia significativa de funcionamiento entre las dos configuraciones.

Esta comparación revela que la reducción de dimensionalidad obtiene mejoras significativas para muchos de los tamaños de n-gramas y sus combinaciones. La única excepción es el caso de los 1-gramas, donde hay una pérdida sistemática de rendimiento, que llega a ser significativa en el caso del inglés. Esto se justifica porque con n-gramas de mayor tamaño (y las combinaciones de éstos) el número de características que introducen rui-

Característica	Ing	Esp	Ita	Cat
1-gramas	80,68	80,47	**78,56**	**80,32**
2-gramas	76,96	74,28	73,13	75,93
3-gramas	59,41	60,57	59,79	65,57
1+2-gramas	**81,74**	**80,50**	78,37	80,12
1+3-gramas	81,38	80,04	78,23	79,94
2+3-gramas	77,25	73,20	72,61	74,90
1+2+3-gramas	81,49	80,11	78,37	79,83

Tabla 4: Precisión obtenida por cada una de las características con la eliminación de *hapax legomena*.

Característica	Ing	Esp	Ita	Cat
1-gramas	$<$	$=$	$=$	$=$
2-gramas	$=$	$=$	$=$	$=$
3-gramas	$\gg$	$\gg$	$\gg$	$\gg$
1+2-gramas	$=$	$=$	$=$	$=$
1+3-gramas	$>$	$=$	$=$	$=$
2+3-gramas	$\gg$	$=$	$\gg$	$=$
1+2+3-gramas	$=$	$=$	$=$	$=$

Tabla 5: Comparación estadística de la precisión entre el experimento original y el experimento de eliminación de *hapax legomena*.

do en la clasificación aumenta de forma considerable. Eliminar aquellos n-gramas poco frecuentes permite la depuración del vector de características y la mejora en la precisión del clasificador. Por ejemplo, para la combinación 1+2+3-gramas en inglés, la precisión sufre una mejora significativa ($p < 0,01$) pasando de 80,36 % a 81,49 % y reduciendo el tamaño del vector de 22244 componentes a tan sólo 3714 (una reducción del 83,39 %).

El método del umbral de frecuencia suele ser considerado como una técnica *ad hoc* para mejorar la eficiencia de los clasificadores, pero no como un criterio fundamentado para la selección de características relevantes. De hecho no suele emplearse como una técnica agresiva de selección, ya que los términos poco frecuentes se consideran como informativos en tareas como la recuperación de información y clasificación de textos.

4.3. χ^2

Los resultados obtenidos con la selección mediante χ^2 se muestran en la figura 2. Por cuestiones de espacio se ofrecen sólo las gráficas para inglés. La figura 2 (a) muestra la precisión obtenida por el sistema empleando 1-gramas y distintas variantes del número m de características seleccionadas. Se observa que cualquier reducción llevada a cabo sobre el conjunto original (tabla 1) provoca un deterioro en el rendimiento del sistema. La única reducción que no implica una pérdida significativa en la precisión se produce para m=3500. En la figura 2 (b) vemos los resultados para 2-gramas. En este caso, con m=7000 se obtiene mejor resultado que en el original (76,85 % frente a 76,78 %) aunque esta mejora no es significativa. Para el resto de casos se observa un comportamiento similar al de los 1-gramas. Con m=5000 se obtiene peor precisión que en el caso original, aunque esta diferencia no es estadísticamente significativa ($p < 0,05$). Por tanto, se puede reducir hasta m=5000 sin que haya una pérdida de precisión significativa en el sistema. La figura 2 (c) muestra los resultados para 3-gramas. Aquí se obtiene un mejor valor para m=8000 (55,50 %) aunque este valor no es significativamente mejor que para el caso original (54,55 %). Para reducciones mayores, el sistema tiene una bajada de rendimiento significativo con respecto al original. En este sentido el comportamiento es bastante similar al caso de 1-gramas y 2-gramas. Con el resto de idiomas tampoco se obtienen mejoras significativas empleando esta técnica sobre los conjuntos de características mencionados.

En la figura 2 (d) vemos los resultados para 1+2-gramas. Aquí se obtiene un mejor valor para m=7000 (81,82 %) aunque este valor no es significativamente mejor que para el caso original (81,70 %). Se puede reducir hasta m=5000 sin que el resultado sea significativamente peor ($p < 0,01$). Los resultados para la combinación de 1+3-gramas se puede ver en la figura 2 (e). Aquí se obtiene un mejor valor para m=9000 (80,92 %) aunque este valor no es significativamente mejor que para el original (80,08 %). Se puede reducir hasta m=6000 sin que el resultado sea significativamente peor ($p < 0,01$). En la figura 2 (f) tenemos los resultados para 2+3-gramas. Aquí se obtiene un mejor valor para m=5000 (76,81 %), siendo significativamente mejor que para el caso original (75,18 %). Se puede reducir hasta m=3000 sin que el resultado sea significativamente peor ($p < 0,01$).

La última combinación de n-gramas, la que se obtiene mediante 1+2+3-gramas, merece un estudio más en detalle. En la figura 3 (a) se muestra la precisión obtenida para esta combinación en inglés. Se obtiene un máximo para m=11000 (menos de la mitad de características del conjunto original) obteniendo una precisión de 81,96 %, significativamente mayor ($p > 0,01$) que el experimento origi-

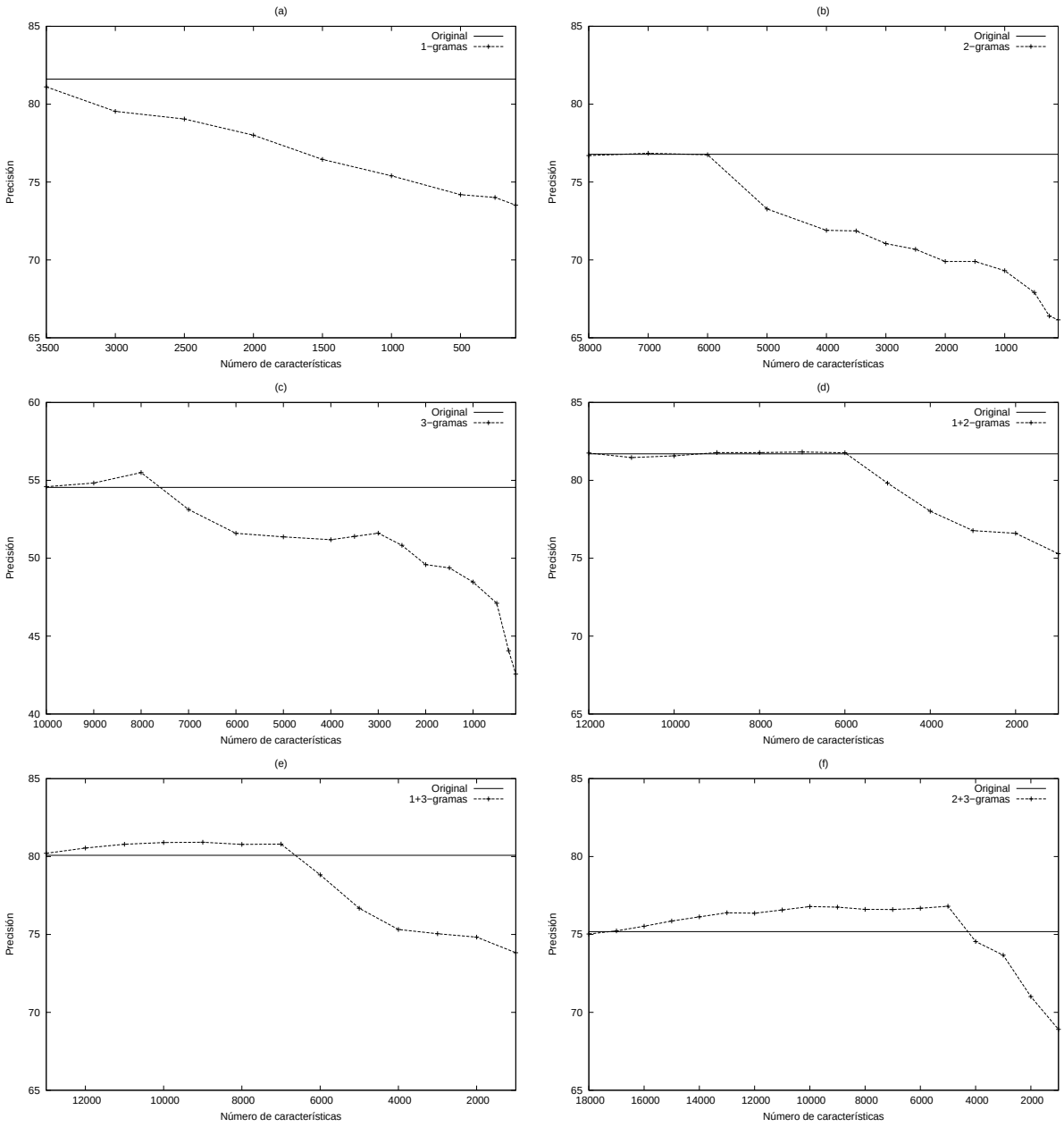

Figura 2: Precisión obtenida con χ^2 con distintas reducciones de dimensionalidad en inglés para (a) 1-gramas, (b) 2-gramas, (c) 3-gramas, (d) 1+2-gramas, (e) 1+3-gramas y (f) 2+3-gramas. *Original* representa la precisión obtenida en el experimento original sin selección.

nal. En la figura 3 (b) podemos ver los resultados para el corpus en español con estas mismas características. Al igual que sucedía con el corpus en inglés, esta combinación de características es la que más se beneficia de la aplicación de χ^2 en lo que a ganancia de rendimiento se refiere. Se obtiene un máximo de precisión (81,28 %) con m=12000 características, siendo este valor significativamente mejor ($p < 0,01$) que con el conjunto original de 23100 características (80,05 %). Para el resto de resultados podemos destacar que con m=5000 la precisión del sistema no es significativamente peor que para el caso original, reduciendo el tamaño del vector un 78,35 %. Los resultados para italiano se pueden ver en la figura 3 (c). En este caso la aportación de la selección es menos efectiva. La máxima precisión se alcanza con m=12000 (79,47 %). Este valor, sin embargo, no es significativamente mejor que el original (78,39 %). La única mejora significativa ($p < 0,05$) se consigue para m=13000, obteniendo un 79,43 %. Podemos reducir hasta m=8000 sin pérdida significa-

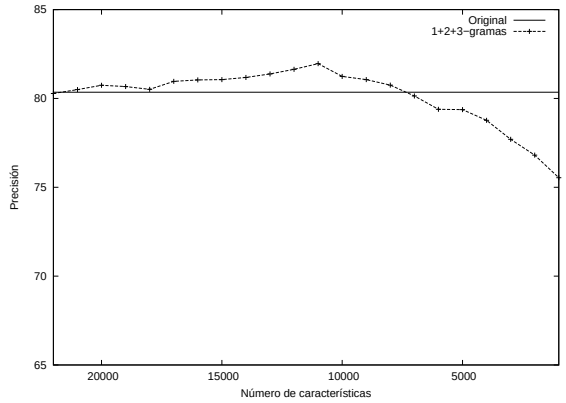

Figura 4: Precisión obtenida con IG para distintas reducciones de dimensionalidad en inglés para la combinación de 1+2+3-gramas. *Original* representa la precisión obtenida en el experimento original sin selección.

tiva ($p < 0,05$) de precisión. Por último, la figura 3 (d) muestra los resultados para catalán. La máxima precisión se alcanza con m=11000 (81,25 %), obteniendo una mejora significativa ($p < 0,05$) con respecto al valor original (80,43 %). Se puede conseguir reducir el número de características hasta m=5000 sin que haya una pérdida significativa de precisión ($p < 0,05$) con respecto al original.

4.4. Information Gain

Tras repetir los experimentos realizados con χ^2, el comportamiento ofrecido por IG ha resultado prácticamente equivalente al método anterior. Como ejemplo, la figura 4 muestra la precisión obtenida por el sistema sobre el corpus en inglés empleando la combinación 1+2+3-gramas. Se observa un máximo para m=11000, obteniendo una precisión de 81,96 %, significativamente mayor ($p < 0,01$) que el experimento original. Se puede observar que la gráfica obtenida con IG es prácticamente idéntica a la obtenida con χ^2 en la figura 3 (a), solapándose para la mayoría de valores de m. En ningún caso existe una diferencia estadísticamente significativa entre ambos métodos.

5. Conclusiones

En este trabajo se ha presentado un estudio sobre la selección de características aplicada a la tarea de clasificación automática de preguntas en un contexto multilingüe. Se ha definido un sistema basado en SVM y n-gramas, experimentando sobre un corpus de preguntas en cuatro idiomas diferentes: inglés, español, italiano y catalán.

La primera técnica de selección estudiada ha sido el *umbral de frecuencia*, que consiste en eliminar las características que aparecen de forma limitada en el corpus. Concretamente, se han eliminado aquellos n-gramas que aparecían una única vez en el conjunto de preguntas (*hapax legomena*). Pese a su simplicidad, esta técnica ha conseguido reducir considerablemente la dimensionalidad del espacio nativo de características, mejorando el rendimiento del sistema para todos los vectores de características con excepción de los 1-gramas. El mejor resultado obtenido para inglés (81,74 % para 1+2-gramas) ha superado al original (81,70 %), aunque esta diferencia no ha sido estadísticamente significativa. En cualquier caso, esta forma de selección ha conseguido descartar el 76,82 % de las características del vector original en este idioma.

Se han empleado otras dos técnicas de selección, χ^2 e IG, más sofisticadas que la anterior. Los experimentos realizados han demostrado que ambas obtienen resultados muy similares, mejorando los valores conseguidos con la técnica del umbral de frecuencia. Los resultados han sido especialmente buenos para la combinación 1+2+3-gramas (81,96 %). Se ha obtenido una mejora significativa para todos los idiomas con respecto a los experimentos originales, reduciendo el tamaño del vector de características a menos de la mitad del inicial. La mejora obtenida sugiere que estos métodos de selección consiguen efectivamente mantener los mejores n-gramas para la clasificación, mejorando el rendimiento final del sistema. Pese a que los mejores resultados con IG y χ^2 (81,96 % en inglés) han superado a los mejores resultados originales (81,70 %), esta diferencia no es significativa. Como ya mencionamos más arriba, la selección de características proporciona dos beneficios. En primer lugar, elimina características ruidosas y mejora el rendimiento del sistema. Aunque esta mejora es patente en los experimentos realizados no es estadísticamente significativa con respecto a los resultados originales. En segundo lugar, la selección de características reduce el número de características de aprendizaje y el coste computacional del entrenamiento y evaluación del clasificador. A diferencia del umbral de frecuencia, IG y χ^2 son procesos computacionalmente costosos, por lo que debe valorarse cuándo el coste de llevar a cabo la selección compensa la reducción de coste que se produce durante el aprendizaje

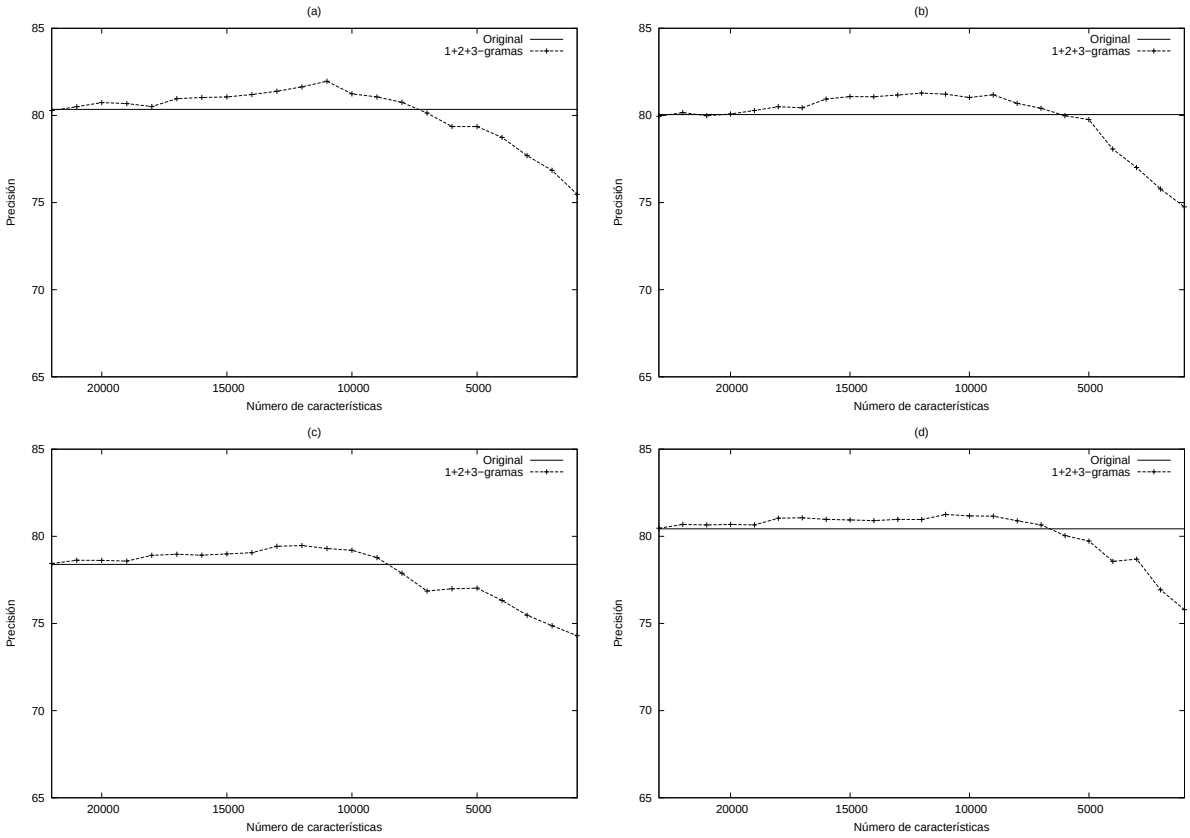

Figura 3: Precisión obtenida con χ^2 para distintas reducciones de dimensionalidad para la combinación 1+2+3-gramas en (a) inglés, (b) español, (c) italiano y (d) catalán. *Original* representa la precisión obtenida en el experimento original sin selección.

y la clasificación.

Bibliografía

Bisbal, Empar, David Tomás, Lidia Moreno, José L. Vicedo, y Armando Suárez. 2005. A multilingual svm-based question classification system. En *MICAI 2005*, volumen 3789 de *LNCS*, páginas 806–815. Springer.

Cardie, Claire. 1996. Automatic feature set selection for case-based learning of linguistic knowledge. En *Conference on Empirical Methods in Natural Language Processing*, páginas 113–126.

Forman, George. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

Li, Xin y Dan Roth. 2002. Learning question classifiers. En *Proceedings of the 19th international conference on Computational linguistics*, páginas 1–7, Morristown, NJ, EEUU. Association for Computational Linguistics.

Sekine, Satoshi, Kiyoshi Sudo, y Chikashi Nobata. 2002. Extended named entity hierarchy. En *LREC 2002*, páginas 1818–1824, Las Palmas, España.

Wang, Xizhao y Qiang He. 2004. Enhancing generalization capability of svm classifiers with feature weight adjustment. En *KES 2004*, volumen 3213 de *Lecture Notes in Computer Science*, páginas 1037–1043. Springer.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edición.

Yang, Yiming y Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. En *ICML '97*, páginas 412–420, San Francisco, CA, EEUU. Morgan Kaufmann Publishers Inc.

Zhang, Dell y Wee Sun Lee. 2003. Question classification using support vector machines. En *SIGIR '03*, páginas 26–32, Nueva York, NY, EEUU. ACM.