Estudio Bidireccional de un Sistema de RI Multilingüe Basado en Traducción de n-Gramas

Jesús Vilares Ferro[1], Adrián Blanco González[2], and David Vilares Calvo[1]

[1] Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 – A Coruña
{jvilares, david.vilares}@udc.es
[2] Departamento de Informática, Universidade de Vigo
Campus As Lagoas s/n, 32004 – Ourense
adbgonzalez@uvigo.es

Resumen Continuando nuestra investigación sobre el empleo de n-gramas de caracteres como unidad de traducción en sistemas de RI Multilingüe, este artículo analiza el comportamiento de nuestra solución en direcciones inversas de traducción a partir de sendos experimentos paralelos con consultas en inglés sobre textos en español y viceversa. Lo positivo de los resultados corrobora la validez de nuestra propuesta.

Keywords: Recuperación de Información Multilingüe; n-gramas de caracteres; Traducción Automática

1. Introducción

El tratamiento a nivel de *n-grama* —entendido como una secuencia de caracteres contiguos dentro de una palabra dada— es ya una técnica estándar del estado del arte en procesamiento de texto [10,12], empleándose en correspondencias aproximadas, tratamiento de errores ortográficos, identificación del idioma, detección de plagio, bioinformática, etc. Este éxito se debe a sus características inherentes: (a) simplicidad, al no precisar de recurso o conocimiento lingüístico alguno; (b) eficiencia, al procesarse en un solo paso; (c) robustez frente a variaciones y errores ortográficos; (d) completud, al conocerse de antemano el alfabeto de símbolos; y (e) independencia del idioma y dominio.

El presente trabajo se centra en su aplicación en *Recuperación de Información Multilingüe (RIM)*, un caso particular de RI en el cual consultas y documentos están en idiomas diferentes, por lo que se hace necesario algún tipo de fase de traducción intermedia que permita el establecimiento de correspondencias, si bien no es necesario obtener una única traducción o que ésta sea sintácticamente correcta. Los sistemas de RIM suelen emplear algún tipo de traducción sencilla a nivel de palabra. Desafortunadamente ésta es sensible a la presencia de errores ortográficos, palabras desconocidas, la falta de recursos lingüísticos apropiados, etc. Sin embargo, si consentimos en relajar todavía más las restricciones a la hora de traducir, de tal forma que ni siquiera precisemos de palabras completas,

este tipo de sistemas se pueden beneficiar de las ventajas que la utilización de n-gramas de caracteres como unidad de procesamiento pueden aportar, no sólo ya como unidad de indexación, sino también como unidad de traducción.

McNamee y Mayfield [9] fueron pioneros en este campo. Para ello empleaban un algoritmo de traducción directa de n-gramas que se basaba en técnicas estadísticas para alinear corpus paralelos a nivel de n-grama permitiendo así evitar algunas de las limitaciones de las técnicas basadas en diccionarios, tales como la necesidad de normalizar las palabras o la imposibilidad de traducir palabras desconocidas. Además, como se trata de una solución que no emplea ningún tipo de procesamiento particular dependiente del idioma, puede ser empleada cuando la disponibilidad de recursos lingüísticos es reducida. Sin embargo, esta primera aproximación pecaba de ser lenta y poco flexible.

En este contexto pretendemos desarrollar soluciones basadas también en traducción de n-gramas pero que sean más flexibles y eficientes.

2. Alineamiento de n-Gramas

Nuestra principal aportación la constituye un nuevo algoritmo de alineamiento de n-gramas, que consta de dos fases. Primeramente el corpus paralelo de entrada es alineado a nivel de palabra mediante GIZA++ [5], obteniendo como salida las probabilidades de traducción entre las palabras de ambos idiomas. Este primer paso actúa como filtro, ya que sólo aquellos pares de n-gramas correspondientes a palabras alineadas serán considerados para su posterior procesamiento, en contraste con la aproximación de [9], de granularidad mucho mayor al partir del corpus alineado a nivel de párrafo para luego procesar todos los n-gramas que contenían dos párrafos alineados. Partiendo de los alineamientos de palabras obtenidos, en la segunda fase del algoritmo se procede a calcular las medidas de alineamiento entre n-gramas empleando para ello medidas estadísticas de asociación [8]. Esta solución permite acelerar el proceso de entrenamiento, al concentrar la complejidad en la fase de alineamiento a nivel de palabra y de este modo facilitar el poder ensayar nuevas medidas de asociación u otros procedimientos en la fase final de alineamiento a nivel de n-gramas propiamente dicho. Otra ventaja de nuestra aproximación es la posibilidad de partir de diccionarios o alineamientos de palabras ya disponibles.

Se ha optado por emplear recursos de libre distribución para minimizar el coste de desarrollo y hacer el sistema más transparente. De este modo utilizamos la plataforma Terrier [11] como motor de recuperación, y el corpus paralelo Europarl [3] como entrada para el alineamiento.

2.1. Alineando Palabras Empleando Medidas de Asociación

Podemos ver nuestro algoritmo de alineamiento de n-gramas como una extensión de cómo utilizar medidas de asociación para generar diccionarios bilingües de palabras a partir de corpus paralelos alineados a nivel de párrafo [4]. En dicho contexto, dado un par de palabras (w_s, w_t) —donde w_s es una palabra

en la lengua origen y w_t su traducción candidata en la lengua destino—, sus frecuencias de coocurrencia en el corpus paralelo de entrada se pueden organizar en una *tabla de contingencia*:

	$T = w_t$	$T \neq w_t$	
$S = w_s$	O_{11}	O_{12}	$= R_1$
$S \neq w_s$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

La primera fila corresponde a las observaciones donde el párrafo en la lengua origen contiene la palabra w_s, y la segunda fila a aquéllas en las que dicho párrafo no la contiene. Lo mismo ocurre para las columnas con respecto al párrafo en la lengua destino y w_t. Las cuentas de estas celdas se denominan *frecuencias observadas*: O_{11}, por ejemplo, corresponde al número de párrafos alineados donde el párrafo en la lengua origen contiene w_s y su correspondiente paralelo contiene w_t; O_{12} correspondería al número de párrafos alineados donde el párrafo en la lengua origen contiene w_s pero su paralelo no contiene w_t, etcétera. R_1 y R_2 son las sumas parciales por fila de dichas frecuencias, y C_1 y C_2 las por columna. El número total de pares considerados N constituye el *tamaño de la muestra*.

Construida la tabla, se calcularían medidas de asociación para cada par de palabras, integrando los más prometedores en el diccionario bilingüe.

2.2. Adaptaciones para Alinear a Nivel de *n*-Grama

En la segunda fase de nuestro algoritmo de alineamiento no partimos de párrafos alineados conteniendo palabras, sino de las palabras alineadas mediante GIZA++ durante la primera fase, las cuales contienen n-gramas de caracteres. Una primera opción sería simplemente adaptar la tabla de contingencia a este nuevo contexto, considerando que se están manejando pares de n-gramas (g_s, g_t) coocurriendo en palabras alineadas; en lugar de pares de palabras (w_s, w_t) coocurriendo en párrafos alineados. De esta forma O_{11}, por ejemplo, pasaría a ser el número de pares de palabras alineadas donde la palabra en el idioma origen contiene el n-grama g_s y la palabra en el idioma destino contiene g_t.

Esta solución parece lógica, pero no es adecuada, ya que en nuestro contexto las coocurrencias de n-gramas en palabras alineadas son *probables*, no ciertas como ocurría con los párrafos, puesto que GIZA++ calcula una *probabilidad* de traducción para cada par de palabras que coocurren en el corpus [5]. Consecuentemente una misma palabra puede ser alineada con diversas traducciones candidatas, cada una con una probabilidad diferente.

Tomemos como ejemplo el caso de las palabras en inglés `milk` y `milky`, y las españolas `leche`, `lechoso` y `tomate`. Un posible alineamiento a nivel de palabra, con sus correspondientes probabilidades y 4-gramas componente, podría ser:

término origen	**traducción candidata**	**prob.**
`milk` = {`-milk-`}	`leche` = {`-lech-`, `-eche-`}	0,98
`milky` = {`-milk-`, `-ilky-`}	`lechoso` = {`-lech-`, `-echo-`, `-chos-`, `-hoso-`}	0,92
`milk` = {`-milk-`}	`tomate` = {`-toma-`, `-omat-`, `-mate-`}	0,15

Según esto podría considerarse, estrictamente hablando, que el 4-grama origen `-milk-` no coocurre realmente con el 4-grama destino `-lech-`, puesto que los alineamientos entre las palabras que los contienen, `milk`–`leche` y `milky`–`lechoso`, no son seguros. Sin embargo, parece mucho más probable que `-milk-` se corresponda con `-lech-` y no con `-toma-`, pues la probabilidad del par que contiene simultáneamente `-milk-` y `-toma-`, `milk`–`tomate`, es mucho menor que la de los pares que contienen a `-milk-` y `-lech-`, `milk`–`leche` y `milky`–`lechoso`. Partiendo de esta idea, proponemos ponderar la probabilidad de coocurrencia de un par de n-gramas de acuerdo a la probabilidad con las que están alineadas las palabras de los contienen. De esta forma, las tablas de contingencia resultantes correspondientes a los pares de 4-gramas (`-milk-`, `-lech-`) y (`-milk-`, `-toma-`) serían:

	$T =$ `-lech-`	$T \neq$ `-lech-`		$T =$ `-toma-`	$T \neq$ `-toma-`	
$S =$ `-milk-`	O_{11} =**1,90**	O_{12} =**4,19**	R_1 =**6,09**	O_{11} =**0,15**	O_{12} =**5,94**	R_1 =**6,09**
$S \neq$ `-milk-`	O_{21} =**0,92**	O_{22} =**2,76**	R_2 =**3,68**	O_{21} =**0**	O_{22} =**3,68**	R_2 =**3,68**
	C_1 =**2,82**	C_2 =**6,95**	N =**9,77**	C_1 =**0,15**	C_2 =**9,62**	N =**9,77**

Podemos apreciar que, por ejemplo, la frecuencia O_{11} correspondiente a *(`milk`, `lech`)* no es 2 como hubiera cabido esperar inicialmente, sino 1,90, ya que el par aparece en dos alineamientos, `milk`–`leche` y `milky`–`lechoso`, ponderando dichas coocurrencias de acuerdo a la probabilidad de sus alineamientos:

$$O_{11} = 0{,}98 \text{ (para } \texttt{milk}\text{–}\texttt{leche}\text{)} + 0{,}92 \text{ (para } \texttt{milky}\text{–}\texttt{lechoso}\text{)} = \mathbf{1{,}90}$$

Construida la tabla, se pueden calcular las medidas de asociación estándar entre cada par de n-gramas: *coeficiente de Dice* (*Dice*), *información mutua* (*IM*) y *log-likelihood* (*LogL*), calculadas de la siguiente manera [8]:

$$Dice = \frac{2\,O_{11}}{R_1 + C_1} \qquad IM = log\frac{NO_{11}}{R_1 C_1} \qquad LogL = 2\sum_{i,j} O_{ij}\; log\frac{NO_{ij}}{R_i C_j}$$

Retomando el caso anterior nos encontraríamos que para el caso del coeficiente de Dice, por ejemplo, el valor de asociación para el par correcto (`-milk-`, `-lech-`) es mucho mayor que para el par incorrecto (`-milk-`, `-toma-`):

$$Dice(\texttt{-milk-,-lech-}) = \frac{2*1{,}90}{6{,}09+2{,}82} = \mathbf{0{,}43} \quad Dice(\texttt{-milk-,-toma-}) = \frac{2*0{,}15}{6{,}09+0{,}15} = \mathbf{0{,}05}$$

Adicionalmente, buscando minimizar el ruido introducido en el sistema y la ambigüedad resultante, desecharemos aquellos alineamientos a nivel de palabra con probabilidad menor que un umbral W=0,15. Esto nos permitirá reducir drásticamente (en 90 % o más) el número de pares de palabras a procesar.

Asimismo analizaremos el efecto de un segundo mecanismo de filtrado de objetivo similar: la aplicación de un alineamiento bidireccional a nivel de palabra [6], es decir, aceptar un alineamiento *idiomaOrigen*→*idiomaDestino* (w_s, w_t), donde w_s y w_t denotan respectivamente la palabra origen y su traducción candidata, sólo si existe también el correspondiente alineamiento *idiomaDestino*→*idiomaOrigen* (w_t, w_s). Esto nos permitirá centrarnos en aquellos pares de traducción menos ambiguos, reduciendo también el número de pares de palabras a procesar.

3. Experimentos

3.1. Diseño de los Experimentos

Nuestro estudio se ha centrado en el caso del español (ES) y el inglés (EN): consultas en inglés sobre textos en español (EN2ES) y consultas en español sobre textos en inglés (ES2EN). De acuerdo con [7], hemos calculado el porcentaje de cognatos en sus *listas de Swadesh* para así obtener una estimación del grado de similitud entre ambos idiomas, obteniendo un 19,30 % (40/207) de similitud.

Los corpus empleados son los de la *robust task* del CLEF 2006 [2]. El corpus inglés contiene 113.005 documentos (425 MB) del Los Angeles Times de 1994 y 56.472 documentos (154 MB) del Glasgow Herald de 1995; el español contiene 454.045 teletipos (1,06 GB) de la agencia EFE de 1994 y 1995. Los *topics* constan de tres campos: *título*, un breve título como su nombre indica; *descripción*, una somera frase de descripción; y *narrativa*, un pequeño texto especificando los criterios de relevancia. Siguiendo la política del CLEF, nuestros experimentos se han realizado empleando únicamente los campos *título* y *descripción*.

Por lo tanto ambos experimentos, EN2ES y ES2EN, emplean corpus diferentes, aunque comparables. Sin embargo, esto no constituye un problema a la hora de analizar el comportamiento del sistema, ya que si los resultados son cualitativamente similares constituirán un claro indicativo de las capacidades o deficiencias de la solución planteada. El conjunto de consultas tampoco es el mismo: en ambos casos empleamos 60 consultas, si bien sólo un tercio de ellas es común. Finalmente, tampoco el corpus paralelo empleado para los alineamientos es el mismo. Para EN2ES, se ha empleado el *release* v2 del Europarl [3], con unos 28 millones de palabras por idioma, mientras que ES2EN emplea el v6, que casi dobla el anterior.

El *proceso de indexación* es muy simple: el texto se pasa a minúscula y se eliminan los signos de puntuación, para ser luego tokenizado en 4-gramas [9] e indexado por el motor Terrier [11] con pesos InL2. En el caso del *proceso de consulta*, la entrada, en el idioma origen, se descompone primero en 4-gramas, para luego sustituir dichos 4-gramas por sus traducciones al idioma destino de acuerdo a uno de nuestros dos *algoritmos de selección*:

1. **Por rango**: que devuelve los N n-gramas traducción candidatos con las medidas de asociación más altas.
2. **Por umbral**: que devuelve los n-gramas traducción cuya medida de asociación es superior o igual a un umbral T dado.

Finalmente, la consulta traducida resultante es lanzada contra el motor.

3.2. Resultados con el Coeficiente de Dice

La Figura 1 recoge los resultados obtenidos aplicando un alineamiento a nivel de palabra convencional (unidireccional), tanto al emplear un algoritmo de selección por rango (gráficas superiores), como por umbral (gráficas inferiores).

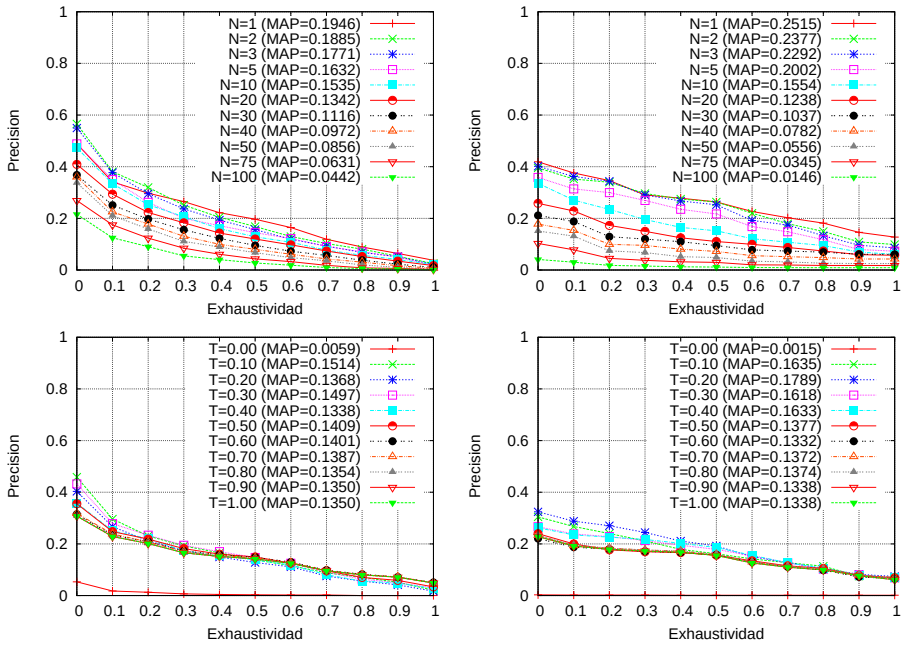

Figura 1. Resultados con coeficiente de Dice y alineamiento unidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

Las gráficas de la izquierda corresponden a las consultas EN2ES y las de la derecha a las inversas, ES2EN.

En el caso de la selección por rango los mejores resultados se obtienen al minimizar el número de traducciones candidatas (N=1). En la selección por umbral, al tomar el coeficiente de Dice valores en el rango [0..1], se han empleado unos umbrales $T \in \{0;\ 0,1;\ 0,2;\ \ldots\ 0,9;\ 1\}$.[3] El mejor rendimiento se obtiene para umbrales bajos (T=0,10 para EN2ES y T=0,20 para ES2EN), siendo los valores obtenidos, en general, más homogéneos. Sin embargo, el rendimiento es significativamente inferior al de por rango[4] debido al ruido generado por los n-gramas extra introducidos por el método basado en umbrales.

La aplicación de un alineamiento bidireccional durante la fase de alineamiento a nivel de palabra —véase Sección 2.2— conlleva una reducción del 50 % en el número de alineamientos de entrada a nivel de palabra y, consecuentemente, en el número de pares de traducciones candidatas generadas a nivel de n-grama (45 %–50 %). Los resultados obtenidos aplicando este mecanismo de filtrado se muestran en la Figura 2, siendo similares a los obtenidos en el caso del alineamiento convencional. En la selección por rango los valores obtenidos son

[3] Advertimos al lector que por problemas de espacio y legibilidad no siempre será posible representar en las gráficas todos los umbrales ensayados.

[4] Usamos en este trabajo tests-t bilaterales sobre las MAP con α=0,05.

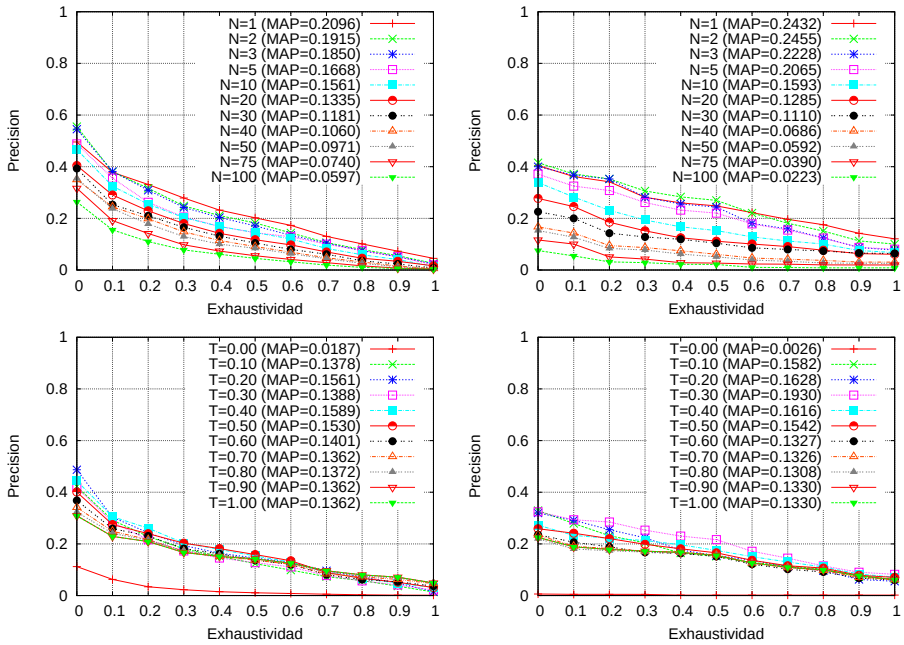

Figura 2. Resultados con coeficiente de Dice y alineamiento bidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

similares, con los mejores resultados para N reducidos (N=1 para EN2ES y N=1-2 para ES2EN). En la selección por umbral los resultados son de nuevo inferiores, sin diferencias significativas con los del alineamiento unidireccional, requiriendo umbrales más altos (T=0,40 para EN2ES y T=0,30 para ES2EN).

Estos datos demuestran que la selección por rango es significativamente superior y que el filtrado por alineamiento bidireccional no influye significativamente en el rendimiento del sistema, a la vez que reduce notablemente el consumo de recursos. Por otra parte, se hace patente que el sistema es robusto frente al ruido introducido por alineamientos ambiguos a nivel de palabra.

3.3. Resultados con Información Mutua

IM puede tomar cualquier valor en el rango $(-\infty..\infty)$, correspondiendo los valores negativos con pares de términos que se evitan mutuamente. De este modo, y para homogeneizar los test, en el caso de la selección por umbral, éstos se calcularán como $T_i = \mu + 0{,}5\, i\, \sigma$ donde T_i es el i-ésimo umbral (con $i \in \mathbb{Z}^+$), μ es la *media* de los valores de asociación obtenidos y σ es su *desviación típica*.

Los resultados con alineamiento unidireccional de la Figura 3 muestran un rendimiento significativamente inferior al de Dice. Para la selección por rango ha sido necesario incrementar el número de n-gramas traducción introducidos,

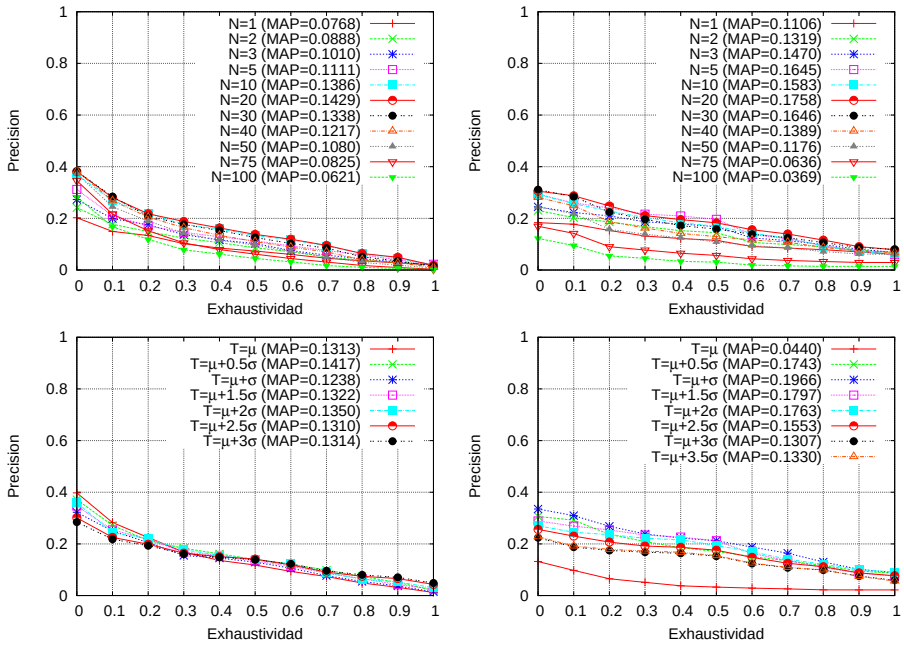

Figura 3. Resultados con IM y alineamiento unidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

hasta N=20 en ambos casos, para obtener un resultado óptimo. Esto se debe a que IM tiende a sobreestimar los datos de menor frecuencia, lo que provoca que alineamientos de n-gramas no válidos pero poco frecuentes tengan asignados valores de IM muy altos, sobrepasando los de pares correctos, introduciendo así ruido en la consulta durante la traducción. En lo que respecta a la selección por umbral, los resultados se mantienen bastante homogéneos entre umbrales, sin diferencias significativas con los de por rango, obteniendo los mejores resultados con T=μ+0,5σ para EN2ES y con T=μ+σ para ES2EN.

El comportamiento en el caso de aplicar alineamiento bidireccional (Figura 4) es similar, salvo una mejora no significativa en ES2EN cuando se emplea selección por rango, ya que la aplicación del alineamiento bidireccional permite reducir el número de pares de n-gramas incorrectos sobrevalorados. Para la selección por umbral los resultados son en general algo mejores que para el caso unidireccional, aunque el valor óptimo es similar. Sus umbrales correspondientes son también similares a los anteriores (T=μ para EN2ES y T=μ+1,5σ para ES2EN).

Queda patente, pues, que para IM no existen diferencias significativas entre algoritmos de selección, y que el alineamiento bidireccional tampoco influye significativamente en el rendimiento. Asimismo, el sistema se ha mostrado de nuevo robusto frente a la presencia de ruido.

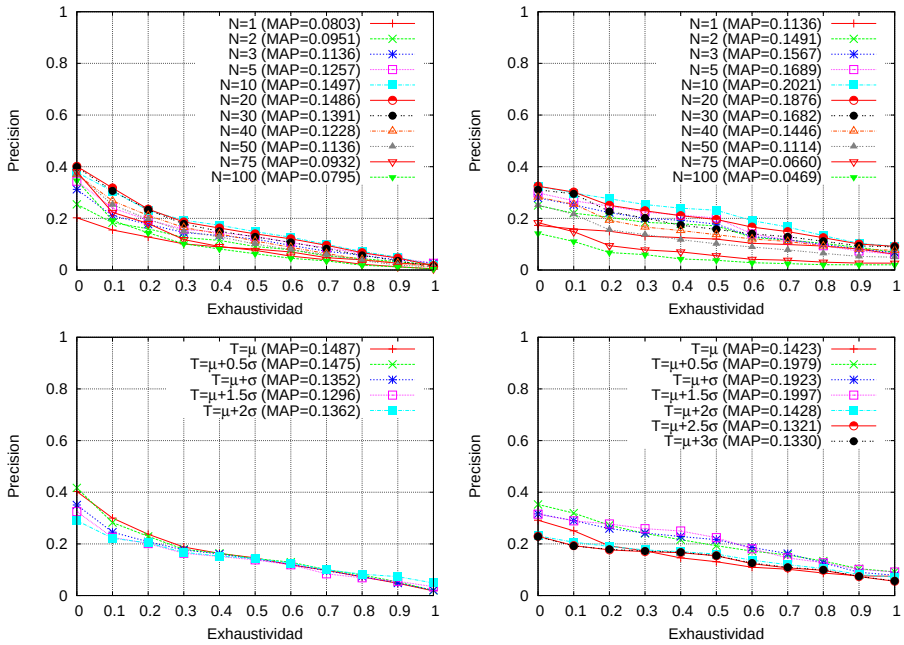

Figura 4. Resultados con IM y alineamiento bidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

3.4. Resultados para *Log-Likelihood*

Log-Likelihood puede tomar también cualquier valor en $(-\infty..\infty)$, por lo que en el caso de la selección por umbral continuaremos fijando dichos umbrales en base a la media y desviación típica. Sin embargo, tras haber estudiado la distribución de los alineamientos de n-gramas obtenidos hemos optado esta vez por una granularidad variable, con umbrales de la forma:

$$T_i = \begin{cases} \mu + 0{,}05\ i\ \sigma & -\infty < i \leq 2 \\ \mu + 0{,}50\ (i-2)\ \sigma & 2 < i < +\infty \end{cases}$$

donde T_i es el i-ésimo umbral (con $i \in \mathbb{Z}$), μ es la *media* de los valores de obtenidos y σ es su *desviación típica*.

Los resultados obtenidos para el alineamiento unidireccional se muestran en la Figura 5. En el caso de selección por rango, los mejores resultados son de nuevo para valores bajos de N, con N=1 como óptimo, superando el rendimiento con Dice. Por contra, los resultados obtenidos para la selección por umbral son significativamente pobres. Asimismo los resultados para el caso del alineamiento bidireccional (Figura 6) no muestran diferencias significativas, demostrándose nuevamente la validez del filtrado, así como la robustez de la solución.

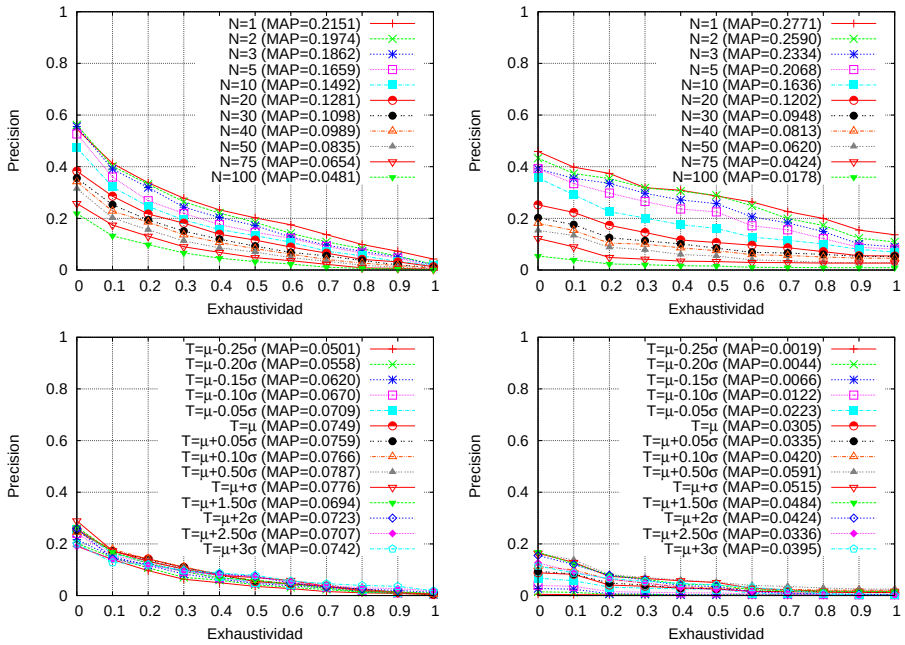

Figura 5. Resultados con *log-likelihood* y alineamiento unidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

3.5. Comparativa de Resultados

Para completar nuestro análisis, la Figura 7 recoge sendas comparativas de los mejores resultados absolutos obtenidos para cada medida de asociación en cada dirección de consulta con respecto a diversas líneas de base de interés:

- Una ejecución monolingüe en el idioma destino empleando 4-gramas como términos (`ES 4-grams` en EN2ES y `EN 4-grams` en ES2EN), la cual constituye nuestro rendimiento objetivo deseado. Es nuestra *cota superior*.
- Otra ejecución con 4-gramas, pero lanzando directamente las consultas sin traducción alguna, en el idioma origen, contra el índice del idioma destino (`EN 4-grams` en EN2ES y `ES 4-grams` en ES2EN), lo que nos permite medir el impacto de las correspondencias casuales. Es nuestra *cota inferior*.
- Otra ejecución monolingüe en el idioma destino, pero empleando *stemming* (`ES stm` en EN2ES y `EN stm` en ES2EN). Se usaron el *stemmer* Snowball,[5] basado en Porter, y la lista de *stopwords* de la Universidad de Neuchâtel.[6]

Como podemos apreciar, *log-likelihood* es superior, superando tanto a Dice como a IM (significativamente inferior). Los mejores resultados se han obtenido con selección por rango, que permite además un procesamiento más

[5] `http://snowball.tartarus.org`
[6] `http://www.unine.ch/info/clef/`

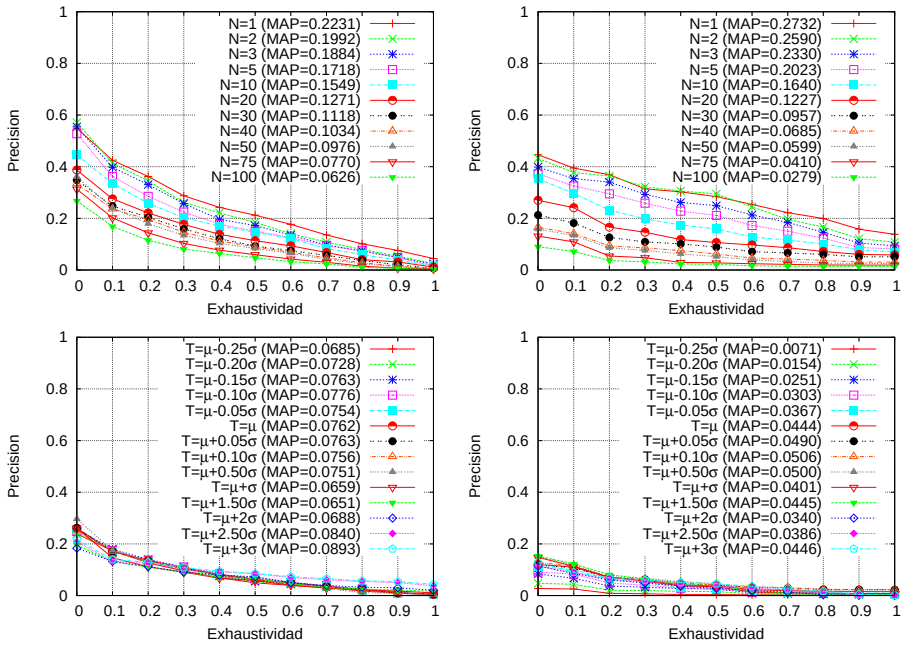

Figura 6. Resultados con *log-likelihood* y alineamiento bidireccional para selección por rango (sup.), por umbral (inf.), EN2ES (izq.) y ES2EN (dcha.)

sencillo y menos costoso. Nótese también que aunque hemos comparado los mejores resultados absolutos, siendo algunos con alineamiento unidireccional, el alineamiento bidireccional no afecta significativamente en modo alguno el rendimiento. Los resultados obtenidos han sido, pues, positivos y permiten confiar en la validez de la aplicabilidad de nuestra solución.

4. Conclusiones y trabajo futuro

Continuando con nuestras investigaciones sobre el empleo de n-gramas de caracteres como unidad de procesamiento en sistemas de Recuperación de Información Multilingüe, hemos presentado una comparativa paralela de su comportamiento en el caso de consultas en inglés sobre textos en español y viceversa. Los resultados positivos que hemos obtenidos nos permiten demostrar la validez de la solución independientemente de la dirección de la traducción, algo que no habíamos hecho hasta ahora.

En lo que respecta al trabajo futuro, abordaremos el problema del gran tamaño de los índices resultantes cuando se emplean n-gramas como términos de indexación, lo que nos permitiría reducir el consumo de recursos así como probablemente mejorar el rendimiento del sistema. Proponemos en este punto estudiar el empleo de técnicas de *pruning* así como extender el concepto

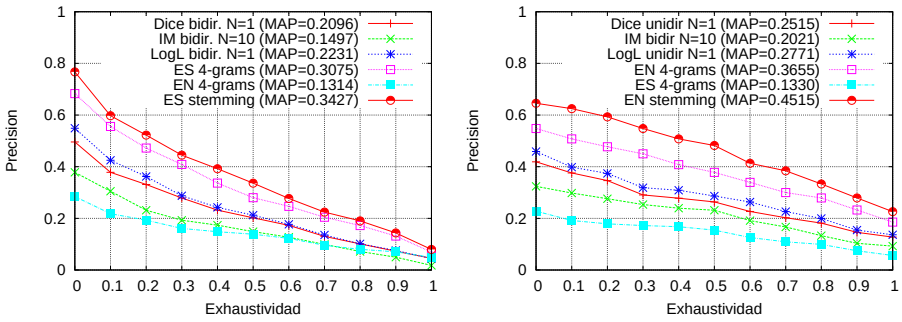

Figura 7. Resumen de resultados para EN2ES (izq.) y ES2EN (dcha.).

de *stopword* al caso de los n-gramas. Sin embargo, de cara a preservar la independencia de nuestra solución respecto a los idiomas y dominios de aplicación, dichos *stop-n-grams* deberían ser generados de forma automática [1].

Agradecimientos. Trabajo parcialmente subvencionado por el ESF Research Networking Programme (red *"Evaluating Information Access Systems – ELIAS"*), Ministerio de Ciencia e Innovación y FEDER (proyectos TIN2010-18552-C03-01 y TIN2010-18552-C03-02), Ministerio de Educación (programa de *Becas de Colaboración* en departamentos universitarios) y Xunta de Galicia (*"Rede Galega de Recursos Lingüísticos para unha Sociedade do Coñecemento"*).

Referencias

1. R. Blanco and A. Barreiro. Static pruning of terms in inverted files. In Vol. 4425 of *LNCS*, pp. 64–75. Springer-Verlag, 2007.
2. Cross-Language Evaluation Forum (CLEF). `http://www.clef-initiative.eu`
3. EUROPARL. `http://www.statmt.org/europarl/`
4. W. A. Gale and K. W. Church. Identifying word correspondence in parallel texts. In *Proc. of the Workshop on Speech and Natural Language*, pp. 152–157, 1991.
5. GIZA++. `http://code.google.com/p/giza-pp/`
6. P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. of NAACL'03*, pp. 48–54, 2003. ACL.
7. W. P. Lehmann. *Historical Linguistics*, Ch. 9. Taylor & Francis, 1992.
8. C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
9. P. McNamee and J. Mayfield. JHU/APL experiments in tokenization and non-word translation. Vol. 3237 of *LNCS*, pp. 85–97. Springer-Verlag, 2004.
10. A. M. Robertson and P. Willett. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48–69, January 1998.
11. TERRIER. `http://www.terrier.org`
12. J. Vilares, M. Vilares, and J. Otero. Managing Misspelled Queries in IR Applications. *Information Processing & Management*, 47(2):263–286, 2011.