Alineamiento de Entidades con Nombre usando distancia léxica[1]

Rafael Borrego Ropero
Víctor Jesús Díaz Madrigal
Dpto. Lenguajes y Sistemas Informáticos (Universidad de Sevilla)
{rborrego, vjdiaz }@lsi.us.es

Resumen: En este artículo se presenta un sistema para asistir en el etiquetado de entidades con nombre y palabras semánticamente relevantes en corpora bilingüe. Además de las heurísticas usadas habitualmente para el alineamiento de corpora basadas en la frecuencia y la posición, se ha aprovechado el alto parecido lexicográfico entre palabras pertenecientes a lenguajes cercanos, como el español y el francés, para usar el algoritmo Edit Distance. Tras estudiar los resultados de estas técnicas, se ha realizado un sistema de votación ponderada que, a partir de una secuencia de palabras en un lenguaje, propone una secuencia equivalente en el otro idioma. Al ser técnicas que no se basan en un conocimiento previo de los lenguajes, pueden ser adaptadas para etiquetar dos lenguajes cualesquiera que tengan un tronco común.
Palabras clave: Alineamiento, Etiquetado de entidades, Edit distance, Corpora bilingüe

Abstract: In this article a system is presented to help in labelling named entities and meaningful words in bilingual parallel corpora. Besides the strategies used frequently in the alignment of corpora based on word frequency and position, it has been taken advantage of the high lexicographical similarity among words in nearby languages as Spanish and French to use the algorithm Edit Distance. After studying the results of these techniques, a system of voting has been defined that given a set of words in a language, an equivalent set in another language is proposed. As these techniques are not based on a previous knowledge of the languages, they can be used to label any pair of languages that share a common branch.
Keywords: Alignment, Name Entity Recognition, Edit distance, Bilingual corpora

1 Introducción

Dado un corpora bilingüe, el alineamiento de textos [3] consiste en identificar qué partes - por ejemplo: párrafos, frases o palabras - de un corpus se corresponden con las del otro. En la literatura se han propuesto diversas técnicas para el alineamiento de corpora ya que es una tarea de gran utilidad en la asistencia a la traducción. En este caso, se propone usar heurísticas aplicadas en el alineamiento de corpora para relacionar entidades con nombre. El reconocimiento de entidades con nombre [5] (NER, Named Entity Recognition) es una tarea de especial relevancia para la compresión de un texto. El objetivo que persigue un sistema NER es la identificación y clasificación en un texto de todos aquellos sintagmas que hacen referencia a nombres de personas, lugares y organizaciones. La tarea NER es crucial en la resolución de otros problemas más complejos como la extracción de información.

En este artículo se mostrará cómo se pueden mejorar los resultados en el alineamiento de entidades con nombre y de palabras relevantes en un determinado dominio aprovechando el gran parecido que hay entre palabras de varios idiomas con un tronco común. El sistema ha sido probado sobre un corpora bilingüe con varios centenares de frases en español y francés. Los documentos originales son noticias en francés de temas económicos que han sido traducidas después al español por un experto.

El sistema sigue un proceso similar al que haría una persona para comparar dos textos en distintos idiomas. Primero se procede a una segmentación del texto en párrafos, frases y palabras apoyándose en la información aportada por distintos signos de puntuación. Después son aplicadas heurísticas para determinar la relación

[1] Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia (TIN 2004-07246-C03-03)

 © Sociedad Española para el Procesamiento del Lenguaje Natural

existente entre elementos lingüísticos. Trabajos anteriores [1] [2] demuestran el éxito obtenido aplicando técnicas basadas en la posición en la que figura cada palabra junto con su frecuencia de aparición. El sistema propuesto se apoya además en la similitud de palabras medida en función del resultado de aplicar el algoritmo Edit Distance [4]. Aunque en un caso general la similitud de palabras no es un factor esencial, ésta si puede ser de gran utilidad cuando tratamos con corpora en dominios muy específicos o cuando tratamos de establecer relaciones entre entidades con nombre. La razón de esta mejora se debe a que dicha terminología tiene en numerosas ocasiones un mismo origen lingüístico o histórico.

Dado que no se han usado conocimientos previos del corpora ni de ninguno de los dos lenguajes, esta técnica es susceptible de ser aplicada a cualquier colección de documentos español-francés o a otros corpora bilingüe con similares características.

2 El córpora utilizado

Para los experimentos se ha usado un corpora bilingüe español-francés obtenido a partir de varios artículos del diario francés *Le Monde* y *Le Monde Economie*. Los artículos fueron re-etiquetados con XML mostrando primero la introducción y a continuación los párrafos en los dos idiomas.

Tras eso, se procedió a dividir el corpora en frases y palabras, indexando sus posiciones de inicio y fin en los párrafos. La división en frases se hizo teniendo en cuenta la posición de los caracteres '.'. La división de las palabras se hizo a partir de la división anterior teniendo en cuenta los espacios en blanco, signos de interrogación, de exclamación, etc. Aunque existen casos en los que esta forma de proceder genera errores en la interpretación de palabras, estos errores fueron ignorados debido a que su número era muy inferior en comparación con el tamaño del corpora.

Tras analizar estos ficheros y ver que todos tenían una estructura común, se comprobó que todos los párrafos del corpora guardaban una correspondencia 1:1 (cada párrafo del corpus en francés se correspondía con un único párrafo en el corpus en español), lo cual es una gran ventaja ya que evitó que se tuviera que usar algoritmos para relacionar párrafos.

Posteriormente, se filtraron los corpus y se hizo una comparación del número de frases de cada fichero para ver si seguían una correspondencia 1:1. Este tipo de proporción nos indica cómo de fiel ha sido la traducción y cómo de cercana es la posición de cada palabra a la de su equivalente en el otro idioma. Se comprobó que de los 23 archivos iniciales, sólo 18 contenían una correspondencia 1:1 en cuanto a frases, por lo que se descartaron los otros cinco. Tras ese primer filtrado, se dividió el corpora en 2 conjuntos de ficheros: uno de entrenamiento, que incluía el 90% de las frases (362 en cada idioma) , y otro de test con el 10% restante (51 frases). Esta división se hizo tanto para hacer las pruebas finales (test) como para ajustar los parámetros del sistema (entrenamiento). Las tablas 1 y 2 muestran el número de ficheros, párrafos y frases para cada uno de los corpus.

	Ficheros	Párrafos	Frases
Corpora	18	134	419
Training	16	121	377
Test	2	13	42

Tabla 1. Datos del corpus en español

	Ficheros	Párrafos	Frases
Corpora	18	134	419
Training	16	121	377
Test	2	13	42

Tabla 2. Datos del corpus en francés

Una vez dividido el corpora, se calculó la frecuencia de cada palabra en cada corpora (ver tabla 3 y 4).

	Palabras	Palabras únicas
Corpora	9425	2509
Training	8488	2346
Test	937	443

Tabla 3. Palabras en el corpus en español

	Palabras	Palabras únicas
Corpora	9389	2496
Training	8440	2328
Test	949	455

Tabla 4. Palabras en el corpus en francés

Además, se hizo un estudio para ver si las frases contenían un número de palabras similar en ambos idiomas. Se vio que las frases del corpus en español contenían de media sólo 0.089 palabras más que las frases del corpus en

francés. En aquellas frases del corpus en español que tenían más palabras que la correspondiente en el corpus en francés, había 0.89 palabras más de media. En aquellas frases del corpus en francés que tenían más palabras que la correspondiente en el corpus en español había 0.80 palabras más de media. Por tanto, el tamaño de las frases era muy parecido.

3 Heurísticas

A continuación, se expondrán las heurísticas utilizadas para buscar la traducción de una palabra en el otro idioma. Primero se indicará la basada en Edit Distance, a continuación la basada en análisis de frecuencia, y finalmente la técnica basada en la posición. Tras ello, se comentarán los resultados iniciales de probar éstas heurísticas, para finalmente mostrar el sistema de votación utilizado que combina dichas tres técnicas.

3.1 Basada en cercanía léxica

El algoritmo *Edit Distance* (también conocido como *Levenshtein distance*) permite averiguar el mínimo coste para transformar una cadena en otra, ya sea añadiendo o eliminando caracteres, o cambiando uno por otro. Esta medida servirá de apoyo para determinar el parecido lexicográfico de pares de palabras. Por ejemplo, para transformar la palabra "semana" en "semaine" se necesita eliminar el carácter 'i' de "semaine", y cambiar el último carácter , "e" por "a".

El algoritmo *Edit Distance* establece una distancia o una métrica de similitud entre palabras. El valor final de esta distancia varía en función del coste asociado a cada operación de edición (añadir, borrar, cambiar). En este caso se ha aplicado un mismo coste a todas las operaciones de edición.

La forma de empleo de esta heurística es la siguiente: dada una palabra en uno de los corpus, el sistema aplica el algoritmo *Edit Distance* a todas las palabras del párrafo equivalente en el otro corpus. El resultado final está constituido por aquellas palabras con mayor similitud, es decir, con un valor de distancia menor.

Esta medida da muy buenos resultados para buscar entidades como lugares, nombres de personas o empresas, u organismos, ya que muchas de ellas son traducidas de forma muy similar o con pocas variaciones. Además, hay numerosas palabras como "semana" ("semaine" en francés) que presentan un alto parecido, por lo que la distancia entre una palabra y su equivalente es muy baja. No obstante, da problemas al traducir palabras que no guardan correspondencia en el otro idioma como "Estados Unidos" ("Etat-Units" en francés) ya que intenta una primera comparación de "Estados" con "Etat-Units", y posteriormente "Unidos" con "Etat-Units", fallando para ambas.

3.2 Basada en análisis de frecuencia

Para realizar esta heurística, se debe calcular para cada palabra *w* el número de veces que aparece *w* en todo el corpora y en cada fichero. Si la traducción tiene un alto parecido con el original, el número de veces que aparece una palabra en ambos idiomas es muy similar. Por ello, cuando se selecciona una palabra en un corpus, se buscan todas aquellas del párrafo correspondiente en el otro corpus que tienen una frecuencia de aparición parecida. Esta heurística da buenos resultados en palabras con un número medio de repeticiones. En palabras con frecuencia baja la heurística devuelve demasiadas palabras candidatas. En el caso de palabras con frecuencia alta aparece el fenómeno de las palabras huecas, como pueden ser los determinantes o preposiciones, cuya traducción da lugar a demasiadas soluciones. Con objeto de mejorar los resultados, esta heurística es sólo aplicable en el caso de palabras cuya frecuencia no está situada en estos dos casos extremos.

3.3 Basada en análisis de posición

En la basada en la posición, tras seleccionar una palabra en un corpus, se busca su posición respecto al párrafo al que pertenece. Tras eso, se busca en el párrafo[2] equivalente la palabra que ocupa su posición. A esta palabra se le añaden las de su entorno (la anterior y la siguiente). Esta medida da buenos resultados cuando existe un gran parecido entre las estructuras sintácticas de ambos idiomas y cuando la traducción ha sido relativamente fiel. Otro factor que influye en la bondad de esta heurística es que el tamaño de las frases en los

[2] Los resultados de esta heurística aplicando la búsqueda sobre la frase equivalente en lugar del párrafo arrojó peores resultados, por lo que se optó finalmente buscar en el párrafo.

corpora sea proporcional (como sucede en este caso, según se dijo anteriormente).

3.4 Sistema de votación

Cada una de estas heurísticas devuelve un conjunto de palabras candidatas. Para mejorar los resultados individuales de cada heurística se definió un sistema de votación ponderada. A una misma palabra se le aplican todas las heurísticas de forma que se genera un conjunto (con posibles repeticiones) de palabras candidatas. Finalmente, la palabra escogida es aquella con mayor multiplicidad (que ha sido seleccionada por más heurísticas). La votación puede ser simple o ponderada dependiendo de que el peso del voto de cada heurística sea igual o no. Con un sistema de votación se intenta escoger aquella palabra que más se ajusta de forma global a las tres características propuestas: cercanía léxica, posición y frecuencia. La votación ponderada permite dar mayor realce a aquellas heurísticas cuyo comportamiento sea más fiable.

Para ajustar el valor de los parámetros de ponderación y para escoger la mejor heurística para cada heurística se hizo un análisis previo aplicando las distintas heurísticas sobre una palabra aleatoria de cada párrafo del conjunto de entrenamiento. A partir de los resultados obtenidos se obtuvieron los factores de ponderación para decidir qué peso se le daba en la votación a las palabras devueltas por cada una de las tres técnicas. Estos factores se obtuvieron calculando el porcentaje de aciertos de cada una.

3.5 Resultados de las técnicas

Se escogieron 60 palabras repartidas de forma aleatoria y uniforme entre todos los párrafos del corpus. Se calcularon las traducciones de cada palabra para cada una de las tres heurísticas y para el sistema de votación, y se comprobó si las traducciones devueltas eran correctas o no.

Los resultados obtenidos por las heurísticas junto con el sistema de votación pueden verse en las tablas 5 y 6. En ellas se ve, para cada técnica, el número total de palabras escogidas, cuántas de ellas fueron aciertos, y el porcentaje de aciertos. La diferencia entre el número de palabras total se debe a que mientras que la votación escoge como mucho un candidato, las otras técnicas pueden devolver varias o incluso ninguna candidata .

Se observa claramente como la heurística que usa el algoritmo *Edit Distance* es la que da mejores resultados, y que el sistema de votación ponderada eleva los resultados del 30% al 70%. Sin embargo, estos resultados deben ser tomados con cautela, y por tanto, no deberían ser generalizados, ya que se ven notablemente influidos por dos razones: la fidelidad en la traducción y que los artículos contienen un conjunto importante de palabras similares relacionadas con el dominio económico.

	Edit distance	Posición	Frecuencia	Votación
Total	96	99	21	60
Aciertos	27	18	3	42
% aciertos	28,13%	18,18 %	14,29 %	70 %

Tabla 5. Resultados heurísticas español a francés

	Edit distance	Posición	Frecuencia	Votación
Total	98	99	80	62
Aciertos	30	17	9	44
% aciertos	30,61%	17,17%	11,25 %	70,97 %

Tabla 6. Resultados heurísticas francés a español

4 Búsqueda de entidades

El sistema de votación anterior permite el alineamiento de palabras. Este sistema puede ser extendido para detectar grupos de palabras que representan entidades como nombres compuestos por varias palabras (como “Pascal Blanqué”), o fechas (como “semaine du 8 décembre”). El objetivo es alinear entidades no triviales donde el orden de las palabras cambie ligeramente al traducirlas, o que contengan un número diferente de palabras, como “mercados de acciones”, que se traduce como “marchés actions”, que tiene una palabra menos.

Dada una secuencia consecutiva de palabras $T=w_1 \ldots w_n$ de un párrafo en uno de los corpus, el método para buscar su alineamiento es el siguiente: sea c_i la palabra ofertada por el sistema de votación para traducir la palabra w_i. De aquí se sigue que la traducción propuesta es un conjunto C de palabras $\{c_1,\ldots,c_n\}$. Ahora bien, ha de tenerse en cuenta que ese conjunto no tiene porqué estar ordenado por posición. Es decir, si $i<j$, la palabra c_i no tiene porqué estar antes que la palabra c_j en el corpus paralelo.

Para obtener un segmento alineado contiguo se toman, de todas las candidatas, la que aparece primera y la que aparece la última respecto a la posición en el texto. El segmento propuesto es el intervalo de todas las palabras contenidas entre ambas dos. De esta forma, si el usuario selecciona un conjunto de palabras que se traduzca como "BNP Paribas Asset Management" , el sistema elegiría "BNP" y "Management", y a partir de ahí obtendría el resto de palabras que están entre esas dos.

Dado que el algoritmo Edit Distance devuelve todas las palabras que se parecen, y que la basada en la frecuencia en ocasiones devuelve muchas palabras candidatas con frecuencia parecida, en numerosos casos se encuentra una palabra cuyo contenido coincide con el buscado, pero en otra posición distinta del párrafo (por ejemplo, la palabra "de" aparece varias veces en muchos párrafos). Para eliminar este efecto, se hace un análisis de las palabras más frecuentes sin contenido semántico (palabras huecas) de forma que al buscar las palabras del intervalo se ignoran.

5 Resultados

Para ayudar a la anotación de entidades se ha desarrollado un sistema que permite al usuario cargar y visualizar los ficheros del corpora. Según se van cargando, el sistema se encarga de indexar los elementos necesarios y generar las estadísticas. Una vez cargados todos, se muestra un árbol del directorio, y al ir pulsando en los distintos elementos se muestra una ventana en la que se puede ver simultáneamente el texto en ambos idiomas. Además, según se va seleccionando el texto en un idioma, el sistema colorea la frase equivalente en el otro idioma, baja el scroll para que muestre el texto coloreado, y se ofrece la traducción del texto seleccionado. Además, en cualquier momento el usuario puede ver estadísticas sobre el corpora y sobre cada fichero (número de frases, palabras, frecuencias...).

En el caso de las pruebas para buscar entidades, se tuvo con la dificultad de que el corpus no estaba previamente etiquetado. Por este motivo, se ha trabajado con un conjunto de entidades y palabras que han parecido relevantes del conjunto de ficheros de test. Las entidades incluyen nombres de personas, empresas, lugares, y otras relacionadas con economía (materia sobre la que versaban todos los archivos del corpora). En el caso de pasar de español a francés, se seleccionaron 147, y hubo 110 aciertos, por lo que el porcentaje de aciertos es 74,83 %. En el caso de pasar de francés a español, se seleccionaron 132, y hubo 98 aciertos, por lo que el porcentaje de aciertos es 74,24 %. Por tanto, en ambos casos se ha obtenido un porcentaje de éxito superior al 74 %.

	Fichero 1	Fichero 2	Total
Total	80	67	147
Aciertos	56	54	110
Porcentaje aciertos	70 %	80,56 %	74,83 %

Tabla 7 Resultados entidades español a francés

	Fichero 1	Fichero 2	Total
Total	68	64	132
Aciertos	49	49	98
Porcentaje aciertos	72,05 %	76,56 %	74,24 %

Tabla 8 Resultados entidades francés a español

6 Conclusiones

Este documento ha presentado una serie de técnicas para realizar alineamiento de palabras en corpora bilingüe mostrando como puede mejorarse los resultados utilizando un algoritmo de comparación de palabras en corpus de idiomas cercanos, respecto a otros que sólo realizan análisis de frecuencia o por posición. Tras ello, se han mostrado los resultados de aplicar con éxito estas técnicas a un corpora bilingüe en español y francés, obteniendo un 70% de éxito en la traducción por palabras, y un 74% en la traducción de entidades. Finalmente, se ha indicado como el sistema desarrollado permite encontrar no solo aquellas palabras equivalentes en el otro idioma, sino que además permite localizar entidades y palabras relevantes dentro de un contexto. Gracias a esto, constituye una útil herramienta para etiquetar corpora bilingües, ya que facilita al usuario la búsqueda de traducciones, y la visualización del contenido de dicho corpora.

7 Líneas de trabajo futuro

Existen numerosas líneas de trabajo relacionadas con las técnicas presentadas en este artículo.

En primer lugar, es necesario un estudio de los resultados con otros corpora, ya que los resultados actuales pueden estar influidos: por el tamaño del corpora (el utilizado en las pruebas es claramente insuficiente), por el contexto (en este caso sólo se han usado artículos de economía), y por el idioma (en este caso se han contemplado dos idiomas cercanos, por lo que los resultados podrían cambiar al comparar otros idiomas como español con inglés).

Por otro lado, sería interesante comprobar qué influencia tendría en los resultados la incorporación de otras técnicas probabilísticas más complejas como modelos de Máxima Entropía o Máquinas de Soporte Vectorial (SVM).

8 Bibliografía

P.F. Brown , J.C. Lai y R.L. Mercer. (1991) **Aligning sentences in parallel corpora**, Proceedings of the 29th annual meeting on Association for Computational Linguistics. pp 169-176

W.A. Gale y K.W. Church. (1993) **A program for aligning sentences in bilingual corpora**. Computational Linguistics archive.Volume 19 , Issue 1 .Special issue on using large corpora. Páginas 75 – 102

C.D. Manning y H. Schuetze. (2001). **Statistical Alignment and Machine Translation**. En *Foundaditons of Statistical Natural Language Processsing*, pp 462-494, MIT Press: Cambridge.

W. J. Masek y M. S. Paterson (1980). A space-economical computing string edit distance. Journal of Computer Science, 20. pp 262—272.

E.F. Tjong (2002): **Introduction to the CoNLL'02 Shared Task: Language-Independent Named Entity Recognition.** Proc. Of CoNLL-2002}. Taiwan pp 155--158