Uso de información morfológica en el alineamiento español-euskera

E. Agirre, A. Díaz de Ilarraza, G. Labaka, K. Sarasola
Euskal Herriko Univertsitatea/Universidad del Pais Vasco
649 pk - 20080 Donostia
e.agirre@ehu.es, jipdisaa@ehu.es, jiblaing@ehu.es, jipsagak@ehu.es

Resumen: En este artículo presentamos un primer estudio para el alineamiento de un corpus español-euskera mediante un alineador token-a-token en el que se consideran diferentes opciones de preprocesamiento morfológico. Usando GIZA++ conseguimos una reducción del error (Alignment Error Rate) de un 12.48 % respecto el baseline (carente de preproceso alguno), llegando al 23.76 %. Este resultado es comparable al obtenido para otros idiomas aglutinantes como el euskera.
Palabras clave: Traducción automática, alineamiento, idiomas aglutinantes

Abstract: In this paper we present a preliminary study for the alignment of a Spanish-Basque parallel corpus using a token-based aligner (GIZA++). We have studied several morphological pre-processing alternatives, and achieved 23.76 % Alignment Error Rate, with a reduction of 12.48 % over the baseline (no pre-processing). The results are comparable to those obtained for others agglutinative languages.
Keywords: Machine Translation, alignment, agglutinative languages

1. Introducción

En el área de la traducción automática, los sistemas estadísticos basados en corpus alineados están obteniendo buenos resultados en todas las competiciones. Además, el hecho de que gran parte del software y los recursos necesarios para entrenar y aplicar estos modelos estadísticos sean libres, hacen que su popularidad vaya en aumento. Estos sistemas estadísticos suelen tomar como entrada un corpus bilingüe donde las sentencias correspondientes a cada idioma están alineadas. El primer paso que llevan a cabo estos sistemas suele ser la alineación palabra por palabra de los testos.

Los grupos que han desarrollado estos sistemas se han centrado en idiomas con un grado de flexión bajo o medio (p.ej. el inglés, el francés, el español o el chino) para desarrollarlos, por lo que la metodología suele tomar palabras flexionadas como unidad básica.

En el caso de idiomas altamente flexionados (como es el euskera) la metodología de tomar la palabra flexionada como unidad básica provoca una muy alta dispersión de los datos, siendo este un escollo insalvable para la traducción estadística. Por ejemplo, la frase "en la casa" en euskera se traduce en una sola palabra "etxean".

Aunque relativamente menos flexionado, el español también presenta problemas en sus verbos, donde la concordancia compleja hace que la misma raíz verbal tenga muchas formas (p.ej. "venimos" que en inglés seria "we come").

Un estudio reciente (Koehn, 2005) da resultados de hasta 30.1 BLEU para pares como el inglés-español, pero de solo 17.6 para el inglés-alemán. A pesar de que en este estudio se han utilizado corpora equivalente para los distintos pares de idiomas, los resultados varían ostensiblemente. Esto muestra que el nivel de flexión de una lengua repercute en la calidad de los sistemas estadísticos.

En este artículo presentamos un estudio de las diferentes opciones para el alineamiento palabra a palabra de un corpus español-euskera. Específicamente, nos centraremos en las diferentes posibilidades de preprocesamiento morfológico (especialmente para el euskera, pero también para el español) con vistas a salvar la dispersión de datos y realizar un alineamiento óptimo. La importancia del alineamiento como tarea separada, viene avalada por las dos últimas competiciones públicas de alineamiento (Martin, Mihalcea, y Pedersen, 2005) y (Mihalcea y Pedersen, 2003).

Hemos realizado un exhaustivo estudio de diferentes posibilidades, que hemos evaluado sobre un corpus manualmente alineado de 100 sentencias. Para llevar a cabo estos experimentos utilizaremos herramientas que están disponibles libremente, tales como GIZA++

(Och y H. Ney, 2003) para el alineamiento, AlignmentSet[1] para la evaluación del alineamiento, HandAlign[2] para alinear manualmente el corpus de referencia, Freeling para analisis del español (Carreras et al., 2004), y Eustagger (Aduriz et al., 1994) para el del euskera (este ultimo disponible bajo licencia).

La siguiente sección presenta el trabajo relacionado. Seguiremos explicando las distintas técnicas que se han utilizado para preprocesar el euskera en la sección 3, y las utilizadas para preprocesar el español en la sección 4. En la sección 5 se presenta el diseño del experimento, para terminar comentando los resultados obtenidos en la sección 6 y las conclusiones y el trabajo futuro en la sección 7.

2. Trabajo relacionado

Ha habido distintos experimentos que utilizaban el preprocesamiento morfológico con el fin de mejorar el alineamiento de palabras de lenguas altamente flexionados. Así, podemos encontrar publicaciones que presentan distintos métodos probados en distintos pares de lenguas.

Para el par alemán-inglés hay varios trabajos que usan la información morfológica para preprocesar los textos mejorando así la alineación. En (Nießen y H. Ney, 2000) usan la información morfosintáctica para armonizar el orden de las palabras en los dos idiomas. Posteriormente (Nießen y H. Ney, 2004) etiquetaron algunas expresiones especialmente ambiguas con su categoría sintáctica, concretamente, combinaron algunas expresiones multi-palabra en un único token y dividieron ciertas palabras sustituyéndolas por el lema y etiquetas morfológicas. Mediante estas técnicas consiguieron mejorar sustancialmente la calidad del alineado.

En el caso del par inglés-checo (Goldwater y McClosky, 2005) preprocesaban el checo de cuatro modos diferentes; tokenizando tan solo, lematizando, creando etiquetas que se tratarán como tokens diferenciados y creando etiquetas que se unen al lema dando lugar a lo que podríamos identificar como unos lemas modificados. En el artículo mencionado, además del preproceso del texto de entrada, modificaban el algoritmo de alineación por lo que no se puede saber exactamente qué parte de la mejora que consiguen se debe al preproceso y qué parte se debe a la modificación del algoritmo de alineación.

No podemos dejar de mencionar dos publicaciones que al igual que nosotros, alinean el euskera con el español. (Caseli, M. Nunes, y Forcada, 2005) define un nuevo método de alineamiento que pretende superar los problemas que las técnicas más utilizadas tienen con las unidades multipalabra y con las diferencias en el orden de las palabras de los distintos idiomas a alinear. En esta publicación no se hace ningún tipo de preprocesamiento del texto y aunque para el par de lenguas español-portugués no consiguen superar los resultados de GIZA++, en el caso del español-euskera, donde los problemas mencionados tienen más importancia, logran mejorar los resultados conseguidos con GIZA++.

En el caso de (Nevado, Casacuberta, y Landa, 2004) se basan en alineamiento a nivel de palabra conseguido por GIZA++ para lograr segmentos más pequeños que una frase para utilizar en memorias de traducción. Este artículo no trata de mejorar el alineamiento en sí.

3. Preprocesamiento del euskera

El preproceso del euskera incluye tokenización, lematización y segmentación.

3.1. Tokenización

Es el prepocesamiento mínimo que necesita GIZA++ ya que éste tokeniza el texto usando tan sólo los espacios en blanco. Es necesario separar los signos de puntuación de las palabras precedentes, y, aunque cabe la posibilidad de unir los términos multipalabra para que GIZA++ los trate como un sólo token, no hemos utilizado esta opción.

Este es un ejemplo de una frase tokenizada, donde los tokens están separados con blancos:

Edukiaren egiturari dagokionez , funtsezko aldaketaren bat egin da

3.2. Lematización

Para procesar el euskera hemos usado Eustagger (Aduriz et al., 1994). Siguiendo con el ejemplo anterior, ésta sería la sentencia lematizada:

eduki egitura egon , funts aldaketa bat egin izan

[1] http://www.lsi.upc.es/lambert/software/AlignmentSet.html
[2] http://www.isi.edu/hdaume/HandAlign/

3.3. Segmentación

Al lematizar el euskera se pierde la información que dan los sufijos que se añaden al lema. Información como el caso gramatical, el número o información de subordinación. Esta información además se debería alinear con algunos elementos del español como las preposiciones, los artículos o las conjunciones. Por lo cual, además de lematizar el texto, hemos pretendido hacer explícitos estos rasgos mediante etiquetas que GIZA++ tratará como tokens diferentes. Nuestra intención es que GIZA++ alinee estas etiquetas que hemos añadido con los elementos del español anteriormente mencionados.

Hemos probado diferentes conjuntos de rasgos y combinación de ellos en etiquetas para ver cúales son las más adecuadas para este par de lenguas. A continuación presentamos los distintos rasgos y combinaciones en etiquetas que hemos utilizado y de qué modo las hemos combinado.

Número (NUM): Esta etiqueta representa tanto la información de definición (definido o indefinido) como el número en el caso de ser definido. Sus valores posibles son MG (indefinido), M,S (definido singular), M,P (definido plural) y M,PH (definido plural cercano). Creemos que esta etiqueta será alineada con el artículo del español. Este es un ejemplo de codificación según esta etiqueta del ejemplo anterior:

eduki M,S egitura M,S egon , funts MG, aldaketa MG, bat MG, egin izan

Caso (CAS): Esta etiqueta representa el caso, en euskera se diferencian 18 casos. Estos casos creemos que se alinearán con las preposiciones del español. Aunque algunas veces como con los casos gramaticales (ABS y ERG) que son los que toman tanto el sujeto como el objeto la etiqueta creada para el euskera no tendrá ninguna correspondencia en español. En este ejemplo se muestra la misma frase que en ejemplos anteriores codificado mediante esta etiqueta.

eduki GEN egitura DAT egon , funts ABS aldaketa GEN bat ABS egin izan

Declinación (DEC): Mediante esta etiqueta hemos unido las dos anteriores, ya que al empezar con los experimentos nos dimos cuenta que cuando por parte del español se utilizaba el texto lematizado la etiqueta que representa el número no se alineaba con los artículos como pensábamos y se alineaba mal o no se alineaba con nada. Por lo que pensamos en unir las dos etiquetas. En el siguiente ejemplo mostramos la misma frase con este método de segmentacion.

eduki GEN,M,S egitura DAT,M,S egon , funts ABS,MG, aldaketa GEN,MG, bat ABS,MG, egin izan

subordinación (SUB): En euskera la subordinación de las oraciones se representa mediante un sufijo que se añade al verbo subordinado. Hay distintos tipos de subordinaciones que según nuestro planteamiento se deberían alinear con las conjunciones o los pronombres relativos del español. La misma frase de ejemplos anteriores con una nueva segmentacion.

eduki egitura egon KAUS , funts aldaketa bat egin izan

Aspecto (ASP): Esta etiqueta representa el aspecto del verbo, es decir, si el verbo es perfecto o inperfecto. Mediante esta etiqueta y el procesamiento que hemos hecho con el español pretendemos dar el mismo formato a las cadenas verbales de las dos lenguas. La misma frase como ejemplo de este nuevo método de segmentación.

eduki egitura egon , funts aldaketa bat egin BURU izan

Modo y tiempo verbal (MOD): Esta etiqueta representa el tiempo verbal. El analizador de euskera da 16 valores distintos para este rasgo. Al igual que con el aspecto verbal, con esta etiqueta pretendemos igualar el formato de las cadenas verbales de las dos lenguas. Ejemplo de la segmentacion que acabamos de presentar.

eduki egitura egon A1 , funts aldaketa bat egin izan A1

Combinacion de etiquetas: Creemos que cada etiqueta por separado mejorará en parte la calidad de la alineación pero utilizando todas las etiquetas a la vez lograremos

los mejores resultados. Para ello hemos creado dos corpus utilizando todas las etiquetas anteriormente mencionadas.

La diferencia de estos dos corpus reside en que en uno de ellos se utilizan las etiquetas caso y número **(TodoC)** y en el otro la etiqueta declinación **(TodoD)** que es la combinación de estas dos.

Ejemplo de la segmentacion que hemos llamado TodoC:

eduki M,S GEN egitura M,S DAT egon A1 KAUS , funts MG, ABS aldaketa MG, GEN bat MG, ABS egin BURU izan A1

Y la que hemos llamado TodoD:

eduki GEN,M,S egitura DAT,M,S egon A1 KAUS , funts ABS,MG, aldaketa GEN,MG, bat ABS,MG, egin BURU izan A1

4. Preprocesamiento del español

Al igual que hemos hecho con el euskera también hemos preprocesado el español intentando que la apariencia final del texto en las dos lenguas sea lo más parecido posible. Para analizar el español hemos utilizado el analizador de código abierto Freeling.

4.1. Tokenización

El primer paso es tokenizar el texto, ya que, como hemos explicado anteriormente, Giza++ tokeniza el texto usando simplemente los espacios en blanco. Mostramos como Ejemplo de texto tokenizado la traduccion de la frase utilizada como ejemplo para el euskera.

Por lo que se refiere a la estructura del contenido , se ha hecho algún retoque sustantivo

4.2. Lematización

Al igual que con el euskera no hemos unido los terminos multi-palabra en un único token, esperando que Giza++ los reconozca por su cuenta. Seguimos con el mismo ejemplo esta vez lematizado.

por el que él referir a el estructura del contenido , él haber hacer alguno retoque sustantivo

4.3. Segmentación

En el caso del español nos hemos limitado a procesar las cadenas verbales, ya que con el procesamiento del euskera hemos logrado igualar el formato de los dos textos para el resto de casos.

Para prepocesar las cadenas verbales en español hemos analizado éstas con el analizador Freeling y por cada token analizado como verbo hemos creado dos tokens, el lema de dicho token y una etiqueta que expresa el modo y el tiempo de dicho verbo. Hemos creado dos corpus distintos usando este procesamiento. En uno de ellos hemos dejado el resto de tokens con su forma original y en el otro el resto de tokens los hemos remplazado por sus respectivos lemas.

Ejemplo del procesamiento de las cadenas verbales, manteniendo la forma en el resto de casos.

Por lo que se referir IP a la estructura del contenido , se haber IP hacer P0 algún retoque sustantivo

Y ejemplo del procesamiento de las cadenas verbales, lematizando el resto de tokens.

por el que él referir IP a el estructura del contenido , él haber IP hacer P0 alguno retoque sustantivo

5. Diseño del experimento

Para poder evaluar cúal es la técnica de preproceso que da mejores resultados para este par de idiomas, hemos diseñado el siguiente marco experimental. Por un lado hemos utilizado un corpus paralelo concreto español-euskera de aproximadamente un millón de palabras y ochocientas mil palabras para cada idioma respectivamente (seccion 5.1). De este corpus, hemos separado 200 frases para su anotación manual, del que hemos utilizado 100 para desarrollo y 100 para la evaluación final (seccion 5.2). En la anotación manual, se alinea token a token, por lo que en la evaluación los corpus preprocesados alineados automáticamente tienen que ser mapeados a los tokens del corpus original (seccion 5.4). Las distintas estrategias de preproceso fueron evaluadas sobre la parte de desarrollo, ya que guardamos la parte de evaluacion para experimentos futuros.

5.1. Corpus paralelo

El corpus paralelo está extraído de una memoria de traducción donde las sentencias fueron alineadas a mano. Se trata de una coleccion de 6 libros universitarios con temática variada, desde fósiles, música, administración a historia. Contiene aprox. 36.000 frases, dando lugar a un millón de palabras en español y 800.000 palabras en euskera.

5.2. Alineamiento de referencia (gold standard)

La parte del corpus que fue manualmente anotada se tomó completamente al azar de los 6 libros. De las 200 sentencias se separaron, también al azar, 100 para desarrollo y 100 para evaluación.

Para la alineación manual se contó con un lingüista que tardo aproximadamente 24 horas en realizar la tarea. Dado que se alineaba token a token, el alineamineto de referencia incluye un número significativo de alineamientos múltiples (tanto 1:N como M:1 o M:N) y de alineamiento nulos (0:1 o 1:0). La herramienta para la anotación es de libre acceso.

Previamente al etiquetado, preparamos un manual de anotación inspirado en el de Melamed (Melamed, 1998), en el que fuimos incorporando los casos más dudosos para este par de idiomas. Por ejemplo, los pronombres átonos como en "**se** tomó un trago" se alineó con el verbo auxiliar en euskera "tragoa edan **zuen**".

5.3. Software de alineamiento

Hemos utilizado el software de libre distribucion GIZA++, con los parámetros por defecto, que utiliza el modelo de alineamiento IBM-4. Este software es el más utilizado para alinear corpus paralelos, paso imprescindible para la traducción automática estadística.

5.4. Evaluación y mapeo de las alineaciones automáticas

El alineamiento automático de las diferentes posibilidades de preproceso, no produce directamente el alineamiento de los tokens originales, sino que da como salida el alineamiento entre los tokens artificiales introducidos por el preproceso. Por tanto para cada token artificial generado en el preproceso, se guarda la posicion del token original, de forma que del alineamiento automático se puede reproducir el alineamiento para los tokens originales.

Por ejemplo, para la frase que hemos utilizado de ejemplo a lo largo del artículo, habiendo procesado el euskera del modo que hemos llamado TodoD (usando todas las etiquetas expuestas en el artículo pero uniendo el caso y el numero un una sola etiqueta) y el español Lema+V (lematizando todas las palabras y ademas añadiendo a los verbos la información de tiempo), el alineamiento que da GIZA++ es el que se puede ver en la figura 1. En cambio, tras el mapeo, el alineamiento que se consigue esta basado en los tokens originales y es el que se puede ver en la figura 2.

Figura 1: Alineamiento original creado por GIZA++ para el par Lema+V TodoD

Figura 2: Alineamiento tras el enlace para el par Lema+V TodoD

Una vez obtenido el alineamiento de los tokens originales, utilizamos el software libre utilizado en las dos últimas competiciones públicas de alineamiento (Martin, Mihalcea, y Pedersen, 2005) y (Mihalcea y Pedersen, 2003).

6. Análisis de los resultados

La tabla 1 presenta los resultados para una de las combinaciones (LEMA+V para español, y TodoD para el euskera). En las columnas tenemos distintas métricas de calidad segun las devuelve el software de evaluacion. Las tres primeras columnas devuelven la precisión, el cobertura y la combinación armónica de ambas. Un resultado más alto indica más calidad. Estas medidas están disponibles para alineamiento seguros o posibles, pero como en nuestro caso sólo hemos producido alineamientos seguros, las dos versiones producen resultados idénticos. En la última columna se muestra la *Alignment Error Rate*, que es la medida más utilizada y donde el valor inferior denota mejor calidad. En un principio deberíamos dar estos valores para cada combinación probada (ver más abajo), pero

	Precisión	Cobertura	Fmeasure	AER
NULL	50.46	56.62	53.36	46.64
NULL (weighted)	76.87	55.84	64.69	35.31
NO-NULL	73.30	73.79	73.54	26.46
NO-NULL (weighted)	86.64	67.35	75.78	24.22
Unión-NULL	42.06	66.13	51.42	48.58
Unión-NULL (weighted)	63.92	80.08	71.10	28.90
Unión-NO-NULL	63.18	64.42	63.79	36.21
Unión-NO-NULL (weighted)	77.29	75.21	76.24	23.76
Intersección-NULL	60.60	43.88	50.90	53.07
Intersección-NULL (weighted)	80.78	55.10	65.51	31.37
Intersección-NO-NULL	94.23	44.59	60.53	39.47
Intersección-NO-NULL (weighted)	96.20	53.60	68.84	31.16

Cuadro 1: Distintos métodos de evaluación para LEMA+V para el español y TodoD para el euskera

en el caso de nuestros resultados, el Fmeasure y la AER siempre han estado directamente relacionados, por lo que nos centraremos en la AER para ahorrar espacio.

Al evaluar los alineamientos, el software da dos posibilidades: una evaluación que toma en cuenta los alineamientos nulos (0:1 o 1:0, que en la tabla corresponden a las líneas con NULL), o la evaluación que descarta estos alineamientos (que corresponde a NO-NULL en la tabla). Además puede dar la misma importancia a todos los alineamientos o dar un peso inferior a los alineamientos múltiples (indicado por weighted en la tabla).

Dado que GIZA++ produce dos alineamientos, uno para cada dirección, las cuatro primeras líneas corresponden a la media de las métricas para cada dirección. Las siguientes líneas corresponden a dos estrategias para combinar las dos direcciones de alineamiento: en el caso de unión se toman todos los alineamientos indistintamente (con lo que se gana en cobertura, pero se pierde precisión), y en el caso de la intersección solamente se toman aquellos alineamientos que se encuentran en las dos direcciones (con lo que se gana precisión, pero se pierde cobertura).

Al igual que para las métricas de calidad, un análisis detallado de los resultados en las filas para todas las combinaciones probadas, demostró que las mejores técnicas se confirmaban en todas y cada una de las posibilidades de evaluación. Dadas las restricciones de espacio, en adelante sólo mostraremos los resultados para el AER de Union(weighted) para las dos variantes (NULL y NO-NULL), que son los que mejores resultados (menor AER) dan en todos los casos.

En la tabla 2 se muestran los resultados para las 40 posibilidades de combinar los preprocesos. En el caso de que no realizemos ningún tratamiento para el español (primera fila), la mejor técnica de preproceso del euskera es TodoC, que usa todos los rasgos morfológicos sin agruparlos, con lo que los artículos, preposiciones, etc. del castellano encuentran un token correspondiente en euskera. La mejora sobre no procesar nada (forma-forma) es de casi 10 puntos.

En el caso de la lematización del español (segunda fila), se pierde informacion para el castellano, pero el uso de lemas y el rasgo que agrupa el caso y el número es el que mejor se ajusta. Aquí se mejoran los resultados sobre no lematizar el castellano en 8 décimas.

En el caso de que a las formas del castellano les añadamos el rasgo del tiempo y concordancia verbales (tercera fila, Forma+V), la mejor codificación para el euskera vuelve a ser TodoC, mejorando los resultados sobre la forma únicamente en un punto y décima y media sobre la lematización.

Finalmente, en el caso de que el castellano incluya el lema y los rasgos verbales (última fila, Lema+V), los mejores resultados corresponden a TodoD, donde se usan todos los rasgos pero uniendo el caso y el número como en la estrategia DEC. Esta estrategia da los mejores resultados, reduciendo en medio punto la anterior, y haciendo que el error sobre forma-forma descienda en 13 puntos aprox.

Los resultados cuando se toman en cuenta los alineamientos nulos (tabla 3) son análogos, pero en este caso el error es superior, quedando en 28.9.

Nuestro trabajo se diferencia de sus an-

	Forma	Lema	ASP	CAS	DEC	SUB	MOD	NUM	TodoC	TodoD
Forma	36.24	30.21	30.40	26.87	26.72	30.16	30.21	27.71	**25.39**	26.63
Lema	32.74	29.07	28.85	24.90	**24.48**	28.76	28.69	27.37	26.32	25.00
Forma+V	33.80	29.33	29.18	26.15	25.54	29.18	28.41	26.39	**24.32**	25.05
Lema+V	32.17	28.77	28.85	24.92	23.96	28.61	28.12	26.67	25.00	**23.76**

Cuadro 2: AER para las diferentes combinaciones (Union-NO-NULL(weighted))

	Forma	Lema	ASP	CAS	DEC	SUB	MOD	NUM	TodoC	TodoD
Forma	41.28	33.42	33.45	31.57	31.67	33.11	33.22	32.46	**31.44**	31.98
Lema	37.40	31.52	31.44	29.19	**29.00**	31.09	30.97	31.80	31.33	29.75
Forma+V	39.58	33.13	32.96	30.96	30.63	32.75	32.36	31.47	**30.15**	30.26
Lema+V	37.53	32.17	32.02	29.58	29.12	31.74	31.51	31.97	30.22	**28.90**

Cuadro 3: Resultados de los distintos experimentos (Union-NULL(weighted))

tecesores en que es el primero que preprocesa el euskera para intentar mejorar el alineamiento. Como hemos explicado anteriormente en (Caseli, M. Nunes, y Forcada, 2005) no hacían ningún preprocesamiento del corpus e intentaban mejorar la alineacion mediante heurísticos independientes de la lengua. En dicho artículo usaban un corpus diferente para evaluar los resultados, por lo que los resultados no se pueden comparar directamente pero para tener una referencia conseguían mejorar el AER de un 35.52 que conseguia GIZA++ usando los parámetros por defecto, a un 26,52 que conseguían usando la técnica de alineamiento propuesta en el artículo.

Nuestros resultados son comparables a los que se han hecho públicos en las competiciones de alineamiento. Aunque una comparacion directa no sea justa, para tener una referencia queremos citar que nuestros resultados son mejores que los disponibles para los vencedores del rumano-inglés (26.55) y el inglés-hindú (32.12), pero peores que el inglés-francés (5.71) y el inglés-esquimal (9.46).

7. Conclusiones y trabajo futuro

Estos son los primeros pasos para lograr un alineamiento entre el español-euskera de una calidad suficiente para ser utilizada en tareas más complicadas, como la traducción estadística o la basada en ejemplos.

En este primer experimento hemos aumentado sensiblemente la calidad de la alineación, dividiendo las palabras en lemas y etiquetas morfológicas para harmonizar las secuencias de tokens en ambas lenguas y así facilitar que cada token tenga su correspondiente en el otro idioma. Usando GIZA++ conseguimos una reducción del error (Alignment Error Rate) de hasta un 12.5 % sobre el baseline forma-forma, llegando al 23.76 %. Este resultado es comparable al obtenido por otros autores en competiciones públicas para otros pares de idiomas aglutinantes como el euskera.

Para el futuro, hemos observado que el orden de los tokens es distinto en cada uno de los idiomas, y esto crea problemas a los alineadores actuales. Para lo cual pretendemos cambiar el orden de los tokens del euskera para asemejarlo al orden del español, y de este modo facilitar el trabajo del alineador. Creemos que así mejoraremos aún más la calidad de la alineación.

Una vez conseguido un alineamiento de una calidad aceptable, pretendemos, además de usarlo para desarrollar un traductor estadístico, usar el corpus alineado automáticamente como corpus de entrenamiento para aprender la traducción más adecuada de una palabra según el contexto.

Bibliografía

Aduriz, I., I. Alegria, J. Arriola, X. Artola, A. Díaz de Ilarraza, y N. Ezeiza. 1994. EUSLEM: un lematizador/etiquetador de textos en euskara. En *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Córdoba, Spain.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: an Open-Source Suite of Language Analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC).*

Caseli, H., M. Nunes, y M. Forcada. 2005. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. En *Proceedings of the XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Granada, Spain, September.

Goldwater, S. y D. McClosky. 2005. Improving Statistical MT Through Morphological Analysis. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *MT Summit X*, September.

Martin, J., R. Mihalcea, y T. Pedersen. 2005. Word Alignment for Languages with Scarce Resources. En *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, USA, June.

Melamed, D. 1998. Annotation Style Guide for the Blinker Project. Informe técnico, Institute for Research in Cognitive Science, Philadelphia, USA.

Mihalcea, R. y T. Pedersen. 2003. An Evaluation Exercise for Word Alignment. En *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, May.

Nevado, F., F. Casacuberta, y J. Landa. 2004. Translation Memories Enrichment by Statistical Bilingual Segmentation. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Nießen, S. y H. Ney. 2000. Morpho-syntactic analysis for reordering in statistical machine translation.

Nießen, S. y H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Comput. Linguist.*, 30(2):181–204.

Och, F. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.