DESCRIPCIÓN DE LOS SISTEMAS PRESENTADOS POR IXA-EHU A LA EVALUACIÓN ALBAYCIN'08

Gorka Labaka, Arantza Díaz de Ilarraza, Kepa Sarasola

Grupo IXA
Universidad del País Vasco
{jiblaing, jipdisaa, jipsagak}@ehu.es

RESUMEN

En este artículo describimos los sistemas presentados por el grupo IXA-EHU a la evaluación ALBAYCIN'08. Dada las características de los pares de lenguas a tratar y la naturaleza aglutinativa del euskara hemos procedido a la segmentación de las palabras en morfemas para, de este modo, facilitar el alineamiento. Además este proceso habilita la posibilidad de aprender pseudosintagmas (secuencias de palabras sin estructura sintáctica) que pueden estar compuestos además de por palabras, por morfemas considerados de manera independiente a la palabra a la que van unidos; por ejemplo, en el caso de 'etxe-ra' noa (voy a casa), el pseudosintagma '-ra noa' se puede alinear con 'voy a'.

Además de la segmentación hemos incorporado a la tabla de traducción pares de pseudosintagmas que han sido extraídos utilizando técnicas de traducción basada en ejemplos de MaTrEx [1]. Estos nuevos pseudosintagmas, a diferencia de los extraídos por las técnicas estadísticas, coinciden con sintagmas desde un punto de vista lingüístico. Al ampliar la tabla de traducción con estos nuevos pseudosintagmas, se amplia la cantidad de pseudosintagmas disponibles por el decodificador, además de favorecer aquellas traducciones sintácticamente correctas.

1. INTRODUCCIÓN

En este artículo describimos los sistemas presentados por el grupo IXA de la Universidad del País Vasco a la evaluación Albaycin'08.

El alto nivel de flexión del euskara, junto con el hecho de que no haya gran cantidad de corpus paralelo accesible, complica la tarea de traducción español-euskara convertiendola en un reto interesante.

Para hacer frente a su alto nivel de flexión hemos segmentado las palabras en euskara dividiéndolas en morfemas, de modo similar al realizado en otros trabajos para otros pares de lenguas de gran flexión, como en el caso del inglés-checo [2] y el inglés-turco [3].

El articulo está organizado del siguiente modo: en la sección 2 explicamos las distintas técnicas que usaremos en nuestros sistemas; la sección 3 está dedicada a mostrar los sistemas que hemos evaluado y que se basan en combinaciones de las técnicas previamente explicadas; en la sección 4 resumimos los resultados conseguidos por cada uno de los sistemas; finalmente comentamos las conclusiones extraídas de esos resultados (sección 5).

Este trabajo ha sido subvencionado por Gobierno Vasco, mediante la ayuda predoctoral concedida a Gorka Labaka (código BFI05.326)

2. TÉCNICAS UTILIZADAS

En esta sección explicamos las técnicas que hemos utilizado en la implementación de los sistemas que presentamos a la evaluación Albaycin'08.

2.1. Segmentación del texto en euskara

Dada la naturaleza aglutinante de la lengua y, continuando con el trabajo presentado en un publicación anterior [4], hemos llevado a cabo la segmentación del texto en euskera. De esta manera, tendremos en tokens independientes los morfemas que conforman una palabra.

Para llevar a cabo esta segmentación hemos analizado el texto en euskara utilizando EusTagger[5] y cada palabra se ha separado en como máximo tres tokenes: los prefijos, el lema y los sufijos. De este modo, para una palabra como 'etxekoa' (el de la casa) se crean dos tokenes: 'etxe' y '+koa'. En una primera aproximación, pensamos en crear un token por cada morfema pero, dadas las características de la salida de EusTagger que genera una segmentación muy fina (con muchos morfemas por palabra), decidimos unir todos los sufijos en un único token; los prefijos también fueron tratados de la misma manera. La razón principal para llevar a cabo esta segmentación es facilitar el alineamiento, ya que de este modo habrá menos alineamientos múltiples a la vez que se reduce la dispersión.

Gracias a este proceso de segmentación podemos aprender pseudosintagmas en los que toman parte sólo algunos de los morfemas de una palabra. De este modo se podría extraer el par de pseudosintagmas 'voy a' '-ra noa', donde la preposición 'a' se traduce con el sufijo '-ra' cuando acompaña al verbo 'ir' independientemente del lema al que esté unido. Sin la segmentación no sería posible esta clase de generalización teniendo que extraer pseudosintagmas distintos para cada ejemplo.

El hecho de usar el texto segmentado para entrenar el traductor estadístico, conlleva que necesitemos generar el texto final en euskara basándonos en la salida del traductor, ya que esta estará segmentada al igual que el corpus utilizado en el entrenamiento. Para generar el texto final hemos utilizado el módulo de generación del traductor basado en reglas matxin[6] (que utiliza en el mismo léxico que el analizador).

A la hora de generar el texto final hay que tener en cuenta que el traductor estadístico puede producir combinaciones de morfemas que no correctas pudiéndole asignar a un nombre la flexión correspondiente a un verbo o incluso llegando a asignarle algún tipo de flexión a tokenes que no se pueden flexionar como los signos de puntuación. En este caso y, como primera aproximación, eliminamos la flexión dejando únicamente el lema.

Finalmente, para poder incorporar un modelo de lenguaje basado en palabras (el decodificador usará uno basado en el texto segmentado), en vez de obtener sólo la mejor traducción que el decodificador es capaz de encontrar, obtenemos una lista de las n traducciones más probables y, tras la generación, reordenamos la lista de traducciones incorporando el modelo de lenguaje basado en palabras como si fuera un modelo más.

2.2. Hibridación SMT-EBMT: sistema MaTrEx

En colaboración con Nacional Centre for Language Technology de la Dublin City University hemos adaptado su sistema MaTrEx[1] para utilizarlo con el euskara. Este sistema consiste en enriquecer la tabla de traducción con pares de pseudosintagmas extraídos usando técnicas de la traducción automática basada en ejemplos.

Para extraer los nuevos pseudosintagmas, se analizan sintácticamente ambas partes del corpus paralelo y se marcan los sintagmas (hemos usado Freeling [7] para procesar el español y Eustagger para el euskara). En un segundo paso y basándose en los alineamientos palabra por palabra se alinean estos sintagmas y se incorporan a tabla de traducción.

3. SISTEMAS PRESENTADOS

Para crear nuestros sistemas hemos utilizado las siguientes herramientas:

- Alineador de palabras GIZA++ [8].
- Modelo de lenguaje SRILM [9]
- Moses SMT Toolkit [10]

Mediante estas herramientas de uso libre y los corpora habilitados por la organización (en la tabla 1 se muestra algunos datos de los corpora) hemos creado un sistema *baseline*, usando los *scripts* y los parámetros que Moses trae por defecto. Hay que tener en cuenta que el sistema *baseline* de Moses incorpora técnicas de reordenación lexicalizada además de la basada en distancia. En la creación del *baseline* se han llevado a cabo la optimización de los pesos de cada modelo usando BLEU y Minimum Error Rate Training.

Basándonos en este *baseline* hemos incorporado las técnicas explicadas en la sección 2 creando distintos sistemas de traducción. Posteriormente hemos evaluado el impacto que tiene cada técnica. Para incorporar los pseudosintagmas correspondientes al sistema MaTrEx, hay que analizar ambos textos, alinear los sintagmas basándose en los alineamientos palabra por palabra e incorporar los nuevos pares de pseudosintagmas a la tabla de traducción antes de calcular los pesos de los modelos de traducción con los *scripts* proporcionados con Moses. Tras este proceso se continua con el entrenamiento del sistema.

Por otro lado a la hora de usar el texto segmentado, además de preprocesar y post-procesar las oraciones en euskara, para segmentar el texto y volver a generar la forma final, hemos tenido que modificar el proceso de optimización para poder optimizar también el peso del modelo de lenguaje basado en palabras. Como hemos explicado anteriormente, el decodificador utiliza un modelo de lenguaje basado en el texto segmentado, y el modelo de lenguaje basado en palabras se incorpora a la traducción después del post-proceso de generación mediante el reordenamiento de una lista n-best. Por lo que en cada paso de la optimización hay que incorporar tanto la generación como el reordenamiento de las listas basándose en la lista n-best.

Además de los sistemas donde probamos las técnicas presentadas individualmente, también hemos entrenado un sistema donde probamos la combinación de ambas.

4. RESULTADOS

Hemos evaluado los sistemas presentados en la sección 3 sobre el corpus de test usando las métricas automáticas más usuales (BLEU, MBLEU, WER, PER). En la tabla 2 se presentan los resultados para dichos sistemas y métricas.

Lo más destacable de los datos presentados es que ambas técnicas individuales (MaTrEx y segmentación del euskara) mejoran los resultados del sistema *baseline* para todas las métricas utilizadas. A su vez, la combinación de técnicas supera a cada una de ellas considerada individualmente, logrando los mejores resultados.

5. CONCLUSIÓN

Las técnicas que hemos utilizado han dado un resultado satisfactorio mejorando ambas el *baseline*. Además la combinación de las misma supera a las técnicas aplicadas individualmente.

Respecto al trabajo futuro, nos proponemos modificar la segmentación del euskara, buscando una forma alternativa para agrupar los morfemas. Actualmente, todos los morfemas que acompañan al lema se agrupan en un único

corpora	lenguaje	oraciones	tokenes	vocabulario-tokenes
entrenamiento	Español	58202	1284212	50927
	Euskara		1010545	95724
	Euskara-segmentado		1546304	40436
development	Español	1456	32743	7073
	Euskara		25778	9030
	Euskara-segmentado		39420	6191
test	Español	1446	31004	6836
	Euskara		24372	8695
	Euskara-segmentado		37347	5976

Tabla 1. Estadísticas de los corpora utilizados.

	BLEU	MBLEU	NIST	WER	PER
baseline	10.82	10.21	4.51	80.44	61.67
MaTrEx	11.03	10.38	4.54	80.13	61.65
segmentación euskara	11.19	10.49	4.65	79.27	60.60
segmentación + MaTrEx	**11.37**	**10.65**	**4.71**	**78.65**	**60.01**

Tabla 2. Evaluación de los distintos sistemas probados.

token ya que se consiguen mejores resultados que manteniendo cada morfema un token pero pensamos que una agrupación intermedia mejoraría los resultados.

Por otro lado, nos planteamos mejorar el proceso de generación de la forma final, modificando la secuencia de morfemas que devuelve el traductor estadístico en aquellos casos que ésta no sea morfológicamente correcta. Esta modificación puede implicar la reordenación de la secuencia o la eliminación de algunos de los morfemas.

6. BIBLIOGRAFÍA

[1] N. Stroppa y A. Way, "MaTrEx: DCU Machine Translation System for IWSLT 2006," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 31–36.

[2] S. Goldwater y D. McClosky, "Improving statistical mt through morphological analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, 2005.

[3] Kemal Oflazer y Ilknur Durgar El-Kahlout, "Exploring different representational units in english-to-turkish statistical machine translation," in *Proceedings of Statistical Machine Translation Workshop at ACL 2007*, Prague, Czech Republic, June 2007.

[4] E. Agirre, A. Díaz de Ilarraza, G. Labaka, y K. Sarasola, "Uso de información morfológica en el alineamiento español-euskara," in *XXII Congreso de la SEPLN*, Zaragoza, septiembre 2006.

[5] I. Aduriz y A. Díaz de Ilarraza, "Morphosyntactic disambiguation and shallow parsing in computational porcessing of basque," in *Inquiries into the lexicon-syntax relations in Basque*, Bernarrd Oyharçabal, Ed., Bilbao, 2003.

[6] I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. Forcada, S. Ortiz, y L. Padró, "An open architecture for transfer-based machine translation between spanish and basque," in *Workshop on Open-Source Machine Translation*, Asia-Pacific Association for Machine Translation (AAMT), Ed., Phuket, Thailand, September 2005, pp. 7–14.

[7] X. Carreras, I. Chao, L. Padró, y M. Padró, "Freeling: an open-source suite of language analyzers," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2004.

[8] F. Och y H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[9] Andreas Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September 2002.

[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, y Evan Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007.