DOI: 10.17234/SRAZ.65.45

UDK: 811.134'322.4
Original scientific paper
Recibido el 30 de junio de 2020
Aceptado para la publicación el 25 de noviembre de 2020

Traducción automática para las lenguas románicas de la península ibérica

Antoni Oliver
Universitat Oberta de Catalunya (UOC)
aoliverg@uoc.edu

En este estudio presentamos una comparación de tres estrategias de traducción automática aplicadas a diversas lenguas románicas de la península ibérica: español, portugués, gallego, catalán, aranés, aragonés y asturiano. En nuestro estudio analizaremos sistemas con el español como lengua de partida. El objetivo del trabajo es evaluar los sistemas de transferencia sintáctica superficial para estos pares de lenguas y determinar si se dispone de corpus paralelos libres de tamaño y calidad suficientes para entrenar sistemas de traducción automática estadísticos y neuronales que puedan ofrecer una calidad similar o superior. Las preguntas de investigación de este trabajo son: ¿Qué calidad ofrece el sistema de traducción Apertium para los pares de lenguas analizados? ¿Cuántos corpus están disponibles libremente para el entrenamiento de sistemas estadísticos y neuronales? ¿Qué calidad se obtiene con los sistemas estadísticos y neuronales entrenados con los corpus disponibles y con los corpus sintéticos?

Palabras clave: traducción automática, lenguas románicas, península ibérica, corpus paralelos

1. Introducción: las lenguas románicas de la península ibérica

En la península ibérica existe una extensa variedad de lenguas románicas. Estas lenguas son muy dispares en lo que hace referencia a su estatus de oficialidad y a su número de hablantes. Estos dos factores, oficialidad y número de hablantes, es determinante también en lo que respecta al número y calidad de los sistemas de traducción automática disponibles. En esta sección vamos a hacer un breve repaso de las lenguas románicas existentes en la península ibérica y a presentar el nivel de oficialidad en tres niveles: oficialidad estatal (la oficialidad en todo un estado de la península ibérica), oficialidad autonómica o regional (oficialidad en una autonomía o región o al menos en parte de ella), y oficialidad internacional (oficialidad en instituciones internacionales como la Unión Europea o las Naciones Unidas). En esta sección ofreceremos también cifras aproximadas de hablantes de estas lenguas en la península ibérica. En la tabla 1 se puede ver reflejada de forma esquemática esta información.

Lengua	ISO	O.E.	O.A.	O.I	Hablantes
español	spa	X	X	X	46.000.000
portugués	por	X	X	X	11.000.000
catalán	cat	X	X		10.000.000
gallego	gal		X		2.500.000
asturiano	ast				110.000
aragonés	arg				30.000
mirandés	mwl		X		15.000
aranés	oc-aran		X		4.500

Tabla 1. Nivel de oficialidad y número de hablantes en la península de las lenguas románicas de la península ibérica (O.E: oficialidad estatal; OA: oficialidad autonómica o regional; O.I.: oficialidad internacional)

2. La traducción automática

Las estrategias de traducción automática analizadas en el presente estudio son las siguientes:

- **Transferencia**: se realiza algún tipo de análisis automático sobre la oración en la lengua de partida (generalmente análisis morfosintáctico). Este análisis se transfiere, mediante el uso de una serie de reglas, a una estructura propia de la lengua de llegada. Una vez transferida la estructura de análisis se utilizan unos diccionarios de transferencia que nos permiten obtener la oración traducida en la lengua de llegada.
- **Sistemas estadísticos**: la traducción automática estadística se basa en el uso de modelos estadísticos obtenidos a partir de corpus paralelos. Estos sistemas se basan principalmente en dos modelos probabilísticos: el modelo de traducción y el de lengua de llegada. También se dispone de un modelos de reordenamiento.
- **Sistemas neuronales**: estos sistemas se basan en el uso de redes neuronales artificiales entrenadas a partir de corpus paralelos.

Se han escogido analizar estas estrategias por los diversos motivos. Existen sistemas de transferencia sintáctica superficial para todos los pares de lenguas bajo estudio y esta estrategia se ha utilizado con notable éxito para el desarrollo de sistemas entre lenguas similares, como son las lenguas implicadas en el presente estudio. Ahora bien, los sistemas estadísticos han reemplazado a la mayoría de sistemas de transferencia y más recientemente los sistemas neuronales están reemplazando a los sistemas estadísticos. La investigación actual en traducción automática se centra en los modelos neuronales.

2.2. Apertium

Apertium[1] (Forcada et al. 2011) es un sistema de transferencia sintáctica superficial que se distribuye bajo una licencia de software libre y ofrece todos los pares de lengua bajo estudio. En la tabla 2 podemos ver los pares de lenguas románicas de la península ibérica que disponen de sistema de traducción automática Apertium.

	spa	por	gal	cat	oc-aran	arg	ast
spa	-	X	X	X	X	X	X
por	X	-	X	X			
gal	X	X	-				
cat	X	X		-	X	X	
oc-aran	X			X	-		
arg	X			X		-	
ast	X						-

Tabla 2. Pares de lenguas bajo estudio disponibles en Apertium

2.3. Toolkits para el entrenamiento de sistemas estadísticos y neuronales

Además de los sistemas ya disponibles existen una serie de toolkits que permiten entrenar sistemas estadísticos y neuronales y traducir con los modelos entrenados. Entre los toolkits disponibles utilizaremos los siguientes:

- Moses[2] (Koehn et al. 2007): es el toolkit estadístico por excelencia que ha sido utilizado intensivamente tanto en entornos de investigación como de producción.
- Marian[3] (Junczys-Dowmunt et al. 2018): es un sistema neuronal programado en C++ que destaca por su velocidad de entrenamiento y traducción.

Ambos sistemas se distribuyen bajo licencias libres permisivas por lo que pueden utilizarse tanto en entornos de investigación como de producción.

3. Los corpus paralelos entre las lenguas románicas de la península ibérica

En esta sección vamos a analizar los corpus paralelos disponibles entre las lenguas románicas de la península ibérica. Para ello utilizaremos el sitio web

[1] https://www.apertium.org/
[2] http://www.statmt.org/moses/
[3] https://marian-nmt.github.io/
[4] http://opus.nlpl.eu/

Opus Corpus[4] (Tiedemann 2012), que es una de las mayores colecciones de corpus paralelos de libre acceso disponibles libremente. En la tabla 3 se pueden observar los corpus disponibles en Opus Corpus junto con el número de segmentos.

	spa	por	gal	cat	*oc	arg	as
spa	-	**49.8 M**	**1.4 M**	**12.6 M**	0.4 M	97.4 K	0.7 M
por	**49.8 M**	-	**1.1 M**	**1.5 M**	0.4 K	33	0.5 K
gal	**1.4 M**	**1.1 M**	-	**1 M**	0.4 M	93.7 K	97.2 K
cat	**12.6 M**	**1.5 M**	**1 M**	-	0.4 M	93.7 K	93.7 K
*oc	0.4 M	0.4 M	0.4 M	0.4 M	-	46.4 K	0.3 M
arg	97.4 K	33	92.7 K	93.7 K	46.4 K	-	86.4 K
ast	0.7 M	0-5 K	0.7 M	0.6 M	0.3 M	86.4 K	-

Tabla 3. Corpus paralelos disponibles en la colección Opus Corpus

En la tabla 3 podemos observar marcados en negrita los corpus paralelos que disponen de más de 1 millón de segmentos. Como podemos observar se trata de los corpus entre las lenguas con más hablantes y además todas ellas gozan de algún tipo de oficialidad. Cabe destacar que no se han encontrado corpus paralelos específicamente para el aranés y que las cifras de la tabla son para el occitano.

Para el par español-gallego se ha utilizado también un subcorpus[5] de 6 millones de palabras (152 K segmentos) del corpus CLUVI[6] (Gómez Guinovart 2019).

4. Parte experimental

Como parte experimental en este trabajo nos proponemos a analizar los siguientes sistemas:

- Apertium
- Moses: sistema estadístico.
- Marian: sistema neuronal.

Para los siguientes pares de lenguas:

- español-portugués
- español-gallego
- español-catalán
- español-aragonés
- español-asturiano

[5] https://repositori.upf.edu/handle/10230/20051

[6] http://sli.uvigo.gal/CLUVI/

Como hemos visto en el apartado anterior, no se dispone de corpus paralelos de tamaño suficiente para todos los pares de lenguas bajo análisis. En estos casos se crearán corpus sintéticos tal y como se explica a continuación.

4.2. Creación de corpus sintéticos

Para la creación de los corpus sintéticos utilizaremos un corpus monolingüe de la lengua diferente del español (por ejemplo, para el par español-asturiano utilizaremos un corpus monolingüe del asturiano) y lo traduciremos al español utilizando el sistema de traducción automática Apertium. Este sistema es capaz de marcar las palabras desconocidas. Eliminamos del corpus traducido todos los segmentos que contengan palabras desconocidas.

Como corpus monolingüe utilizaremos las Wikipedias aragonesa y asturiana. Descargamos el *dump* XML[7] más reciente, lo convertimos a texto y lo segmentamos para obtener un corpus segmentado. Estos segmentos se traducen al español utilizando Apertium y eliminando los segmentos que contienen palabras desconocidas.

4.3. Evaluación automática de los sistemas

Para evaluar los sistemas de traducción automática utilizaremos una serie de métricas de evaluación automática. Presentaremos los resultados de tres métricas automáticas:

- BLEU (Papineni et al. 2002): es una métrica que se calcula a partir de n-gramas de palabras. Un número más alto indica mayor calidad.
- NIST (Doddington et al. 2002): esta métrica se calcula a partir de n-gramas ponderados por frecuencia. Los valores más altos indican mayor calidad.
- WER (*Word Error Rate*): Se calcula a partir del número de sustituciones, inserciones y borrados de palabras. Los valores más bajos indican mayor calidad.

Para obtener estas métricas utilizamos el programa MTUOC-eval[8] (Oliver 2020). En los pares de lenguas que disponen de sistema Google Translate ofrecemos los datos de calidad a modo de comparación.

4.4. Resultados

4.4.a. Español-portugués

En la tabla 4 podemos observar los resultados de la evaluación automática de los sistemas. La calidad obtenida para todos estos sistemas es muy similar.

[7] https://dumps.wikimedia.org

[8] https://sourceforge.net/project/mtuoc

Los mejores resultados se han obtenido para el sistema Marian. Estos buenos resultados para el sistema neuronal se pueden explicar por el gran tamaño de los corpus utilizados.

Sistema	**BLEU**	**NIST**	**WER**
Apertium	0.2622	7.464	0.5971
Google T.	0.2795	7.836	0.5693
Marian	0.2995	7.993	0.5661
Moses	0.2886	7.722	0.5839

Tabla 4. Resultados de la evaluación para el par español-portugués.

4.4.b. Español-gallego

En la tabla 5 se pueden observar los resultados de la evaluación. En este caso, aún siendo el corpus de entrenamiento de mucho menor tamaño que el corpus del español-portugués, los resultados de evaluación para los sistemas neuronal y estadístico entrenados son muy buenos, siendo el sistema neuronal Marian el mejor de todos los analizados.

Sistema	**BLEU**	**NIST**	**WER**
Apertium	0.6759	12.016	0.2938
Google T.	0.2959	7.672	0.5396
Marian	0.7187	12.547	0.2869
Moses	0.6614	9.97	0.3574

Tabla 5. Resultados de la evaluación para el par español-gallego.

4.4.c. Español-catalán

Si observamos los valores de evaluación de los sistemas (tabla 6) veremos que esta vez el mejor sistema dependerá de la medida que consideremos. Así, considerando BLEU, el mejor sistema es Moses, seguido muy de cerca de Google Translate. En cambio, considerando NIST el mejor es Google Translate, seguido de Apertium. Y por último, considerando WER el mejor sería Apertium, seguido de Google Translate. Esto nos puede indicar que los tres sistemas, Apertium, Google Translate y Moses tiene una calidad, según las métricas automáticas, muy similar.

Sistema	**BLEU**	**NIST**	**WER**
Apertium	0.8392	14.670	0.1093
Google T.	0.8504	14.899	0.1168
Marian	0.7802	13.909	0.1896
Moses	0.8552	12.574	0.1742

Tabla 6. Resultados de la evaluación para el par español-catalán

4.4.d. Español-aragonés

Al no disponer de corpus paralelos de tamaño suficiente se ha creado uno sintético utilizando la Wikipedia aragonesa y Apertium. De esta manera se ha creado un corpus sintético de 116.679 segmentos. En la tabla 7 se pueden observar los resultados de la evaluación, de los que cabe destacar los bajísimos resultados obtenidos para Marian.

Sistema	**BLEU**	**NIST**	**WER**
Apertium	0.5765	9.366	0.6819
Marian	0.1399	2.7687	0.901
Moses	0.4965	8.594	0.7286

Tabla 7. Resultados de la evaluación para el par español-aragonés.

4.4.e. Español-asturiano

De manera similar al que ocurre con el español-aragonés, el español-asturiano no dispone de corpus paralelos de tamaño suficiente. Hemos creado un corpus sintético utilizando la Wikipedia en asturiano y Apertium, alcanzando un tamaño de 1.203.572 segmentos.

Sistema	**BLEU**	**NIST**	**WER**
Apertium	0.2885	6.28	0.4167
Marian	0.2231	4.917	0.722
Moses	0.1173	3.412	0.8452

Tabla 8. Resultados de la evaluación para el par español-asturiano.

En la tabla 8 podemos observar los resultados de la evaluación. El sistema Moses entrenado con corpus sintéticos es el que ofrece unos peores resultados de evaluación. Marian, aún quedando por debajo de Apertium, obtiene unos resultados más competitivos.

6. Conclusiones

En este artículo hemos presentado una panorámica sobre los sistemas de traducción automática disponibles para las lenguas románicas de la península ibérica. Hemos podido comprobar que la cantidad y la calidad de los sistemas disponibles guardan una estrecha relación tanto con el número de hablantes de las lenguas implicadas como con el nivel de oficialidad de las lenguas.

Los resultados de este estudio confirman que los sistemas de traducción automática por transferencia ofrecen un nivel de calidad destacable para lenguas similares, como las del presente estudio. Por otro lado, se confirma que para las lenguas que disponen de corpus paralelos de tamaño y calidad suficientes, la

calidad de los sistemas neuronales y estadísticos son en general similares y en algún caso superior a la de los sistemas de transferencia. Para las lenguas que no disponen de corpus paralelos suficientes, la estrategia de creación de corpus paralelos sintéticos no funciona bien para entrenar sistemas estadísticos y neuronales.

Como trabajo futuro pretendemos utilizar técnicas no supervisadas (Artetxe/Labaka/Agirre 2018) y analizar maneras de combinar estas técnicas con técnicas supervisadas. También queremos experimentar con técnicas de *transfer learning* (Zoph et al. 2016), aprovechando la gran similitud que presentan las diferentes lenguas románicas de la península ibérica.

Creemos sinceramente que el hecho de disponer de sistemas de traducción automática que ofrezcan una buena calidad y cobertura puede potenciar el uso social de algunas lenguas románicas de la península ibérica, como el aragonés y asturiano, por ejemplo, y asegurar de este modo su supervivencia.

Bibliografía

Artetxe, Mikel / Labaka, Gorka / Agirre, Eneko (2018). Unsupervised Statistical Machine Translation, en: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 3.632-3.642.

Doddington, George (2002). Automatic evaluation of machine translation quality using n-grams co-occurrence statistics, en: *Proceedings of the second international conference on Human Language Technology Research*, pp. 138-145.

Forcada, Mikel L. / Ginestí-Rosell, Mireia / Nordfalk, Jacob / O'Regan, Jim / Ortiz-Rojas, Sergio / Pérez-Ortiz, Juan Antonio / Sánchez Martínez, Felipe / Ramírez-Sánchez, Gema / Tyers, Francis M. (2011). Apertium: a free/open-source platform for rule-based machine translation, en: *Machine translation*, 25(2), pp. 127-144.

Gómez Guinovart, Xavier (2019). Enriching parallel corpora with multimedia and lexical semantics, en: From the CLUVI Corpus to WordNet and SemCor. *Parallel Corpora for Contrastive and Translation Studies: New resources and applications* [ed. John Benjamin], pp. 141-158.

Junczys-Dowmunt, Marcin / Grundkiewicz, Roman / Dwojak, Tomasz / Hoang, Hieu / Heafield, Kenneth / Neckermann, Tom / Seide, Frank / Germann, Ulrich / Fikri Aji, Alham / Bogoychev, Nikolay / Martins, André F. T. / Birch, Alexandra (2018). Marian: Fast Neural Machine Translation in C++, en_ *Proceedings of ACL 2018, System Demonstrations*, pp. 116-121.

Koehn, Philipp / Hoang, Hieu / Birch, Alexandra / Callison-Burch, Chris / Federico, Marcello / Bertoldi, Nicola / Cowan, Brooke / Shen, Wade / Moran, Christine / Zens, Richard / Dyer, Chris / Bojar, Ondřej / Constantin, Alexandra / Herbst, Evan (2007). Moses: Open source toolkit for statistical machine translation, en: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177-180.

Oliver, Antoni (2020). MTUOC: easy and free integration of NMT systems in professional translation environments, en: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 467-468.

Papineni, Kishore / Roukos, Salim / Ward, Todd / Zhu, Wei-Jing (2002). BLEU: a method for automatic evaluation of machine translation, en: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318.

Tiedemann, Jörg (2012) Parallel Data, Tools and Interfaces in OPUS, en: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pp. 2214-2218.

Zoph, Barret / Yuret, Deniz / May, Jonathan / Knight, Kevin (2016). Transfer learning for low-resource neural machine translation, en: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568-1575.

Machine Translation for Romance Languages of the Iberian Peninsula

In this paper, a comparison of three strategies of machine translation applied to several Romance languages of the Iberian Peninsula is presented. The involved languages are Spanish, Portuguese, Galician, Catalan, Aranese, Aragonese and Asturian. In our study we analyse systems having Spanish as source language.

The goal of this study is to evaluate shallow syntactic transfer systems for these language pairs and determine if there are freely available parallel corpora of enough size and quality to train statistical and neural machine translation systems able to deliver similar or higher translation quality.

The research questions of this work are: Which level of quality deliver the Apertium machine translation systems for the language pairs under study? How many parallel corpora are freely available to train statistical and neural systems? Which level of quality can be achieved with the statistical and neural systems trained with the available and synthetic parallel corpora?

Key words: machine translation, Romance languages, Iberian Peninsula, parallel corpus