

Traducción automática de un conjunto de entrenamiento para extracción semántica de relaciones*

JEFFERSON A. PEÑA-TORRES **


VÍCTOR BUCHELI ***

RAÚL E. GUTIÉRREZ DE PIÑÉREZ REYES ****


Recepción: 06 de noviembre de 2021


Aprobación: 02 de febrero de 2022


Forma de citar este artículo: Peña, J. Bucheli, V. & Guitiérrez, R, J. (2022). Traducción automática de un conjunto de entrenamiento para extracción semántica de relaciones, (39), e13436

 <https://10.19053/0121053X.n39.2022.13436>

* Artículo de investigación.

** Magister en Ingeniería con énfasis en Sistemas y Computación de la Universidad del Valle. Profesor auxiliar en la Escuela de Ingeniería de Sistemas y Computación (EISC) de la Universidad del Valle, Cali, Colombia. Grupo de Univalle en Inteligencia Artificial (GUIA). jefferson.amado.pena@correounivalle.edu.co  <https://orcid.org/0000-0002-3879-3320>

*** Doctor en Ingeniería de la Universidad de los Andes. Profesor asistente en la Escuela de Ingeniería de Sistemas y Computación (EISC) de la Universidad del Valle, Cali, Colombia. Grupo de Univalle en Inteligencia Artificial (GUIA). victor.bucheli@correounivalle.edu.co  <https://orcid.org/0000-0002-0885-8699>

**** Doctor en Ingeniería de la Universidad del Valle. Profesor asistente en la Escuela de Ingeniería de Sistemas y Computación (EISC) de la Universidad del Valle, Cali, Colombia. Grupo de Univalle en Inteligencia Artificial (GUIA). raul.gutierrez@correounivalle.edu.co  <http://orcid.org/0000-0001-9941-6206>

Resumen

La traducción automática (TA) se utiliza para obtener corpus anotados a partir de corpus provenientes del idioma inglés, los cuales pueden ser aplicables a diferentes tareas de procesamiento de lenguaje natural (PLN). Teniendo en cuenta que existen más recursos o conjuntos de datos para entrenamiento de modelos de PLN en idioma inglés, en este trabajo se explora la aplicación de la TA para automatizar tareas de PLN en el idioma español. De esta forma, en el artículo se describe un conjunto de datos para la extracción de relaciones genéricas (reACE) y la construcción de un modelo extracción semántica de relaciones en español (ER), basado en el conjunto de muestras traducidas del idioma inglés al español. Los resultados muestran que para la tarea de TA es necesario implementar un proceso de preedición del corpus en inglés, con el fin de evitar errores de traducción, posesición y mantener las anotaciones del corpus original. Los modelos ER en español alcanzan medidas de precisión, exhaustividad y valor-F comparables con las obtenidas por el modelo en el lenguaje de inglés, lo que sugiere que la traducción automática es una herramienta útil para realizar tareas de PLN en el idioma español.

Palabras clave: lingüística informática, traducción automática, lingüística de corpus, extracción de relaciones.

Machine Translation of a Training Set for Semantic Extraction of Relations

Abstract

Machine translation (MT) is used to obtain annotated corpus of English corpus which can be applicable to different natural language processing (NLP) tasks. Considering that there are more resources or data sets for training NLP models in English language, this paper explores the application of MT to automate NLP tasks in Spanish. Thus, the article describes a dataset for the extraction of generic relations (reACE) and the construction of a semantic extraction model of relations in Spanish (ER), based on the set of samples translated from English to Spanish. The results show that for the MT task it is necessary to implement a corpus pre-editing process in English to avoid translation and post-editing errors and maintain the original corpus annotations. The ER models in Spanish achieve measures of accuracy, completeness, and F-value comparable to those obtained by the model in the English language, which suggests that machine translation is a useful tool to perform NLP tasks in the Spanish language.

Keywords: computer linguistics, machine translation, corpus linguistics, relations extraction.

Traduction automatique d'un ensemble de formation pour prélèvement sémantique de relations.

Résumé

La traduction automatique (TA) est utilisée pour obtenir des corpus annotés à partir de corpus de langue anglaise, qui peuvent être applicables à différents travaux de traitement du langage naturel (NLP). En tenant compte du fait qu'il existe davantage de ressources ou d'ensembles de données pour la formation de modèles PLN en langue anglaise, cet article explore l'application de la TA pour automatiser les travaux PLN en langue espagnole. Ainsi, l'article décrit un ensemble de données pour le prélèvement de relations génériques (reACE) et la construction d'un modèle de prélèvement de relations sémantiques en espagnol (ER), basé sur l'ensemble des échantillons traduits de l'anglais à l'espagnol. Les résultats montrent que pour le travail de TA, il est nécessaire de mettre en œuvre un processus de pré-édition du corpus anglais, afin d'éviter les erreurs post-édition de traduction et de garder les annotations du corpus original. Les modèles ER en espagnol atteignent des mesures de précision, de complétude et de valeur F comparables à celles obtenues par le modèle en langue anglaise, ce qui suggère que la traduction automatique est un outil utile pour accomplir des travaux PLN en langue espagnole.

Mots-clés: linguistique informatique, traduction automatique, linguistique de corpus, prélèvement de relations

Tradução automática de um conjunto de treinamento para extração semântica de relações

Resumo

A tradução automática (TA) é usada para obter corpus anotados partindo de corpus da língua inglesa, que podem ser aplicáveis a diferentes tarefas de processamento de linguagem natural (PLN). Levando em conta que existem mais recursos ou conjuntos de dados para treinamento de modelos PLN em inglês, este artigo explora a aplicação da TA para automatizar tarefas PLN em espanhol. Desta forma, o artigo descreve um conjunto de dados para extração de relações genéricas (reACE) e a construção de um modelo de extração semântica de relações em espanhol (ER), baseado no conjunto de amostras traduzidas do inglês para o espanhol. Os resultados mostram que para a tarefa de TA é necessário implementar um processo de pré-edição do corpus em inglês, a fim de evitar erros de tradução e pós-edição e manter as anotações do corpus original. Os modelos ER em espanhol alcançam medidas de acurácia, completude e valor F comparáveis às obtidas pelo modelo na língua inglesa, o que sugere que a tradução automática é uma ferramenta útil para realizar tarefas de PLN na língua espanhola.

Palavras-chave: linguística computacional, tradução automática, linguística de corpus, extração de relações.

Introducción

El procesamiento de lenguaje natural (PLN) se ha convertido en los últimos años en uno de los campos más relevantes de las ciencias de la computación y la lingüística aplicada. Con las herramientas del aprendizaje automático, las máquinas aprenden a leer, descifrar e interpretar los lenguajes humanos, a describir, resumir, traducir e incluso a responder de manera coherente en lenguaje natural. Así, el PLN permite hoy en día desarrollar soluciones de *software* que entienden, analizan y responden de manera similar a la humana: en formato de texto o voz.

El lenguaje natural (o humano) es la fuente de información para que las máquinas aprendan a leer o hablar. En este sentido estas requieren de un corpus o un conjunto de muestras para entrenarse en las tareas de procesamiento de lenguaje natural. Es de esta forma que una máquina es capaz de aprender a comunicarse. Sin embargo, los seres humanos se expresan de infinitas formas, hay cientos de idiomas, dialectos y cada uno de ellos con sus propias reglas. Entonces, el reto del PLN radica en construir una máquina capaz de identificar automáticamente los elementos del lenguaje de orden morfológico, sintáctico y semántico.

Así, las máquinas de PLN tienen en cuenta varios componentes del lenguaje: sintácticos, gramaticales y semánticos. El estudio de estos componentes construye un vínculo entre la lingüística aplicada, la lingüística computacional y el procesamiento de lenguaje natural, y en este sentido, el PLN busca construir modelos que comprendan aspectos del lenguaje humano y automaticen tareas o extraigan información relevante (Sánchez, 2010).

En este trabajo nos enfocamos en una de las tareas del PLN, conocida como extracción de relaciones (ER), la cual permite el reconocimiento de patrones entre entidades nombradas (EN) tales como nombres propios, definiciones, abreviaturas, entre otras. De esta manera, a través del reconocimiento automático de las relaciones se reconoce cualquier tipo de relación semántica entre dos aspectos relevantes e identificables en el texto. Según Pawar *et al.* (2017), para construir una máquina de ER basada en aprendizaje supervisado se deben incluir aspectos morfológicos, sintácticos e incluso semántico-léxicos tales como la categoría gramatical (POS, del inglés *Part-Of-Speech*). Nasar *et al.* (2021) afirman que la tarea de ER requiere un análisis sintáctico basado en segmentación y lematización. Para los autores citados, la tarea de ER es descrita como dependiente del corpus de entrenamiento, donde el éxito de esta se relaciona con la calidad del conjunto de datos de entrenamiento (Pawar *et al.*, 2017; Nasar *et al.*, 2021). De forma similar, la tarea de ER se puede aplicar a diferentes dominios: química farmacéutica, salud, derecho, entre otros (Virmani *et al.*, 2017; Kumar, 2017).

A continuación, presentamos un ejemplo de la tarea de extracción de relaciones. El texto original: *La irritación del estómago ocurre en personas que usan Aspirina u otros medicamentos antiinflamatorios no esteroideos con regularidad*. La extracción indica que Aspirina e irritación son elementos de interés y que están presentes en una relación léxico-semántica. Esta relación es ADR (del inglés *Adverse Drug Reaction*), reacciones adversas a los medicamentos.

Castillo (2020) dice que para construir un modelo de ER es necesario contar con una base de conocimiento que puede ser el resultado de la recopilación de textos y la anotación de un experto o de un proceso automatizado que anote o etiquete texto, así como también de un modelo computacional capaz de tomar las muestras recopiladas y etiquetadas para encontrar patrones lingüísticos que revelen si una oración posee o no una relación semántica. Sin embargo, el autor señala la limitación actual de disponibilidad de conjuntos de muestras o corpus lingüísticos, en particular para algunos idiomas.

En este documento se propone la utilización de la traducción automática (TA) como una solución a las limitaciones de los corpus lingüísticos para el idioma español. Algunas de las herramientas de TA han alcanzado resultados de calidad, los cuales pueden ser útiles para traducir muestras de un idioma a otro. Según Carrino *et al.* (2020), la construcción de conjuntos de datos usando la TA no es nueva, pero su aplicación requiere de la experimentación y de las implementaciones necesarias. En este trabajo se implementa la TA como una estrategia para abordar el problema de la extracción de relaciones (ER) en el idioma español.

La revisión de literatura muestra que los conjuntos de entrenamiento para ER en español son escasos, costosos o no poseen etiquetas para el entrenamiento de modelos computacionales. En otros trabajos, para tratar este problema se han propuesto varios enfoques tales como modelos basados en el aprendizaje para varios idiomas (*cross-lingual learning*), aprendizaje profundo, extracción de relaciones abiertas (Ananthram *et al.*, 2020; Lin *et al.*, 2017; Mesquita *et al.*, 2013; Ni & Florian, 2019; Rodrigues & Branco, 2020; Verga *et al.*, 2015; Zhila & Gelbukh, 2013), entre otros.

Este artículo presenta la traducción automática del conjunto de entrenamiento conocido como: reACE (Hachey *et al.*, 2012); posteriormente, se entrena un modelo de extracción de relaciones genéricas y se presentan los resultados, los cuales muestran que la estrategia de TA funciona, pues los modelos ER en español alcanzan medidas de precisión, exhaustividad y valor-F comparables con las obtenidas por el modelo en lenguaje inglesa.

El documento está organizado de la siguiente forma: en la sección 1 se describen los fundamentos teóricos; en la sección 2 se explica el conjunto de datos y el proceso de traducción automático; en la sección 3, la concepción del conjunto de datos en el español y su utilización en el entrenamiento para la tarea de ER, y en la sección 4 se exponen las conclusiones.

Fundamentos teóricos

A continuación, se presentan la traducción automática, los modelos de ER y los dos servicios web utilizados para la traducción automática en línea, y en último término se describe la tarea de ER basada en el conjunto de datos traducidos.

Traductor de Google y DeepL

El traductor de Google (TG) (Google Translate es su nombre en inglés) es uno de los servicios en línea más populares para traducción automática de artículos, textos cortos, oraciones e incluso páginas web. Esta herramienta fue lanzada en 2006 y actualmente incorpora un traductor automático neuronal llamado Google Neural Machine Translation (GNTN) en inglés, basado en redes neuronales recurrentes (Wu et al., 2016; Yamada, 2019).

El traductor DeepL es una herramienta gratuita en línea similar a TG, lanzada en 2017 y que se ha convertido en referente durante los últimos años. Una de sus ventajas es soportar la traducción de más de 21 idiomas e incorporar técnicas de aprendizaje profundo tales como redes neuronales convolucionales. La calidad de las traducciones es comparable con TG. Así, TG y DeepL son los mejores exponentes de la traducción automática (Cheng, 2019; Hidalgo-Terreno, 2021).

En esta investigación se describen los errores encontrados en el proceso de traducción del conjunto de entrenamiento reACE, conjunto de datos creado para la tarea de extracción de relaciones (ER) en el idioma inglés. Además, se entrena un modelo computacional para la extracción de relaciones semánticas en español utilizando este corpus de muestras traducidas.

Extracción de relaciones (ER)

La ER es una tarea de extracción de información (EI) que reconoce relaciones semánticas entre entidades nombradas previamente definidas (Pawar et al., 2017). Es común describir la ER como un modelo de clasificación automática o de aprendizaje supervisado. El proceso inicia con la identificación de entidades tales como personas, lugares, organizaciones, proteínas, genes o enfermedades. Luego, el modelo computacional debe indicar las relaciones existentes entre las entidades nombradas. El modelo de ER reconoce las relaciones a través de los vínculos semánticos que se presentan entre las entidades nombradas en el texto. Algunos ejemplos de relaciones son: “A está casado con B”, “A es uno de los trabajadores de B”, “La A tiene un efecto adverso a consumirse con B”, donde A y B son las entidades nombradas o elementos de interés y presentan en el texto una relación semántica.

La ER ha sido ampliamente estudiada para el idioma inglés, utilizando enfoques supervisados o semisupervisados, donde existen datos de entrenamiento etiquetados como ACE 2005, ACE 2004, reACE, ADE, BioInfer (Gamallo & García, 2017; Guan et al., 2020; Smirnova & Cudré-Mauroux, 2018, Walker et al., 2006, Mitchell et al., 2005, Hachey et al., 2012, Gurulingappa et al., 2012, Pyysalo et al., 2007). De los

anteriores conjuntos, en este artículo se utiliza el conjunto para la extracción genérica de relaciones llamado reACE (Hachey et al., 2012).

Metodología

Para la traducción automática se tomó el conjunto de muestras reACE del idioma inglés; posteriormente se hizo la traducción automática (TA) desde los servicios en línea descritos en la sección anterior. Este proceso permitió identificar errores en la traducción, los cuales se catalogan en pre y posesición. Con los conjuntos de muestras en español se entrenan modelos computacionales para la extracción de relaciones semánticas (modelos ER). Finalmente, para la evaluación de los modelos se utilizan las métricas de precisión, exhaustividad y valor-F. Los modelos en el idioma inglés y español son comparados con las métricas anteriormente listadas.

El corpus reACE de Hachey *et al.* (2012) tiene etiquetadas relaciones semánticas entre personas, organizaciones, genes y proteínas. Con un total de 5984 oraciones en inglés, es un compilado de muestras de los conjuntos de entrenamiento ACE2004 y ACE2005. El conjunto está en formato XML (Extensible Markup Language, su nombre en inglés) a partir del cual se extraen secuencias que se describen más adelante. Para el trabajo se construye un analizador (*parser*), que obtiene las secuencias ordenadas en un archivo de texto plano.

Las fases llevadas a cabo en la experimentación del TA son las siguientes:

Primera o preprocesamiento del conjunto de muestras y creación de la lista de oraciones por traducir.

Segunda o transformación de listas a secuencias ordenadas: ($R_{\text{Inglés}}$, $E1_{\text{Inglés}}$, $E2_{\text{Inglés}}$, $S_{\text{Inglés}}$). La secuencia representa: la relación en la oración ($R_{\text{Inglés}}$), entre la entidad nombrada número 1 ($E1_{\text{Inglés}}$) y entidad nombrada número 2 ($E2_{\text{Inglés}}$), relación que se encuentra en la oración del idioma origen ($S_{\text{Inglés}}$).

Fase tres o traducción de muestras, el conjunto de todas las secuencias es la entrada para los servicios de TA en línea. Los traductores de Google y DeepL poseen servicios en web, que permiten la traducción de textos a través de una interfaz accesible desde internet. Para este trabajo el proceso de traducción se automatiza, pero necesita la verificación constante, dado que la mayoría de los errores en la traducción se deben a una mala interpretación del texto de partida; algunos considerados como errores de comprensión (Anastasopoulos, 2019), otros problemas relacionados con el contexto, la estructura de la oración y su gramática, la ortografía y el sentido mismo del texto (Bahr et al., 2020; Haque et al., 2020; Mikelenić & Tadić, 2020; Popović, 2020). Los traductores automáticos devuelven la versión traducida al idioma español del conjunto de secuencias. Cada secuencia obtenida tiene la forma: ($RE_{\text{Español}}$, $E1_{\text{Español}}$, $E2_{\text{Español}}$, $SE_{\text{Español}}$). La cual es su versión correspondiente a la secuencia en el idioma inglés.

Fase cuatro, análisis de errores de traducción e impacto en la tarea ER.

Fase cinco o preedición y posesición para corrección de errores. Para el caso de las correcciones de preedición se construye un subconjunto de secuencias con preedición y se lleva a cabo la fase tres. Finalmente, se obtiene un conjunto de secuencias reACE en español y sin errores de traducción.

En este trabajo no se incluye una posesición exhaustiva, ya que esto puede conllevar una edición o construcción de todo el corpus, con una inversión de tiempo mayor (Collantes et al., 2018). La Tabla 1 presenta los tipos de errores obtenidos en la evaluación de la traducción generada por las herramientas de TA.

Tabla 1. Errores considerados generados por la traducción automática

Error	Descripción	Abreviatura
Sin sentido	Traducción que carece de sentido o es absurda en el idioma español.	SS
Desarticulación sintáctica	Pérdida del significado de la oración o su contexto por segmentación, <i>tokenizado</i> u otro análisis sintáctico.	DS
Ortografía	Mala traducción por error de escritura de la palabra	OT
Traducción de entidad nombrada	Traducción de la entidad nombrada	TN
Terminología	Pérdida de la traducción de una palabra en su contexto por ser propia de una ciencia	TR
Adición	La herramienta de traducción agrega una palabra que no se encontraba en el texto de partida	AD
Omisión	La herramienta no traduce una palabra de forma injustificada	OM

Entrenamiento de un modelo para ER en español

En la literatura sobre modelos de ER se encuentran diferentes trabajos que muestran que los modelos basados en el aprendizaje supervisado obtienen mejores resultados que otros enfoques (Belinkov & Glass, 2019). Para el idioma español se han realizado algunos acercamientos utilizando técnicas de aprendizaje profundo, enfoques abiertos y multilingües (Torres et al., 2018; Zhila & Gelbukh, 2013). En este trabajo se entrenan tres modelos de ER basados en máquina de soporte vectorial (SVM). Esta técnica es utilizada frecuentemente en la literatura para construir modelos de ER (Zelenko, 2003; Bach & Saamer, 2007; Zhang, 2017; Torres et al, 2018). El primer modelo es una SVM-ER para el idioma inglés, el cual está basado en el conjunto de datos reACE original; el segundo, una SVM-ER, basado en el conjunto de datos reACE traducido automáticamente; y el tercero, una SVM-ER, basado en el conjunto de datos

reACE traducido automáticamente, verificados los errores de traducción y preeditado. Los modelos SVM-ER en inglés y español utilizan las características que se presentan en la Tabla 2.

Tabla 2. Características consideradas para los modelos de aprendizaje automático SVM-ER inglés y SVM-ER español

Categoría	Característica
Características sintácticas	Categoría gramatical
	Árbol sintáctico
Características de las palabras	Palabra en mayúscula
	Palabra etiquetada en corpus
Características de las oraciones	Persona del verbo en la oración (1. Primera, 2. Segunda, 3. Tercera, 0. Cualquier otro)
	Modo del verbo en la oración (1. Infinitivo, 2. Gerundio, 3. Participio, 0. Cualquier otro)

Resultados y discusión

La Tabla 3 muestra la frecuencia de errores de traducción de cada una de las herramientas de traducción. Los errores de TA se presentan a través de la frecuencia de aparición del tipo de error. De acuerdo con los errores considerados, estos fueron tratados como errores de preedición. Los resultados muestran que de las 5984 oraciones del conjunto reACE, en 465 oraciones se encontraron problemas de traducción. Los errores más comunes son por ortografía, terminología y omisión.

Tabla 3. Frecuencia de errores de traducción de cada una de las herramientas de traducción

Errores TG		Errores DeepL	
SS	1	SS	1
DS	2	DS	1
OT	42	OT	30
TN	5	TN	4
TR	8	TR	6
AD	1	AD	2
OM	3	OM	9
Total	62	Total	53

La calidad de las muestras en el idioma español influye en el aprendizaje del modelo, debido a que los modelos de SVM-ER se estiman a partir de un proceso de entrenamiento. El proceso de predicción es no exhaustivo y mantiene el uso del lenguaje y su naturaleza, se corrigen errores que pueden impactar la tarea como la traducción de las entidades nombradas o como la ortografía que impide que las oraciones conserven su sentido. Algunos ejemplos de correcciones realizadas al corpus se listan a continuación:

- La abreviaturas y acrónimos fueron reemplazados usando un diccionario (*thesaurus*) del idioma inglés, por ejemplo, palabras como CEO, fueron convertidas a *Chief Executive Officer*.
- Las contracciones que son ampliamente utilizadas en inglés fueron expandidas; por ejemplo, palabras como *He <d*, *You <re* o *She <s*. Estas contracciones no fueron traducidas correctamente y por lo tanto se utiliza la expresión completa en inglés. Así, fueron expandidas directamente en el texto origen a palabras como *He had*, *You are* o *She is*.
- En inglés es común utilizar el carácter (-) para las palabras compuestas, el símbolo indica que no deben ser separadas o deben leerse juntas, de otra forma perdería el sentido en la oración. En esta investigación se realizó la separación de este tipo de palabras y caracteres. Por ejemplo, *wheelchair-bound*, que sería traducido como «silla de ruedas-atada», fue automáticamente traducida como «en silla de ruedas».

Como consecuencia, los errores de ortografía (OT) se reducen en 40 % para el traductor de Google y en 50 % para el traductor DeepL. Esto sugiere que la revisión o predicción de los textos en el idioma origen es una tarea importante. A continuación, se listan otros errores que pueden impactar el rendimiento del proceso de entrenamiento del modelo ER y a los que no se dio solución en el corpus de español obtenido:

- La traducción errónea del verbo *to be*, que puede tener un impacto en el proceso de ingeniería de características para la construcción de un modelo ER.
- La modificación de los artículos indeterminados o indefinidos como: *un*, *uno*, *unos*, *unas*. Que se considera como un problema de adición en la traducción. Este cambio puede impactar el contexto de la oración o referencia a una Entidad Nombrada.

Aunque la traducción al español suele presentar errores de terminología y de pérdida del sentido en la oración, ya sea por adición u omisión (Pastor, 2018), las herramientas de traducción seleccionadas para este trabajo permitieron traducir el corpus reACE, manteniendo el sentido y el significado de las palabras de las 5519 oraciones. Esto sugiere que las muestras pueden ser utilizadas para el entrenamiento de una tarea específica de procesamiento de lenguaje natural (PLN) o extracción de información (EI) en el idioma español.

Para el desarrollo de la tarea de extracción semántica de relaciones se presentan tres modelos de acuerdo con su fuente de entrenamiento, i) el corpus de datos de reACE en su versión en inglés, ii) la TA al español y iii) la TA al español con preedición de errores. Los modelos predictivos se construyen en una técnica tradicional para la solución de la tarea de extracción de relaciones, las máquinas de soporte vectorial (SVM, por sus siglas en inglés) (Zelenko, 2003; Bach & Sameer, 2007; Zhang et al., 2017; Torres et al., 2018).

Las SVM se construyeron mediante un proceso estándar de entrenamiento y evaluación de un modelo de aprendizaje supervisado (para más información ver Kramer et al., 2016). La Tabla 4 presenta los resultados del modelo SVM, para el conjunto reACE 2004 y reACE 2005; así se obtiene una perspectiva completa de todo el conjunto de datos reACE (Hachey et al., 2012). En la Tabla 4 se exponen los valores de las medidas de precisión (P), exhaustividad (E) y valor-F (F1) (precision, recall y F-value, por sus nombres en inglés). Los valores relevantes para la comparación de las métricas se marcan con **negrita**.

Tabla 4. Resultados del modelo SVM

	reACE 2004			reACE 2005		
	P	E	F1	P	E	F1
Inglés	81 %	47 %	60 %	73 %	37 %	49 %
Español	81 %	39 %	53 %	75 %	33 %	47 %
Español (preedición)	80 %	47 %	59 %	73 %	36 %	48 %

Las métricas para los modelos en español son comparables a las obtenidas para el idioma inglés, lo que indica que el conjunto de datos y las muestras traducidas son representativas en los dos idiomas. Adicionalmente, las métricas muestran que el modelo tiene un excelente desempeño para la tarea de extracción de relaciones semánticas entre entidades nombradas para el idioma español.

De esta manera, la métrica de precisión (P) mide la calidad del modelo en la clasificación de relaciones; los tres modelos oscilan entre el 73 y el 81 %, se destaca el modelo para el idioma español sin preedición. Y la métrica de exhaustividad (E) mide la cantidad de relaciones que es capaz de extraer; los tres modelos oscilan entre el 33 y el 47 %, y se destaca que los modelos para el idioma español e inglés obtienen valores similares. El valor-F es una métrica que combina las anteriores, precisión y exhaustividad. Los valor-F del modelo para el idioma español son comparables con el modelo para el inglés.

Conclusiones

En este artículo se presenta la traducción de un conjunto de muestras para el entrenamiento de modelos computacionales que realicen la tarea de extracción de relaciones semánticas entre entidades nombradas (ER). Se utiliza el corpus reACE traducido, como conjunto etiquetado para la tarea de ER en el idioma español. Este conjunto de muestras automáticamente traducidas permitió construir un modelo para la tarea ER en el idioma español, lo que muestra la eficiencia y el potencial que tiene la TA, así como también las facilidades de automatización con las TA en línea. El potencial de este trabajo deja ver que la TA es aplicable a otros corpus de datos propios de otras tareas de PLN, en los cuales no hay conjuntos de datos en el idioma español, pero sí existen para el idioma inglés.

Las herramientas de Google y DeepL muestran que tienen un buen rendimiento en la comprensión e interpretación de las oraciones para su posterior traducción. A pesar de su alta tasa de precisión en la traducción, ambas herramientas tuvieron errores al traducir correctamente abreviaturas, términos y artículos. La mayoría de los errores fueron tratados con preedición no exhaustiva, pero existe una gran variedad de imprecisiones relacionadas con las acepciones de algunas palabras y el contexto donde son utilizadas, falso sentido, contrasentido, anglicismos, entre otras. Sin embargo, para el conjunto reACE estos errores tienen una ocurrencia mínima.

El modelo SVM implementado para este trabajo muestra que la técnica de extracción de relaciones descrita ofrece buenos resultados, lo que implica la viabilidad para desarrollar otras tareas de PLN basado en ER, como por ejemplo resumidores automáticos. Así como también poner en marcha tecnologías basadas en PLN para el idioma español. Esto muestra que el corpus traducido permite la extracción de relaciones semánticas entre entidades nombradas de una forma eficiente. Los resultados dejan ver que las métricas de evaluación están cercanas entre los dos idiomas, es decir, reACE es un conjunto de datos que posee una muestra relevante para la tarea de ER en el idioma español. Sin embargo, el rendimiento de los modelos de ER puede tener mejoras si se aplica ingeniería de características o se efectúa una revisión exhaustiva de características léxico-semánticas propias del idioma español; por otra parte, se puede elaborar un modelo de aprendizaje profundo. Estas mejoras y un modelo con aprendizaje profundo se consideran como trabajo futuro.

Referencias

- Ananthram, A., Allaway, E., & McKeown, K. (2020). Event Guided Denoising for Multilingual Relation Learning. *arXiv preprint: arXiv:2012.02721*. <https://doi.org/10.18653/v1/2020.coling-main.131>
- Anastasopoulos, A. (2019). An Analysis of Source-Side Grammatical Errors in NMT. *arXiv preprint: arXiv:1905.10024*.

- Bach, N., & Sameer, B. (2007). *A Survey on Relation Extraction*. Language Technologies Institute, Carnegie Mellon University 178. https://doi.org/10.1007/978-981-10-7359-5_6
- Bahr, R. H., Lebby, S., & Wilkinson, L. C. (2020). Spelling Error Analysis of Written Summaries in an Academic Register by Students with Specific Learning Disabilities: Phonological, Orthographic, and Morphological Influences. *Reading and Writing*, 33(1), 121-142. <https://doi.org/10.1007/s11145-019-09977-0>
- Belinkov, Y., & Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7, 49-72. https://doi.org/10.1162/tacl_a_00254
- Carrino, C. P., Costa-Jussà, M. R., & Fonollosa, J. A. (2020). Automatic Spanish Translation of SQuAD Dataset for Multi-lingual Question Answering. In *Proceedings of the 12th Language Resources and Evaluation Conference* (5515-5523).
- Castillo, M. N. (2020). Corpus Básico del Español de Chile ©: metodología de procesamiento y análisis. *Lexis*, 44(2), 483-523. <https://doi.org/10.18800/lexis.202002.004>
- Cheng, Y. (2019). Neural Machine Translation. In *Joint Training for Neural Machine Translation* (1-10). Springer. https://doi.org/10.1007/978-981-32-9748-7_1
- Collantes, C., Mallo, J., Parra, C., Quiñones, H. & Serrano, R. (2018). Pásate al lado oscuro: ventajas de la traducción automática para el traductor profesional. *La linterna del Traductor*, 17, 33-39.
- Gamallo, P., & García, M. (2017). LinguaKit: Uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, 9(1), 19-28. <https://doi.org/10.21814/lm.9.1.243>
- Guan, H., Li, J., Xu, H., & Devarakonda, M. (2020). Robustly Pre-trained Neural Model for Direct Temporal Relation Extraction. *arXiv preprint: arXiv:2004.06216*. <https://doi.org/10.1109/ICHI52183.2021.00090>
- Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-Related Adverse Effects from Medical Case Reports. *Journal of Biomedical Informatics*, 45(5), 885-892. <https://doi.org/10.1016/j.jbi.2012.04.008>

- Hachey, B., Grover, C., & Tobin, R. (2012). Datasets for Generic Relation Extraction. *Natural Language Engineering*, 18(1), 21–59. <http://dx.doi.org/10.1017/S1351324911000106>
- Haque, R., Hasanuzzaman, M., & Way, A. (2020). Analysing Terminology Translation Errors in Statistical and Neural Machine Translation. *Machine Translation*, 34(2), 149-195. <https://doi.org/10.1007/s10590-020-09251-z>
- Hidalgo-Tertero, C. M. (2021). Google Translate vs. DeepL. *MonTI. Monografías de Traducción e Interpretación*, 154-177.
- Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies. Studies in Big Data, vol 20* (pp. 45-53). Springer, Cham. https://doi.org/10.1007/978-3-319-33383-0_5
- Kumar, S. (2017). A Survey of Deep Learning Methods for Relation Extraction. *arXiv preprint: arXiv:1705.03645*.
- Lin, Y., Liu, Z., & Sun, M. (2017). Neural Relation Extraction with Multi-Lingual Attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 34–43. Association for Computational Linguistics. <http://dx.doi.org/10.18653/v1/P17-1004>
- Mesquita, F., Schmidek, J., & Barbosa, D. (2013). Effectiveness and Efficiency of Open Relation Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 447-457. Association for Computational Linguistics.
- Mikelenić, B., & Tadić, M. (2020). Building the Spanish-Croatian Parallel Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3932-3936. European Language Resources Association.
- Mitchell, A., Strassel, S., Huang, S., & Zakhary, R. (2005). Ace 2004 Multilingual Training Corpus. *Linguistic Data Consortium, Philadelphia*, 1, 1-1.
- Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named Entity Recognition and Relation Extraction: State-of-the-Art. *ACM Computing Surveys (CSUR)*, 54(1), 1-39. <https://doi.org/10.1145/3445965>
- Ni, J., & Florian, R. (2019). Neural Cross-Lingual Relation Extraction Based on Bilingual Word Embedding Mapping. *arXiv preprint: arXiv:1911.00069*. <https://doi.org/10.18653/v1/D19-1038>
- Pastor, G. C. (2018). Laughing One's Head Off in Spanish Subtitles: A Corpus-Based Study on Diatopic Variation and Its Consequences for Translation1. *Fraseología, Diatopía y Traducción/Phraseology, Diatopic Variation and Translation*, 17, 32. <https://doi.org/10.1075/ivitra.17.03cor>

- Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation Extraction: A Survey. *arXiv preprint: arXiv:1712.05191*.
- Popović, M. (2020). Relations Between Comprehensibility and Adequacy Errors in Machine Translation Output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, (pp. 256-264). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.conll-1.19>
- Pysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: A Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics*, 8(1), 50. <https://doi.org/10.1186/1471-2105-8-50>
- Rodrigues, J., & Branco, A. (2020). Argument Identification in a Language Without Labeled Data. In *International Conference on Computational Processing of the Portuguese Language*, (pp. 335-345). https://doi.org/10.1007/978-3-030-41505-1_32
- Sánchez, A. (2010). Traducción automática, corpus lingüísticos y desambiguación automática de los significados de las palabras. En R. Rabadán, M. Fernández & T. Guzmán (coords.), *Lengua, traducción, recepción: en honor de Julio César Santoyo*, vol. 1 (pp. 555-587). Universidad de León, Área de Publicaciones.
- Smirnova, A., & Cudré-Mauroux, P. (2018). Relation Extraction Using Distant Supervision: A Survey. *ACM Computing Surveys (CSUR)*, 51(5), 1-35. <https://doi.org/10.1145/3241741>
- Torres, J. P., De Piñérez Reyes, R. G., & Bucheli, V. A. (2018). Support Vector Machines for Semantic Relation Extraction in Spanish Language. *Colombian Conference on Computing*, 326-337. https://doi.org/10.1007/978-3-319-98998-3_26
- Verga, P., Belanger, D., Strubell, E., Roth, B., & McCallum, A. (2015). Multilingual Relation Extraction Using Compositional Universal Schema. *arXiv preprint: arXiv:1511.06396*. <https://doi.org/10.18653/v1/N16-1103>
- Virmani, C., Pillai, A., & Juneja, D. (2017). Extracting Information from Social Networks Using NLP. *International Journal of Computational Intelligence Research*, 13(4), 621-630.
- Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium. <https://doi.org/10.35111/mwxc-vh88>

- Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Machery, W., Krikun, M. et al. (2016). Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv preprint: arXiv:1609.08144*.
- Yamada, M. (2019). The Impact of Google Neural Machine Translation on Post-Editing by Student Translators. *The Journal of Specialised Translation*, 31, 87-106.
- Zelenko, D., Chinatsu, A., and Anthony, R. (2003, Feb.). Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3, 1083-1106. <https://dl.acm.org/doi/10.3115/1118693.1118703>
- Zhang, Q., Mengdong C., and Lianzhong, L. (2017). A Review on Entity Relation Extraction. In *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. IEEE. <https://doi.org/10.1109/ICMCCE.2017.14>
- Zhila, A., & Gelbukh, A. (2013). Comparison of Open Information Extraction for Spanish and English. *Computational Linguistics and Intellectual Technologies*, 12(1), 794-802.