Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 31-38 recibido 31-03-2017 revisado 15-05-2017 aceptado 22-05-2017

Similitud español-inglés a través de word embeddings

Spanish-English similarity through word embeddings

Fernando Enríquez, Fermín Cruz, F. Javier Ortega, José A. Troyano
Universidad de Sevilla
Escuela Técnica Superior de Ingeniería Informática, Av. Reina Mercedes s/n
{fenros, fcruz, javierortega,troyano}@us.es

Resumen: En este trabajo hemos afrontado la tarea de similitud de textos multilingüe mediante representaciones vectoriales de las palabras. Hemos experimentado con varias colecciones de textos con pares de frases en español e inglés, adaptando dos técnicas basadas en *word embeddings* que han mostrado su eficacia en la similitud de textos monolingüe: la agregación de vectores y el alineamiento. La agregación permite construir una representación vectorial de un texto a partir de los vectores de las palabras que lo componen, y el algoritmo de alineamiento aprovecha los *word embeddigs* para decidir el emparejamiento de palabras de los dos textos a comparar. En el proceso se han utilizado dos estrategias distintas: usar traductores automáticos para poder aplicar directamente las técnicas de similitud monolingüe, y aplicar una técnica de transformación de modelos para trasladar los vectores de un idioma al espacio del otro. Las dos estrategias han funcionado razonablemente bien por separado, y los resultados mejoran cuando las salidas de los dos tipos de sistemas se integran mediante técnicas de *ensemble learning*.
Palabras clave: Similitud bilingüe, *word embeddings*, alineamiento de textos, transformación de modelos

Abstract: In this paper we have faced the cross-lingual text similarity task using vector representations of words. We have experimented with several collections of texts with pairs of sentences in Spanish and English, adapting two techniques based on word embeddings that have shown their effectiveness in the similarity of monolingual texts: vector aggregation and vector-based text alignment. The aggregation allows to construct a vector representation of a text from the vectors of the words that compose it, and the algorithm of alignment takes advantage of word embeddigs to decide the pairing of words of the two texts to be compared. Two different strategies have been used in the process: using automatic translators to be able to directly apply monolingual similarity techniques, and applying a model transformation technique to translate the vectors of one language into the space of the other. Both strategies have worked reasonably well separately, and the results improve when the outputs of the two types of systems are integrated by means of ensemble learning techniques.
Keywords: Cross-lingual similarity, word embeddings, text alignment, model transformation

ISSN 1135-5948

1 Introducción

Las técnicas de *feature learning* están adquiriendo cada vez más relevancia por lo que aportan a la hora de aplicar algoritmos de aprendizaje automático sobre información no estructurada. Se entiende por *feature learning* (también llamado *representation learning*) el proceso de obtención de atributos (*features*) de forma automática desde datos no estructurados, para que puedan posteriormente ser usados como entrada a algoritmos de aprendizaje automático. Son muchos los dominios en los que este tipo de técnicas son de utilidad: procesamiento de imágenes, vídeos, audio, series temporales, lenguaje natural, etc.

Las protagonistas indiscutibles del *feature learning* en el procesamiento del lenguaje natural son las técnicas de *word embeddings* (Mikolov et al., 2013), que proporcionan mecanismos para obtener, a partir de una palabra, una representación vectorial en un espacio continuo con números reales. La potencia de esta transformación de palabras a vectores reside en que dichos vectores capturan relaciones de similitud semántica entre palabras. Además, estas relaciones se calculan de forma totalmente no supervisada en base a los contextos en los que las palabras son usadas en grandes colecciones de textos. Se aplica la idea de que el significado de una palabra viene determinado por su contexto (Firth, 1957). Cuando los modelos se entrenan sobre colecciones suficientemente grandes como la Wikipedia los resultados son realmente sorprendentes, como es el caso del famoso ejemplo de la siguiente ecuación de vectores $king - man + woman \approx queen$.

Este tipo de representaciones se han utilizado, por ejemplo, para la ordenación temporal de eventos (Saquete y Navarro-Colorado, 2017), la identificación de grupos de palabras relacionadas semánticamente (Kovatchev, Salamó, y Martí, 2016), la inducción de la polaridad de palabras de opinión (Pablos, Cuadros, y Rigau, 2015; López-Solaz et al., 2016) e incluso la detección de la ironía (López y Ruiz, 2016). Otra de las tareas en las que se han aplicado las técnicas de *word embeddings* es el cálculo de la similitud de textos. Los sistemas de similitud de textos son de gran ayuda para distintas tareas PLN. Se puede distinguir entre dos tipos de cálculo de similitud: a nivel de palabras y a nivel de textos. En ambas tiene cabida la utilización de *word embeddings* pero, dada su mayor complejidad, es en la segunda donde hay más margen de aplicación. Para calcular la similitud de textos, se han utilizado *word embeddings* con éxito con dos estrategias distintas. Por un lado, agregando los vectores de las palabras de los textos para que puedan ser comparados mediante algún tipo de distancia. Y por otro lado, utilizándolos como información de entrada para algoritmos de alineamiento, que es otra de las estrategias clásicas usadas en los sistemas de cálculo de similitud.

Uno de los aspectos más atractivos de las técnicas de *word embeddings* es que son no supervisadas. Sólo es necesario disponer de un gran volumen de textos en un idioma para construir los modelos y, a partir de ellos, podremos beneficiarnos de las relaciones semánticas entre palabras que se pueden derivar de la comparación de sus correspondientes vectores. Pero esta representación vectorial aún puede dar más de sí. No solo es útil para obtener atributos significativos desde palabras, sino que abre la puerta a la conexión entre palabras de distintos idiomas. Es lo que se conoce como *cross lingual embeddings*. Estas técnicas plantean la definición de un espacio de representación común a varios idiomas, en el que modelos de distintos idiomas pueden proyectar en puntos cercanos palabras que tengan significados similares.

La motivación de este trabajo es la de utilizar la información que proporcionan los *word embeddings* como base para calcular sistemas de similitud de textos de dos idiomas distintos. Nuestra experimentación se ha centrado en la pareja de idiomas español-inglés. La idea es intentar aprovechar la eficacia que han demostrado estas técnicas en la tarea de cálculo de similitud por un lado, y en la definición de espacios vectoriales comunes a distintos idiomas por otro. Hemos experimentado con dos enfoques distintos. En primer lugar hemos utilizado traductores automáticos para traducir las frases de un idioma a otro y así poder aplicar directamente técnicas ya probadas para calcular similitud de textos en un escenario monolingüe. En segundo lugar hemos aplicado una técnica de transformación de modelos para trasladar los vectores de un idioma al espacio del otro, y poder así calcular distancias entre palabras de los dos idiomas. Las dos estrategias han funcionado razonablemente bien por separado, y los resultados mejoran sensiblemente cuando

se aplican técnicas de *ensemble learning* para integrar las salidas de los dos tipos de sistemas. Con estos esquemas de combinación se obtienen resultados muy competitivos (una correlación de Pearson de entre 0,420 y 0,899 para las distintas colecciones de textos con las que hemos experimentado), lo que muestra la eficacia de los *word embeddings* a la hora de capturar la relación entre palabras de distintos idiomas.

Para evaluar nuestros sistemas hemos recurrido tanto a recursos ya preparados para la tarea que están disponibles públicamente, como a la adaptación, por nuestra parte, de otros recursos de tareas cercanas. Los nuevos corpus desarrollados han sido publicados para que estén a disposición de la comunidad investigadora [1].

El resto del artículo se organiza de la siguiente forma: la sección 2 describe la tarea abordada y la metodología seguida para construir los corpus de evaluación, la sección 3 presenta las distintas estrategias de cálculo de similitud implementadas, la sección 4 incluye los resultados experimentales y, por último, en la sección 5 se extraen las conclusiones y se plantean algunas líneas de trabajo futuro.

2 La tarea

En esta sección detallaremos la tarea que hemos afrontado, definiendo los objetivos y explicando la procedencia y contenidos de los recursos que se han utilizado, que incluyen tanto conjuntos de datos ya preparados para la tarea como la adaptación de otros recursos "cercanos".

2.1 Definición

El objetivo del sistema que se ha desarrollado es determinar el nivel de interrelación existente entre dos frases desde el punto de vista semántico. Llevar a cabo esta tarea con éxito es complicado debido a los múltiples factores que influyen en el significado de una frase. La manera en que se relacionan las formas léxicas con los pronombres, el uso de sinónimos o hiperónimos, los nexos oracionales que implican refuerzo, contradicción o matización, etc. son ejemplos de algunos de esos factores que dificultan la tarea desde el punto de vista lingüístico. La propia ambigüedad y versatilidad del lenguaje puede dar lugar a interpretaciones diferentes, por lo que es difícil obtener un marco de evaluación para determinar la precisión de los sistemas que abordan esta tarea. En nuestro caso nos hemos basado en uno de los foros internacionales más importantes en este ámbito, como son las conferencias SemEval (*Semantic Evaluation*) que se celebran desde 1998 en diferentes ciudades del mundo (anualmente desde 2012). Con ellas se persigue estudiar y analizar la naturaleza del significado en el lenguaje, lo cual se lleva a cabo proponiendo tareas o retos de diferente índole. Algunas se centran en la similitud entre palabras o entre elementos de distinto nivel (palabras, frases, documentos,...), aunque en este caso nos centraremos en la tarea *Semantic Textual Similarity* (STS), la cual aparece integrada en SemEval desde 2012. Los sistemas desarrollados para esta tarea devuelven como salida un nivel de equivalencia semántica en un rango entre 0 y n para cada pareja de frases que reciben como entrada. El valor más alto estará asociado a una pareja de frases que comparten el mismo significado, mientras que el valor 0 se asociará a un par de frases con significado totalmente diferente.

Para este trabajo hemos considerado una dificultad añadida, que consiste en procesar pares de frases en diferentes idiomas, concretamente español e inglés. Siguiendo el mismo esquema de niveles de equivalencia entre frases antes mencionado, vemos en la Tabla 1 tres ejemplos con diferentes grados de similitud.

En el siguiente apartado se explican los recursos que se han construido para poder llevar a cabo la experimentación.

2.2 Recursos para la evaluación

Para poder construir y evaluar un sistema de aprendizaje automático supervisado que resuelva la tarea STS, necesitamos un conjunto de datos de entrenamiento en el que los pares de frases hayan sido previamente clasificados. En concreto hemos recurrido tanto a recursos ya preparados para la tarea que están disponibles públicamente, como a la adaptación propia de otros recursos para que puedan servir de entrenamiento y prueba en la tarea STS multilingüe.

Los datos de las ediciones pasadas de SemEval han sido nuestro punto de partida[2]. En concreto comenzamos con la adaptación de los conjuntos de datos en español con

[1] http://www.lsi.us.es/~fermin/index.php/Datasets

[2] http://ixa2.si.ehu.es/stswiki/index.php

Valor	Frases
3.8	Esta licencia fue creada originalmente por Richard Stallman fundador de la Free Software Foundation (FSF) para el proyecto GNU (GNU project).
	The GPL license was created by Richard Stallman in 1989 to protect programs released as part of the GNU project.
2	La "Región de Los Lagos" es una de las quince regiones en las que se encuentra dividido Chile.
	The Region of the Lakes is a region of Chile, created in 1974, by Decree Law No. 575, in a process known as regionalization.
0	Como en la economía de todos los países europeos, el sector terciario o sector servicios es el que tiene un mayor peso.
	Of the crustaceans, the shrimp, and of the mollusks the squid and the octopus.

Tabla 1: Ejemplos de pares de frases para cada nivel de similitud

frases extraídas de Wikipedia para la tarea STS de SemEval 2014 (Agirre et al., 2014) y 2015 (Agirre et al., 2015). Para poder utilizarlos en nuestra tarea multilingüe hemos traducido manualmente la segunda frase de cada pareja del *dataset* al inglés, obteniendo finalmente la versión español-inglés.

Una vez obtenidos los conjuntos de datos multilingües derivados de la tarea STS de SemEval, consideramos la posibilidad de construir un nuevo conjunto de datos de procedencia distinta que nos permitiese comparar los resultados con los anteriores. Para encontrar frases equivalentes, una posible fuente de datos son los corpus paralelos, mientras que para hallar pares de frases con similitud cero se podrían seleccionar frases de dominios diferentes, pero la dificultad radica en generar pares de frases de los niveles intermedios de similitud. Eso nos hizo fijarnos en otra tarea relacionada con el significado de las frases, como es el reconocimiento de *textual entailment*. Esta tarea consiste en determinar si de una de las frases se puede inferir que la otra es cierta o no. En concreto, nos hemos apoyado en la tarea 8 de SemEval 2012 (Agirre et al., 2012) que consistió precisamente en clasificar pares de frases en español e inglés en cuatro tipos diferentes, en función de si existe una relación de *entailment* entre ellas y su direccionalidad (ver tabla 2). Para adaptar este conjunto de datos (Negri et al., 2011) a la tarea de similitud que nos ocupa, consideramos la siguiente asociación entre tipo de *entailment* y valor de similitud: Bidirectional $\Rightarrow$ 3, *Forward* $\Rightarrow$ 2, *Backward* $\Rightarrow$ 2 y *No entailment* $\Rightarrow$ 1. En este caso contamos con menos niveles de similitud que en la tarea STS original, destacando especialmente la ausencia del nivel cero, ya que todos los pares contienen información relacionada aunque no exista *entailment*.

Por último, hemos tomado los datos facilitados para la tarea *Cross Lingual STS* de la edición 2016 (Agirre et al., 2016) de SemEval, seleccionando exclusivamente los pares español-inglés y adaptándolos al formato utilizado hasta ahora. En esta edición de 2016 se encuentran disponibles dos conjuntos de datos diferentes, uno basado en fuentes de noticias multilingües (*news*) y otro con datos provenientes de fuentes diversas (*multi*).

Para todos los conjuntos de datos mencionados, provenientes tanto de la tarea STS como de la tarea TE, hemos obtenido también una traducción automática[3] de cada frase, que será utilizada en algunos de los experimentos. El resultado final es un formato de cuatro columnas para cada conjunto de datos: frase 1 en español, frase 2 en inglés, traducción automática de la frase 1 al inglés y traducción automática de la frase 2 al español.

Resumiendo, estos son los recursos que se han utilizado para la fase de experimentación y su origen:

- *SE-12-TE*: Adaptación de los datos de la tarea 'Cross-lingual textual entailment' de SemEval 2012.
- *SE-14-STS*: Traducción manual de los datos de la tarea 'Semantic Textual Similarity' de SemEval 2014.
- *SE-15-STS*: Traducción manual de los datos de la tarea 'Semantic Textual Similarity' de SemEval 2015.

[3]Google Translate: `https://translate.google.com/`

Entailment	Frases
Bidireccional	Mozart nació en la ciudad de Salzburgo Mozart was born in Salzburg
Forward	Mozart nació el 27 de enero de 1756 en Salzburgo Mozart was born in 1756 in the city of Salzburg
Backward	Mozart nació en la ciudad de Salzburgo Mozart was born on 27th January 1756 in Salzburg
No entailment	Mozart nació el 27 de enero de 1756 en Salzburgo Mozart was born to Leopold and Anna Maria Pertl Mozart

Tabla 2: Ejemplos de pares de frases para cada tipo de *textual entailment*

- *SE-16-STS-news* y *SE-16-STS-multi*: Datos de la tarea 'Cross Lingual STS' de SemEval 2016.

3 *Métricas de similitud*

En esta sección explicaremos en detalle las métricas utilizadas para medir la similitud entre dos frases dadas, todas ellas basadas en última instancia en los modelos de palabras de *word embeddings*. En primer lugar, nos centraremos en la similitud monolingüe, para luego abordar las dos propuestas desarrolladas en relación a la similitud multilingüe, a través del uso de traductores automáticos y mediante la transformación de modelos, respectivamente.

3.1 Similitud monolingüe

A la hora de calcular la similitud monolingüe a nivel de frases, el primer paso es definir el proceso mediante el cual aplicamos los modelos a nivel de palabras de *word embeddings*. En nuestro caso proponemos dos mecanismos:

- **Agregación:** obtenemos la similitud entre frases construyendo un vector para cada una de ellas, resultante de calcular la media de los vectores de las palabras que conforman la frase. Para obtener el valor de la similitud entre los dos vectores hemos realizado experimentos aplicando dos métricas: la distancia del coseno y la distancia euclídea.
- **Alineamiento:** realizamos un alineamiento de las frases haciendo corresponder a cada palabra de una frase la palabra más similar (según el modelo de *word embeddings*) de la otra frase, y viceversa. Una vez alineadas, calculamos la media de esas similitudes. Se puede leer una explicación más detallada en (López-Solaz et al., 2016).

3.2 Traducción automática

Para resolver el cálculo de la similitud entre textos de distintos idiomas, nuestra primera aproximación consiste en traducir automáticamente cada texto al idioma contrario, obteniendo una pareja de frases en cada idioma (una frase en su idioma original y otra resultado de la traducción). De esta forma, dadas dos frases, s_{en} y s_{sp}, y sus respectivas traducciones, $trad_{en \to sp}$ y $trad_{sp \to en}$, calculamos la similitud aplicando las métricas vistas en la sección anterior a las representaciones vectoriales de las frases en el mismo idioma.

3.3 Transformación de modelos

Nuestra segunda aproximación a la hora de calcular la similitud entre textos de distintos idiomas se basa en la idea propuesta en (Mikolov, Le, y Sutskever, 2013). En ese trabajo, los autores parten de la intuición de que conceptos similares, expresados en distintos idiomas, deben tener distribuciones geométricas similares en un espacio vectorial. De esta forma, dados dos modelos entrenados uno en cada lenguaje, es posible aprender una matriz de transformación lineal entre ambos modelos (ver Figura 1).

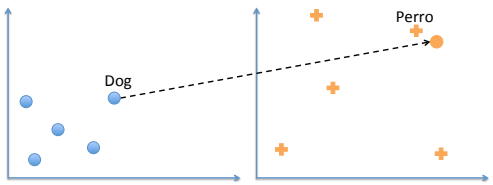

Figura 1: Transformación de modelos

Dado un conjunto de pares de palabras, $\{x_i, z_i\}$, donde x_i es la representación vectorial de la palabra i del idioma X y z_i la representación vectorial de la palabra i del idioma Z, buscamos una matriz de transformación, W, que aproxime W_{x_i} a z_i. Dicha matriz se

calcula a través del siguiente problema de optimización:

$$min_W \sum_{i=1}^{n} ||W_{x_i} - z_i||^2 \quad (1)$$

que se ha resuelto mediante gradiente descendente estocástico.

De esta forma obtenemos dos matrices de transformación, una de español a inglés y otra de inglés a español. Dadas dos palabras en esos idiomas, w_{en} y w_{sp}, nos basamos en la representación vectorial de cada una y en las matrices de transformación para obtener proyecciones de dichas palabras hacia los modelos vectoriales contrarios.

Contamos así con dos representaciones vectoriales para las palabras de cada frase: una en el modelo original y otra, aplicando la transformación de modelos, en el modelo correspondiente al otro idioma.

Para calcular la similitud entre las frases, aplicamos las técnicas de agregación y alineamiento vistas en la sección 3.1 de la siguiente forma:

- **Agregación:** se aplica el mecanismo de agregación a las dos parejas de frases proyectadas sobre un mismo modelo, obteniendo dos posibles valores de similitud para las distancias del coseno y euclídea, según se proyecten las frases del español al inglés, o viceversa.
- **Alineamiento:** obtenemos un único valor, de manera que se proyectan las palabras del español al inglés o viceversa según se estén buscando las similitudes mayores en un sentido o el contrario.

4 Experimentación

Nuestro objetivo principal al diseñar los experimentos es la comparación de los dos métodos de similitud propuestos (agregación y alineamiento) y sus variantes adaptadas a la tarea de similitud multilingüe (usando traducción automática o matrices de transformación). Con este fin, el desarrollo experimental consiste en el cómputo de las métricas de similitud para cada uno de los conjuntos de datos, y la evaluación mediante validación cruzada de las predicciones de un modelo regresional de tipo *Random Forest*[4]. Aunque podríamos haber optado por un esquema basado en entrenamiento y evaluación a partir de conjuntos de datos distintos, hemos decidido usar validación cruzada por dos razones: en primer lugar, facilita el análisis acerca de la eficacia de las distintas métricas y métodos propuestos, más allá de las diferencias inherentes a los distintos corpus; en segundo lugar, usar un esquema basado en entrenamiento y evaluación implicaría un gran número de combinaciones posibles entre los distintos conjuntos de datos, lo cual complica el análisis de los resultados. Se han utilizado 10 particiones para la evaluación cruzada y 500 estimadores para el algoritmo de entrenamiento *Random Forest* (el resto de parámetros se han mantenido con los valores por defecto).

Para obtener los modelos de *word embeddings* se han utilizado "dumps" de la Wikipedia [5] para ambos idiomas, eliminando las etiquetas y anotaciones HTML. Una vez limpios, hemos obtenido dos corpus, uno de 5,589,342,425 palabras para el inglés y otro de 1,116,015,489 palabras para el español.

Se han llevado a cabo dos tipos de experimentos. En primer lugar, para estudiar la bondad de cada una de las métricas por separado, se han ejecutado experimentos individuales usando cada una de las métricas de manera independiente (Tabla 3). En segundo lugar, se han entrenado y evaluado modelos a partir de cuatro conjuntos de métricas (ver Tabla 4): las basadas en matrices de transformación (acrónimo *MAT* en las tablas de resultados), las basadas en traducción automática al español (*TRAD_ES*), las basadas en traducción automática al inglés (*TRAD_EN*) y el conjunto total de métricas (*COMB*). En todos los casos se muestran valores de correlación de Pearson (ρ) entre los valores de similitud estimados y reales.

4.1 Análisis de resultados

Observando los resultados obtenidos usando cada una de las métricas de manera individual (Tabla 3), en la mayoría de los casos es la métrica obtenida mediante alineamiento la que consigue los mejores resultados. En general, el método de alineamiento funciona mejor cuando está basado en traducción automática ($\bar{\rho} = 0,563$) que cuando lo está en las matrices de transformación ($\bar{\rho} = 0,488$), siendo preferible además realizar la traducción de las frases del español al inglés ($\bar{\rho} = 0,574$) frente

[4] Se emplea la implementación disponible en *scikit-learn* (Pedregosa et al., 2011).

[5] Extraídos de `https://dumps.wikimedia.org`

	MAT					TRAD_ES			TRAD_EN		
	cos es→en	euc es→en	cos en→es	euc en→es	ali	cos	euc	ali	cos	euc	ali
SE-12-TE	0,006	**0,088**	-0,033	0,062	0,080	0,182	0,177	**0,205**	0,186	**0,227**	0,138
SE-14-STS	0,499	0,460	0,396	0,374	**0,627**	0,628	0,631	**0,665**	0,633	0,630	**0,716**
SE-15-STS	0,248	**0,480**	0,305	0,382	0,447	0,466	**0,490**	0,485	0,457	0,480	**0,531**
SE-16-STS-multi	0,333	**0,471**	0,276	0,452	0,467	0,530	**0,641**	0,540	**0,674**	0,673	0,611
SE-16-STS-news	0,596	0,306	0,416	0,322	**0,818**	0,759	0,725	**0,864**	0,819	0,722	**0,876**
(promedio)	0,336	0,361	0,272	0,318	**0,488**	0,513	0,533	**0,552**	0,554	0,547	**0,574**

Tabla 3: Resultados (ρ) usando cada métrica de similitud de manera independiente

	MAT	TRAD_ES	TRAD_EN	COMB
SE-12-TE	0,153	0,323	0,283	**0,420**
SE-14-STS	0,711	0,764	0,735	**0,772**
SE-15-STS	0,589	0,602	0,538	**0,606**
SE-16-STS-multi	0,665	0,727	0,651	**0,778**
SE-16-STS-news	0,846	0,884	0,878	**0,899**
(promedio)	0,593	0,660	0,617	**0,695**

Tabla 4: Resultados (ρ) usando los distintos conjuntos de métricas de similitud

a la traducción inglés-español ($\bar{\rho} = 0,552$). En cuanto a las métricas obtenidas mediante agregación, no es posible asegurar cuál de ellas (la distancia del coseno o la euclídea) es un mejor estimador, pues obtienen resultados similares o se imponen de manera alternativa según el conjunto de datos y el método multilingüe utilizado.

Si nos fijamos en los resultados obtenidos por los conjuntos formados por las métricas relativas a cada uno de los métodos de similitud multilingüe propuestos (Tabla 4), los datos muestran claramente que el método consistente en la traducción automática de las frases en inglés al español es el más efectivo ($\bar{\rho} = 0,66$), seguido del método basado en la traducción contraria ($\bar{\rho} = 0,617$). Parece claro por tanto que es más efectivo llevar a cabo un proceso previo de traducción que utilizar el método de transformación del espacio vectorial ($\bar{\rho} = 0,66$ frente a $\bar{\rho} = 0,593$). Sin embargo, ambos métodos, traducción y transformación, aportan información complementaria de cara a la resolución de la tarea de similitud textual, como se desprende de los resultados obtenidos al utilizar todas las métricas de manera conjunta. En este escenario es en el que se obtienen los mejores resultados ($\bar{\rho} = 0,695$), con incrementos sustanciales con respecto a los resultados anteriores; en promedio, se obtienen más de 3 puntos porcentuales de mejora con respecto al mejor de los resultados anteriores.

Analizando los resultados por corpus, se confirma la mayor dificultad del conjunto de frases adaptadas de la tarea TE de 2012; es posible que los peores resultados obtenidos se deban a la ausencia de frases con similitud 0. En el resto de los conjuntos de datos se obtienen resultados claramente mejores. Tres de ellos obtienen un resultado de $\rho > 0,77$, siendo especialmente reseñable el resultado obtenido para el corpus de la tarea STS de 2016, en su versión *news* ($\rho = 0,899$), a menos de dos puntos de distancia del mejor resultado de la competición (0.912).

5 *Conclusiones y trabajo futuro*

En este trabajo hemos explorado de qué forma los modelos de *word embeddings* pueden ser aplicados para calcular la similitud semántica de textos escritos en español e inglés, respectivamente. Hemos seguido dos estrategias para poder comparar palabras de idiomas distintos: usando un traductor automático para poder aplicar posteriormente técnicas de similitud monolingüe, y mediante matrices de transformación que permitan trasladar los vectores de las palabras de un idioma al espacio del otro. En cuanto a las técnicas para calcular métricas de similitud, nos hemos apoyado en dos aproximaciones que han demostrado ser efectivas para el caso monolingüe: agregación de vectores para obtener una representación vectorial de los textos, y uso de los vectores para decidir el mejor alineamiento entre las palabras de los dos textos a comparar. Los experimentos muestran que tanto la traducción automática como la transformación de modelos son de utilidad como base para el cálculo de la similitud, obteniéndose mejores resultados con la traducción automática. Cuando se integran las salidas de ambos tipos de sistemas mediante técnicas de *ensemble learning* los resultados mejoran sensiblemente, lo que demuestra un alto grado de complementariedad en la información que se obtiene con cada una de las dos estrategias.

Como trabajo futuro estamos especialmente interesados en investigar otras alter-

nativas para aprovechar los modelos de *word embeddings* en el cálculo de similitud de textos bilingües. Los buenos resultados que hemos obtenido con la combinación de distintos sistemas nos animan a seguir por esta vía introduciendo nuevos *inputs* que aporten información complementaria. Además de la traducción automática y la transformación lineal de modelos, que hemos usado en este trabajo, estamos valorando otras opciones como la creación de modelos híbridos que integren palabras de dos idiomas o la transformación de modelos mediante la aplicación de alguna técnica de aprendizaje automático.

Bibliografía

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, y others. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, páginas 252–263.

Agirre, E., C. Banea, C. Cardie, y J. Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 81–91.

Agirre, E., M. Diab, D. Cer, y A. Gonzalez-Agirre. 2012. Semeval-2012 task 6: a pilot on semantic textual similarity. En *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, páginas 385–393.

Agirre, E., A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, y L. Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. En *Proceedings of SemEval-2016*, páginas 512–524.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, páginas 1–32.

Kovatchev, V., M. Salamó, y M. A. Martí. 2016. Comparing distributional semantics models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural*, 57:109–116.

López, G. J. y I. M. Ruiz. 2016. Character and word baselines systems for irony detection in spanish short texts. *Procesamiento del Lenguaje Natural*, 56:41–48.

López-Solaz, T., J. A. Troyano, F. J. Ortega, y F. Enríquez. 2016. Una aproximación al uso de word embeddings en una tarea de similitud de textos en español. *Procesamiento del Lenguaje Natural*, 57:67–74.

Mikolov, T., Q. V Le, y I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. En C J C Burges L Bottou M Welling Z Ghahramani, y K Q Weinberger, editores, *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., páginas 3111–3119.

Negri, M., L. Bentivogli, Y. Mehdad, D. Giampiccolo, y A. Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, páginas 670–679. Association for Computational Linguistics.

Pablos, A. G., M. Cuadros, y G. Rigau. 2015. Unsupervised word polarity tagging by exploiting continuous word representations. *Procesamiento del Lenguaje Natural*, 55:127–134.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saquete, E. y B. Navarro-Colorado. 2017. Cross-document event ordering through temporal relation inference and distributional semantic models. *Procesamiento del Lenguaje Natural*, 58:61–68.