Una aproximación basada en redes neuronales para el problema de implicación textual translingüe

Darnes Vilariño, David Pinto,
Mireya Tovar, y Beatriz Beltrán

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla, México
{darnes, dpinto, mtovar, bbeltran}@cs.buap.mx
Paper received on 20/07/12, Accepted on 02/08/12.

Resumen. En este artículo de investigación se presenta una aproximación basada en Redes Neuronales para resolver el problema de Implicación Textual entre diferentes idiomas. Se parte de la hipótesis de que la similitud entre dos textos y su correspondiente longitud puede ayudar a resolver el problema de implicación textual. Se ofrecen dos formas para calcular la similitud textual (léxica y por variación léxica-semántica). Dado que la pareja de textos a evaluar se encuentran en idiomas diferentes, se lleva a cabo un proceso de traducción para posteriormente calcular los grados de similitud entre las oraciones en un mismo idioma. Se usan los resultados reportados en la literatura para comparar aquellos obtenidos por la aproximación propuesta en este trabajo. Los experimentos reportados muestran que sí existe relación entre el tamaño de las oraciones, el grado de similitud y su correspondiente juicio de implicación textual (bidireccional, hacia adelante, hacia atrás o sin implicación).

Palabras Clave: Implicación textual translingüe, Redes Neuronales, Aprendizaje Automático**.**

1 Introducción

La implicación textual translingüe (Cross-Lingual Textual Entailment o CLTE por sus siglas en inglés) ha sido propuesta recientemente por [4, 5] como una extensión de la tarea de implicación textual [1], la cual se puede definir como sigue: Dado un texto (*T*) y una hipótesis (*H*) escritos en diferentes idiomas, la tarea de CLTE consiste en determinar si el significado de *H* puede ser inferido a partir del significado de *T*. En este artículo se reportan los resultados obtenidos al evaluar dos características básicas de los textos en cuestión: longitud y grado de similitud. Se determina el rendimiento de ambas características usando un enfoque basado en conocimiento y un enfoque supervisado. Los textos son también analizados cuando se aplica un procedimiento de expansión semántica para verificar si este proceso ayuda o no a mejorar el rendimiento de los algoritmos para la tarea planteada. Parte

Advances in Intelligent and Information Technologies.
Research in Computing Science 60, 2012, pp. 285 -293.

de estos experimentos fueron reportados en la Tarea 8 del foro de evaluación semántica SemEval 2012 y que llevó por nombre *"Cross-lingual Textual Entailment for Content Synchronization*" [6].

Para los experimentos llevados a cabo en este artículo, se define la tarea de implicación textual translingüe de manera formal como sigue:

Dado un par de fragmentos de texto tópicamente relacionados (T_1 y T_2) escritos en diferentes idiomas, la tarea consiste en anotar este par automáticamente con alguno de los siguientes juicios de implicación textual:

- Bidireccional (*Bidirectional*) ($T_1 \rightarrow T_2$ & $T_1 \leftarrow T_2$): los dos fragmentos se implican mutuamente (equivalencia semántica).
- Hacia adelante (*Forward*) ($T_1 \rightarrow T_2$ & T_1 ! $\leftarrow T_2$): implicación unidireccional de T_1 a T_2.
- Hacia atrás (*Backward*) (T_1 ! $\rightarrow T_2$ & $T_1 \leftarrow T_2$): implicación unidireccional de T_2 a T_1.
- Sin implicación (*No Entailment*) (T_1 ! $\rightarrow T_2$ & T_1 ! $\leftarrow T_2$): no hay implicación textual entre T_1 y T_2.

En esta tarea, se asume que tanto T_1 como T_2 son declaraciones verdaderas (*TRUE*); de aquí que no existen pares contradictorios. Los conjuntos de datos de entrenamiento y prueba se encuentran disponibles en las siguientes combinaciones de idiomas:

- Español/Inglés (SPA-ENG)
- Alemán/Inglés (DEU-ENG)
- Italiano/Inglés (ITA-ENG)
- Francés/Inglés (FRA-ENG)

A manera de ejemplo, considere los siguientes pares de oraciones (Español/Inglés) que cumplen con uno de los juicios de implicación textual:

Juicio de implicación: Forward
T_1 : La unificación italiana fue el movimiento político y social que aglomeró diferentes estados de la península itálica en el único estado de Italia en el siglo XIX.
T_2 : Italian unification was the political and social movement which agglomerated different states of the Italian peninsula into the single state of Italy.
Juicio de implicación: Backward
T_1 : News Corporation, que pertenece a Rupert Murdoch, ha decidido rechazar un contrato para tomar el control total de la cadena de transmisión BSkyB.
T_2 : News Corporation, held by Rupert Murdoch, has decided to give up the plan of acquiring full control of BSkyB, a broadcasting company headquartered in London.
Juicio de implicación: Bidirectional
T_1 : Ratko Mladic, también conocido como "El carnicero de Bosnia", fue arrestado después de haber sido buscado durante más de una década.
T_2 : Ratko Mladic, nicknamed "The Butcher of Bosnia," has been taken imprisoned after being sought for more than 10 years.
Juicio de implicación: No entailment
T_1 : Un tsunami que se generó en el Pacífico Sur por un poderoso terremoto submarino ha matado a 110 personas.

T_2 : A strong undersea earthquake started a tsunami in the Pacific, leading to the death of at least 110 people with the majority of fatalities in Samoa.

Las oraciones anteriormente mostradas fueron tomadas aleatoriamente del conjunto de datos de entrenamiento empleado en los experimentos presentados en este trabajo de investigación. Son usadas únicamente con el propósito de mejorar la comprensión del artículo, pero de ninguna manera reflejan un comportamiento generalizado sobre los pares de oraciones y los juicios de implicación textual.

El resto de este artículo se encuentra estructurado de la siguiente manera: la sección 2 describe las diferentes aproximaciones presentadas en este trabajo. Los resultados obtenidos son mostrados y discutidos en la sección 3. Finalmente, se discuten las conclusiones y trabajo a futuro en la sección 4.

2 Descripción de las propuestas

Para el experimento llevado a cabo en este trabajo, hemos considerado atacar el problema de CLTE usando principalmente dos características: tamaño de las oraciones (longitud textual) y la similitud textual entre oraciones del mismo idioma. Para el caso de similitud textual, hemos considerado un enfoque léxico y uno por variación léxica-semántica basado en expansión de términos (de manera concreta, sinónimos).

El juicio de implicación textual es llevado a cabo de dos maneras diferentes: 1) usando reglas de decisión planteadas de manera empírica, y 2) usando una red neuronal entrenada a priori.

Con la finalidad de encontrar el grado de similitud entre dos oraciones en el mismo idioma, fue necesario traducir las oraciones originales (escritas en dos idiomas distintos) al idioma contraparte, es decir, si el par de oraciones (T_1, T_2) es Alemán-Inglés, entonces T_1 se traduce al inglés y T_2 se traduce al alemán. Para este propósito hemos usado Google Translate[1].

Antes de presentar los modelos de determinación del juicio de implicación textual usando reglas empíricas y redes neuronales, presentamos a continuación dos formas de calcular la similitud textual entre dos oraciones.

2.1 Cálculo de similitud textual

Sea T_1 la primera oración del par a analizar su implicación textual y escrita en el idioma origen (Español, Alemán, Italiano o Francés) y T_2 el fragmento de texto tópicamente relacionado con T_1 (dado en idioma Inglés), entonces, se obtiene T_3 que es la traducción al Inglés de T_1, y T_4 que es la traducción de T_2 al idioma origen (Español, Alemán, Italiano o Francés). El cálculo de similitud textual a nivel léxico y por variación léxico-semántica se describe a continuación. En ambos casos, se usa la oración original, es decir, no se realiza preprocesamiento alguno de las mismas.

[1] http://translate.google.com.mx

2.1.1 Similitud léxica

Primeramente se determina la similitud léxica entre dos textos escritos en el idioma origen (*SimS*), es decir, entre T_1 y T_4. Adicionalmente, se calcula la similitud léxica entre las oraciones que se encuentran escritas en inglés (*SimT*), es decir, entre T_2 y T_3.

Para este propósito se ha utilizado el coeficiente de similitud de Jaccard [7]. La ecuación (1) muestra la similitud léxica para dos textos escritos en el idioma origen. De la manera similar se define *SimT* como el grado de similitud para los dos fragmentos de textos escritos en el idioma inglés.

$$SimS = simJaccard(T_1, T_4) = \frac{|T_1 \cap T_4|}{|T_1 \cup T_4|} \qquad (1)$$

2.1.2 Similitud léxico-semántica

Para calcular el grado de similitud léxico-semántica se considera una expansión de los términos originales en las oraciones usando los sinónimos de cada palabra (en el idioma origen y en el idioma destino). Se emplearon cinco diccionarios que contienen los sinónimos de los idiomas considerados en el marco del SemEval 2012 (inglés, español, alemán, italiano y francés)[2] para este propósito.

En la Tabla 1 se muestra tanto el número de términos, como el número de sinónimos en promedio por término considerados para cada idioma.

Tabla 1. Diccionarios de sinónimos usados para la expansión de términos.

Idioma	Términos	Num. promedio de sinónimos por término
Inglés	2,764	60
Español	9,887	45
Alemán	21,958	115
Italiano	25,724	56
Francés	36,207	93

Sea $T_1 = w_{1,1}w_{1,2} \dots w_{1,|T_1|}$ y $T_2 = w_{2,1}w_{2,2} \dots w_{2,|T_2|}$ las oraciones origen y destino, respectivamente. Sea $T_3 = w_{3,1}w_{3,2} \dots w_{3,|T_3|}$ y $T_4 = w_{4,1}w_{4,2} \dots w_{4,|T_4|}$ las versiones traducidas de las oraciones en el idioma origen y destino, respectivamente. El conjunto de sinónimos para la palabra/término $w_{i,k}$ se expresa como $Synset(w_{i,k})$, y se obtiene a partir de los diccionarios anteriormente mencionados. Para lograr un mejor grado de similitud entre ambos textos, cada palabra se ha generalizado utilizando el truncador de Porter [8].

Para calcular la similitud léxico-semántica entre dos palabras escritas en el idioma origen ($w_{1,i}$ y $w_{4,j}$), se usa una ecuación similar a la ecuación (1). De la misma

[2] http://extensions.services.openoffice.org/dictionaries

manera se puede obtener *Sim*($w_{2,i}$, $w_{3,j}$), que es la similitud por variación léxico-semántica entre las palabras escritas en el idioma inglés.

$$Sim(w_{1,i}, w_{4,j}) = \begin{cases} 1, & \text{si } (w_{1,i} = w_{4,j})\ OR\ w_{1,i} \in SynSet(w_{4,j}) \\ & \quad OR\ w_{4,j} \in SynSet(w_{1,i}) \\ 0, & \text{en otro caso} \end{cases} \quad (2)$$

Ambas oraciones consideran la existencia de similitud cuando dos palabras son idénticas, o cuando al menos alguna de las dos palabras aparece en el conjunto de sinónimos de la otra palabra.

La similitud entre los fragmentos de texto T_1 y T_4 (*SimS*) se calcula por medio de la ecuación (3), y de la misma manera se obtiene *SimT* (T_2, T_3) que es la similitud léxico-semántica completa entre los fragmentos de texto T_2 y T_3.

$$SimS(T_1, T_4) = \frac{\sum_{i=1}^{|T_1|} \sum_{j=1}^{|T_4|} Sim(w_{1,i}, w_{4,j})}{|T_1 \cup T_4|} \quad (3)$$

2.2 Determinación del juicio de implicación textual

En esta sección se presentan dos aproximaciones para determinar el correspondiente juicio de implicación textual. Para ambos casos se utilizan dos formas de calcular la similitud textual (léxica y léxico-semántica), obteniendo cuatro experimentos que son evaluados y comparados en la Sección 3.

2.2.1. Aproximación empírica

En este caso, se ha considerado el tamaño de la oración como una característica en la determinación del juicio de implicación textual. De manera particular, se compara la longitud de la oración T_2 con respecto a la oración T_3 (ambas escritas en el idioma inglés). Un análisis empírico nos ha llevado a considerar que si la longitud de T_2 es menor que la longitud de T_3 y existe además un grado de similitud textual (léxica o semántica) superior a *0.5* para los pares de oraciones escritos en los mismos idiomas, entonces, se establece un juicio tipo *Forward*. Por otro lado, si la longitud de T_2 es mayor que la longitud de T_3 y existe además un grado de similitud textual (léxica o léxico-semántica) superior a *0.5* para los pares de oraciones escritos en los mismos idiomas, entonces, se establece un juicio tipo *Backward*. En el caso en el que, la longitud de T_2 sea igual que la longitud de T_3, la longitud de T_1 sea igual que la longitud de T_4, y existe además un grado de similitud textual (léxica o léxico-semántica) superior a *0.5* para los pares de oraciones escritos en los mismos idiomas, entonces, se establece un juicio tipo *Bidirectional*. Finalmente, si ninguno de los casos anteriores se cumple, se establece un juicio de *No entailment*. Estas reglas de decisión pueden verse más claramente en el Algoritmo 1.

Algoritmo 1.

If $|T_2| < |T_3|$ then
  If (*simT* > 0.5 and *simS* > 0.5)
  then **Forward**

ElseIf $|T_2| > |T_3|$ then
If (*simT* > 0.5 and *simS* > 0.5)
then **Backward**
ElseIf ($|T_1| = |T_4|$ and $|T_2| = |T_3|$) then
If (*simT* > 0.5 and *simS* > 0.5)
then **Bidirectional**
Else **No entailment**

Dado que se usan dos tipos de similitud textual, se han obtenido dos aproximaciones basadas en la propuesta empírica anterior. Hemos llamado *EMP_Lex*, a la propuesta que usa la similitud textual léxica, mientras que aquella que usa la similitud por variación léxico-semántica se ha denomidado *EMP_Sem*.

2.2.2. Aproximación supervisada

Para esta aproximación se utilizó una red neuronal tipo perceptrón multicapa de propagación hacia atrás (implementado en Weka [2]), con 20 neuronas en la capa de entrada, 7 neuronas en la capa intermedia y 4 de salida. Las neuronas de la capa de entrada combinan las seis características consideradas para la entrada de la red que son: cuatro longitudes de oración (para T_1, T_2, T_3 y T_4), y los dos valores de similitud textual (*SimS* y *SimT*). Los parámetros de la red neuronal fueron ajustados logrando un 95.7% de instancias correctamente clasificadas en el proceso de entrenamiento.

Dado que se usan dos tipos de similitud textual, se han obtenido dos aproximaciones basadas en la propuesta supervisada anterior. Hemos llamado *NN_Lex*, a la propuesta que usa la similitud textual léxica, mientras que aquella que usa la similitud por variación léxico-semántica se ha denomidado *NN_Sem*.

3. Experimentos

Los corpora usados en los experimentos provienen del conjunto de datos para implicación textual translingüe presentados en [3] y proporcionados por los organizadores de la Tarea 8 de SemEval 2012 [6]. En el caso de las reglas de decisión empíricas, se emplearon los conjuntos de entrenamiento únicamente para ajustar los parámetros, mientras que en el caso de la red neuronal, dichos conjuntos se usaron para entrenar los pesos de la red. La cantidad de oraciones contempladas tanto en los conjuntos de entrenamiento como de prueba fue de 500 para cada idioma.

En la tabla 2 se muestran los resultados globales obtenidos por las cuatro aproximaciones propuestas en este artículo. Adicionalmente, hemos incluido los puntajes máximo (*Mejor*), mínimo (*Peor*), promedio (*Promedio*) y mediana (*Mediana*), reportados en la competencia de la Tarea 8 de SemEval 2012; esto con fines de comparación.

Tabla 2. Resultados globales obtenidos por las propuestas y su comparación con aquellos obtenidos en la Tarea 8 de SemEval 2012

	SPA-ENG	**ITA-ENG**	**FRA-ENG**	**DEU-ENG**
Mejor	**0.632**	**0.566**	**0.57**	**0.558**
Promedio	0.407	0.362	0.366	0.357
Mediana	0.346	0.336	0.336	0.336
Peor	0.266	0.278	0.278	0.262
EMP_Lex	0.350	0.336	0.334	0.330
EMP_Sem	0.366	0.344	0.342	0.268
NN_Lex	**0.476**	**0.456**	**0.486**	**0.458**
NN_Sem	0.398	0.398	0.392	0.386

Los puntajes muestran que la aproximación supervisada se encuentra por arriba de la mediana de los resultados reportados en la literatura. Si bien, aun no se alcanza el mejor valor obtenido en la competencia, se ha observado que las dos características básicas que se han considerado son sumamente importantes, pues se acercan considerablemente al mejor puntaje. Cabe aclarar que no se han hecho análisis semánticos profundos, lo cual es claramente una ventaja de las propuestas planteadas. Sin embargo, se considera que es un buen punto de partida y que al agregar un análisis más profundo de las oraciones en cuestión, seguramente se obtendrá un mejor rendimiento.

Tabla 3. Resultados obtenidos por la red neuronal (*NN_Lex*) para cada uno de los juicios de implicación textual (sin expansión de términos).

Idiomas	***Backward***			***Forward***			***No-Entailment***			***Bidirectional***			**Promedio general**
	P	***R***	***F_1***	***P***	***R***	***F_1***	***P***	***R***	***F_1***	***P***	***R***	***F_1***	
SPA-ENG	0.573	0.504	0.536	0.574	0.528	0.550	0.392	0.376	0.384	0.400	0.496	0.443	**0.476**
ITA-ENG	0.551	0.520	0.535	0.542	0.468	0.502	0.382	0.432	0.405	0.370	0.408	0.391	**0.456**
FRA-ENG	0.561	0.552	0.556	0.567	0.576	0.571	0.393	0.382	0.389	0.425	0.432	0.429	**0.486**
DEU-ENG	0.544	0.544	0.544	0.510	0.584	0.545	0.430	0.296	0.351	0.349	0.408	0.376	**0.458**

Tabla 4. Resultados obtenidos por la red neuronal (*NN_Sem*) para cada uno de los juicios de implicación textual (con expansión de términos).

Idiomas	***Backward***			***Forward***			***No-Entailment***			***Bidirectional***			**Promedio general**
	P	***R***	***F_1***	***P***	***R***	***F_1***	***P***	***R***	***F_1***	***P***	***R***	***F_1***	
SPA-ENG	0.511	0.576	0.541	0.536	0.536	0.536	0.226	0.196	0.208	0.281	0.288	0.285	**0.398**
ITA-ENG	0.541	0.580	0.560	0.500	0.376	0.429	0.270	0.248	0.258	0.308	0.384	0.342	**0.398**
FRA-ENG	0.460	0.460	0.460	0.617	0.568	0.592	0.220	0.296	0.252	0.318	0.248	0.279	**0.392**
DEU-ENG	0.515	0.400	0.450	0.492	0.512	0.502	0.294	0.280	0.287	0.286	0.352	0.315	**0.386**

Con la finalidad de analizar a profundidad los resultados obtenidos con las aproximaciones basadas en redes neuronales (*NN_Lex* y *NN_Sem*), en las Tabla 3 y 4 se muestran los valores de precisión (***P***), *recall* (***R***) y $\boldsymbol{F_1}$ obtenido para el juicio de implicación textual. Ahí se puede observar que al menos para la aproximación plan-

teada, el uso de sinónimos generaliza demasiado, incrementando el valor de similitud textual y confundiendo los juicios de implicación finales. Analizando los valores obtenidos, se puede ver que las carácterísticas usadas son significativas para el caso de juicios *Forward* y *Backward*, sin embargo, no son suficientes para determinar con buen grado de precisión los juicios *Bidirectional* y *No entailment*. Como trabajo a futuro se plantea revisar cuáles son las características más relevantes para estos dos últimos juicios de implicación textual y así mejorar los resultados obtenidos.

4 Conclusión

En este trabajo se presentan cuatro aproximaciones para resolver el problema de implicación textual translingüe. Dos aproximaciones usan un enfoque empírico empleando similitud textual léxica y por variación léxica-semántica. Las otras dos aproximaciones usan un enfoque supervisado contemplando también ambos tipos de similitud textual.

Se ha observado que la determinación del juicio *No entailment* es complicada, pues existen oraciones que no se implican textualmente, pero comparten términos comunes. Para el caso del juicio *Bidirectional*, se ha obligado a que las longitudes de las oraciones sean exactamente iguales, lo cual es una restricción muy estricta. Por lo que, se debe buscar una estrategia para relajar esta restricción.

A pesar de las limitaciones anteriores, se considera que el uso de las longitudes de las oraciones en la determinación de un juicio de implicación textual translingüe es una característica interesante, pues es relativamente simple de calcular y que no requiere un análisis semántico profundo.

Se está en proceso de evaluar el efecto de la traducción en los porcentajes obtenidos cuando se determina el juicio de implicación. Como trabajo a futuro, se piensa usar corpus paralelos para obtener diccionarios estadísticos que puedan mapear de mejor manera el significado de las oraciones.

Referencias

1. Dagan, Ido and Oren Glickman: Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, January 2004.
2. Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005).
3. Manning, Christopher D., Schütze, Hinrich: Foundations of Statistical Natural Language Processing. MIT Press, 6a Edición (2003)
4. Mehdad, Yashar, Matteo Negri, and Marcello Federico: Towards Cross-Lingual Textual Entailment. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, 321–324 (2010).
5. Mehdad, Yashar, Matteo Negri, and Marcello Federico: Using bilingual parallel corpora for cross-lingual textual entailment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT ’11, Stroudsburg, PA, USA, Volume 1, 1336–1345 (2011).

6. Negri, M., A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo: Semeval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. First Joint Conference on Lexical and Computational Semantics (*SEM), pages 399–407, Montreal, Canada, June 7-8, (2012).
7. Negri, Matteo, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti: Divide and conquer: crowdsourcing the creation of crosslingual textual entailment corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, Stroudsburg, PA, USA, 670–679 (2011).
8. Porter, M: An algorithm for suffix stripping. Program, 14(3):130–137 (1980).