Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 99-102 recibido 12-11-2018 revisado 07-01-2019 aceptado 23-01-2019

# Contribuciones a la comprensión lectora: mecanismos de atención y alineamiento entre n-gramas para similitud e inferencia interpretable *

## *Contributions to language understanding: n-gram attention and alignments for interpretable similarity and inference*

**Iñigo Lopez-Gazpio**
Grupo Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
inigo.lopez@ehu.eus

**Resumen:** Tesis doctoral titulada "Hizkuntza-ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarrirako / Contribuciones a la Comprensión Lectora: Mecanismos de Atención y Alineamiento entre N-gramas para Similitud e Inferencia Interpretable", defendida por Iñigo Lopez-Gazpio en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de los doctores Eneko Agirre (Departamento de Lenguajes y Sistemas Informáticos) y Montse Maritxalar (Departamento de Lenguajes y Sistemas Informáticos). La defensa tuvo lugar el 30 de octubre de 2018 ante el tribunal formado por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Gorka Azkune (Secretario, Universidad de Deusto) y David Martínez (Vocal, IBM). La tesis obtuvo la calificación de sobresaliente Cum Laude y mención internacional.
**Palabras clave:** Similitud semántica, inferencia textual, redes neuronales, mecanismos de atención, alineación n-gramas, composicionalidad

**Abstract:** Ph. D. thesis entitled "Hizkuntza-ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarrirako / Contributions to Language Understanding: N-gram Attention and Alignments for Interpretable Similarity and Inference", written by Iñigo Lopez-Gazpio at the University of Basque Country (UPV/EHU) under the supervision of Dr. Eneko Agirre (Languages and Computer Systems Department) and Dr. Montse Maritxalar (Languages and Computer Systems Department). The viva voce was held on October 30 2018 and the members of the commission were Dr. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Dr. Gorka Azkune (Secretary, University of Deusto) and Dr. David Martínez (Vocal, IBM). The thesis obtained Cum Laude excellent grade and international mention.
**Keywords:** Textual similarity, language inference, neural networks, attention mechanisms, n-gram alignments, compositionality

### *1 Introducción de la tesis*

Esta tesis doctoral se ha realizado en el grupo Ixa de la Universidad del País Vasco (UPV/EHU) y presenta contribuciones relacionadas con la comprensión lectora de sistemas inteligentes por medio de las cuales estos sistemas incrementan su capacidad para entender el texto en entornos educacionales. Principalmente trata la línea de investigación de la composicionalidad de textos en vectores distribucionales, y la línea de investigación de la identificación e interpretación de similitudes y diferencias entre textos.

La sociedad actual muestra cierto interés por el aprendizaje continuo incluso en etapas avanzadas de la vida, y esto resulta en un creciente interés de cursos de diversas áreas de los cuales muchos se imparten *online*. Una gran ventaja de estos cursos en línea o del *e-learning* en general reside en su capacidad para favorecer la expansión y llegar a muchos es-

*Esta tesis doctoral ha sido realizada con una beca predoctoral del Ministerio de Educación, Cultura y Deporte. Referencia: MINECO FPU13/00501.

ISSN 1135-5948. DOI 10.26342/2019-62-12

tudiantes sin ninguna restricción geográfica. Como consecuencia de abarcar un espectro tan amplio de estudiantes potenciales es habitual que estos cursos lleguen a tener un número masivo de estudiantes. No en vano estos cursos son conocidos con el acrónimo *MOOC* del inglés *Massive Open Online Course*. El principal problema de los cursos MOOC es que los docentes de dichos cursos no son capaces de afrontar las necesidades individualizadas de los estudiantes inscritos en sus cursos, debido a su gran número. Como consecuencia de esta sobrecarga emplean evaluaciones tipo test para corregir las actividades de los estudiantes.

La motivación principal de esta tesis es desarrollar técnicas de procesamiento de lenguaje natural (*PLN*) con las que poder evaluar de forma automática a los estudiantes con respecto a una respuesta de referencia de un experto docente. Además, nuestra motivación es que los sistemas expertos de PLN sean capaces de identificar y relacionar segmentos entre el texto escrito por un estudiante y la respuesta de referencia, de forma que identifiquen explícitamente similitudes y diferencias entre ambos textos. Identificar estas relaciones es clave para pode producir retroalimentación en tiempo real con respecto a una respuesta de un estudiante que se desea evaluar.

## *2 Estructura de la tesis*

La presente tesis tiene dos objetivos principales: 1) el desarrollo de sistemas inteligentes de PLN que sean capaces de evaluar respuestas de estudiantes contra respuestas de referencia de expertos docentes, y 2) que estos sistemas inteligentes sean capaces de producir retroalimentación útil para que los estudiantes puedan continuar su labor de aprendizaje.

Para organizar estos objetivos de forma secuencial hemos dividido la tesis en cinco secciones, la cual se presenta como compilación de artículos. En una primera sección introductoria presentamos la motivación, los objetos de estudio y las líneas de investigación que utilizaremos a lo largo de la tesis. Esta sección introductoria también contiene un resumen de todos los artículos relacionados. En la segunda sección nos centramos en realizar un análisis profundo del estado del arte con respecto a tecnologías del PLN, así como a analizar tareas y sistemas del ámbito educacional llevadas a cabo hasta la fecha (Agirre et al., 2015a). Realizamos un énfasis especial en las arquitecturas basadas en redes neuronales y en las tareas de *Similitud Textual Semántica* (STS) (Cer et al., 2017) e *Inferencia Lógica* (NLI), ya que dichas arquitecturas y tareas forman la base para nuestro desarrollo de nuevos sistemas inteligentes.

En la tercera sección presentamos nuestro primer artículo (actualmente bajo revisión) que aborda el primer objetivo de la tesis: el desarrollo de sistemas inteligentes de PLN que sean capaces de evaluar un par de textos de entrada. Para el desarrollo del sistema inteligente analizamos diversas técnicas de modelado y representación de texto en vectores distribucionales, como sistemas basados en agrupaciones de palabras (*Bag-of-Words*), en redes neuronales recurrentes (*Recurrent Neural Networks*) y en redes convolucionales (*Convolutional Neural Networks*). En esta tercera sección proponemos una arquitectura novedosa en el estado del arte basada en redes neuronales capaz de modelar, representar y alinear n-gramas arbitrarios entre los textos de entrada. Si bien la alineación entre pares individuales de palabras es algo conocido y explotado con éxito en el estado del arte (Artetxe et al., 2018), la alineación entre n-gramas es una línea de investigación novedosa explorada en el marco de esta tesis.

En la cuarta sección presentamos nuestro segundo artículo (Lopez-Gazpio et al., 2017) que aborda el segundo objetivo de la tesis: explorar la capacidad de los sistemas inteligentes de PLN de forma que sean capaces de identificar las similitudes y diferencias entre un par de textos. De forma que esta capacidad adquirida en los sistemas permita generar retroalimentación útil a los estudiantes. Para implementar esta capacidad desarrollamos una nueva capa un nivel por encima de la Similitud Textual Semántica y la Inferencia Lógica, de forma que este nuevo nivel de anotación permite identificar y relacionar pares de agrupaciones de palabras. Llamamos Similitud Textual Semántica Interpretable (*interpretable STS* o *iSTS*) a esta nueva capa, y con ella es posible entrenar sistemas inteligentes de PLN para que sean capaces de reconocer de manera detallada las diferencias y similitudes entre textos. No sólo hemos diseñado la capa iSTS dentro del marco de esta tesis, sino que también hemos estado activos organizando dicha tarea en SemEval durante diversos años (Agirre et al., 2015b; Agirre et

al., 2016). Además hemos implementado distintos sistemas capaces de resolver la tarea (Agirre et al., 2015c; Lopez-Gazpio, Agirre, y Maritxalar, 2016) y finalmente sometido estos sistemas a evaluación con el objetivo de indagar si la retroalimentación es útil para los humanos en un entorno educacional.

Finalmente, en la quinta sección se presentan las contribuciones de la tesis divididas según los principales objetivos tratados y el trabajo futuro.

## *3 Contribuciones más relevantes*

Para continuar presentamos en este apartado las contribuciones más relevantes divididas en dos apartados acorde con los principales objetivos enumerados en el contexto de la tesis.

### 3.1 Modelos basados en la atención sobre n-gramas

En lo referente al desarrollo de modelos basados en la atención sobre n-gramas, hemos partido sobre la hipótesis inicial en la que considerábamos que la modelización de segmentos mayores que palabras individuales en vectores distribucionales y su respectiva alineación debería de ser superior a la modelización y alineamiento de palabras individuales.

Para llevar a cabo esta labor hemos partido de la implementación propia de un sistema de terceros descrito en el estado del arte basado en un modelo Bag-of-Words. Es decir, un sistema que en sí mismo es capaz de modelar interacción entre oraciones por medio de la interacción individual entre las palabras que conforman la oración. Tomando este sistema como *baseline* hemos realizado ciertas modificaciones para que sea capaz de modelar n-gramas en vectores distribucionales y también sea capaz de alinear dichas representaciones. También hemos diseñado otras dos variantes del modelo inicial que utilizan redes recurrentes y redes de convolución para extender la representación vectorial de las palabras, y poder evaluar nuestra propocición del modelo basado en n-gramas contra otras dos alternativas que utilizan sistemas supervisados complejos para modelas la composicionalidad.

Los resultados obtenidos en cinco conjuntos distintos de test de tareas relativas a STS (STS Benchmark y SICK-TS) y NLI (SNLI, MNLI y SICK-TE) muestran claramente la superioridad del modelo basado en n-gramas contra otras alternativas. Los resultados varían dependiendo del conjunto de test utilizado, en el que con respecto al sistema inicial hemos llegado a obtener una reducción del error relativo del 41 % en SICK-TS, del 38 % en STS Benchmark y del 29 % en SICK-TS. Con respecto a los conjuntos de test de SNLI y MNLI hemos obtenido reducciones del error relativo más limitadas en torno al 8 % y 11 %. También hemos observado que los modelos empleando redes neuronales recurrentes y redes convolucionales para extender la representación distribucional de las palabras obtienen mejores resultados que el sistema inicial, que no utiliza ningún mecanismo complejo de composicionalidad.

Con esta línea de investigación demostramos que el alineamiento entre n-gramas es útil de cara a la representación de oraciones, ya que es capaz de introducir contexto en la representación distribucional de los segmentos de la oración. Desde nuestro punto de vista modelar n-gramas es un paso intermedio entre los sistemas basados en agrupaciones de palabras (Bag-of-Words) y los sistemas basados en árboles de dependencias (Tree-RNN) que efectivamente son capaces de incorporar parte de la estructura sintáctica de las oraciones.

### 3.2 Capacidad de interpretación entre textos

En lo referente al desarrollo de la capacidad de interpretabilidad ya hemos mencionado que nuestra principal aportación ha sido la de diseñar una capa encima de STS y NLI capaz de modelar explícitamente las similitudes y diferencias entre un par de oraciones. Para ello hemos diseñado una nueva tarea (iSTS) en la cual segmentamos las oraciones de entrada, y después realizamos alineamientos entre los segmentos identificados, estableciendo una etiqueta y un valor numérico para cada alineación. Con las etiquetas podemos especificar si un segmento es equivalente, similar, más o menos específico, contradictorio o está relacionado con otro segmento; y con el valor numérico podemos establecer la fuerza de esta etiqueta mediante un valor numérico dentro de una escala, teniendo $valor \in [0, 5]$ Con esta anotación detallada los sistemas son capaces de aprender a identificar y diferenciar las relaciones de grano fino entre las oraciones, y producir retroalimentación útil a estudiantes. Diversos experimentos ponen de manifiesto un aumento en la correlación cuando

los humanos tenían a mano verbalizaciones producidas por sistemas expertos de PLN entrenados en iSTS.

### 3.3 Recursos

Dentro del marco de esta tesis se han creado y liberado una serie de recursos materiales y de software con el objetivo de aportar nuevas herramientas a la comunidad científica.

En relación con el primer objetivo de la tesis se han liberado sistemas basados en redes neuronales capaces de realizar las tareas de STS y NLI[1].

En relación con el segundo objetivo de la tesis se han liberado diversos sistemas basados en aprendizaje automático y redes neuronales capaces de realizar la tarea de iSTS. Además, con respecto a la organización de la tarea en SemEval[2][3] también se han liberado herramientas para realizar la evaluación de sistemas, una aplicación de anotación de datos para facilitar la tarea de creación de nuevos conjuntos de datos, directrices para la anotación de datos, y un total de tres conjuntos de entrenamiento y otros tres conjuntos de test para entrenar nuevos sistemas en la tarea de iSTS. Los conjuntos de datos pertenecen al dominio de titulares de noticias, descripciones de imágenes y respuestas de estudiantes, entre los que suman más de 2500 pares de oraciones anotadas.

### *Agradecimientos*

Agradecemos el apoyo de la corporación NVIDIA por la donación de dos unidades de procesamiento gráfico utilizadas para esta investigación (Tesla K40 y Pascal Titan X).

### *Bibliografía*

Agirre, E., I. Aldabe, O. L. de Lacalle, I. Lopez-Gazpio, y M. Maritxalar. 2015a. Erantzunen kalifikazio automatikorako lehen urratsak. *EKAIA Euskal Herriko Unibertsitateko Zientzia eta Teknologia Aldizkaria*, (29).

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, y J. Wiebe. 2015b. 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)*, Denver, CO, June. Association for Computational Linguistics.

Agirre, E., A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, y L. Uria. 2015c. Ubc: Cubes for english semantic textual similarity and supervised approaches for interpretable sts. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 178–183, Denver, Colorado, June. Association for Computational Linguistics.

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 the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 512–524, San Diego, California, June. Association for Computational Linguistics.

Artetxe, M., G. Labaka, I. Lopez-Gazpio, y E. Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. En *Proceedings of the 22nd Conference on Computational Natural Language Learning*, páginas 282–291.

Cer, D., M. Diab, E. Agirre, I. Lopez-Gazpio, y L. Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. En *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, páginas 1–14. Association for Computational Linguistics.

Lopez-Gazpio, I., E. Agirre, y M. Maritxalar. 2016. iubc at semeval-2016 task 2: Rnns and lstms for interpretable sts. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 771–776, San Diego, California, June. Association for Computational Linguistics.

Lopez-Gazpio, I., M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, y E. Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

[1] `https://github.com/lgazpio`

[2] `http://alt.qcri.org/semeval2015/task2/`

[3] `http://alt.qcri.org/semeval2016/task1/`