Aprendizaje Profundo: Teoría y aplicaciones (INF659)

Detección de Intenciones y Reconocimiento de Entidades

[1] Daniel Arteaga*, Ruben Cordova*, Juan Carlos Tovar*

***Resumen*—**

***Index Terms*—**

I. Introducción

La detección de intenciones y reconocimiento de entidades, son elementos importantes en soluciones como chatbots, que generan un valor agregado en las organizaciones para la satisfacción de los usuarios. Es por ello, que el objetivo de este estudio es la creación de un modelo de deep learning que integre técnicas de sequence-to-sequence para la detección de intenciones y reconocimiento de entidades, automatizando tareas y ejecutando acciones presentes en un diálogo. Sequence-to-sequence (seq2seq) es una técnica empleada ampliamente en los problemas de traducción de procesamiento de lenguaje natural (NLP) donde, como en los modelos de traducción de máquina, se trata de definir los bloques de modelado, aprendizaje de hiper parámetros y predicción. Específicamente, para la etapa de modelado se emplea el paradigma de Encoder-Decoder, donde se lee la entrada y se produce su representación (encoder) para luego emplear dicha representación y generar la secuencia objetivo (decoder). Adicionalmente, se incluye el concepto de atención, el cual le permite al modelo centrarse en partes específicas de la entrada en los diferentes pasos de la red neuronal. Con esto, el modelo obtiene la capacidad de centrarse solamente en las partes relevantes de la entrada.

En el presente trabajo, se propone utilizar la técnica de seq2sec junto con el mecanismo de atención para la detección de intenciones y reconocimiento de entidades. El trabajo está organizado de la siguiente manera: en la sección 2 se presenta el estado de arte, donde se analizan las técnicas utilizadas en trabajos similares. En la sección 3 se detalla el diseño del experimento, específicamente la descripción del conjunto de datos, el preprocesamiento y la metodología empleada.

[1]INF659

*Pontificia Universidad Católica del Perú, Escuela de Posgrado, Maestría de Informática (e-mail: author@pucp.edu.pe).

II. Estado del Arte

Estudios proponen el uso de técnicas de deep learning, como Redes Neuronales Recurrentes (RNN) basados en la atención, para la detección de intenciones y reconocimiento de entidades, pasos importantes para la comprensión del habla y sistemas de diálogo. En [4] Mesnil et al. proponen el uso de RNNs para realizar reconocimiento de entidades en escenarios de compresión de lenguaje hablado (SLU). Este trabajo es uno de las primeras propuestas en este tipo de redes neuronales, cuyos resultados obtenidos muestran una mejora respecto a las técnicas tradicionalmente empleadas como Conditional Random Field (CRF), específicamente en la reducción del 2% en el error absoluto. Con el paso de los años, la atención de las investigaciones se orientó a desarrollar modelos que puedan realizar de forma conjunta la detección de intención y reconocimiento de entidades. En [1] Bing utiliza ATIS (Airline Travel Information Systems) data set, ampliamente utilizado en la investigación de la comprensión del lenguaje hablado. Donde utiliza un modelo de red neuronal recurrente bidireccional con buenos resultados, un F1 score del 95.78 y un error del 5.6% para el reconocimiento de entidades y detección de intenciones, respectivamente. En [2] Schumann, además de utilizar un modelo recurrente basado en la atención, se incorpora la corrección del error del reconocimiento automático del habla (ASR) y con el uso del ATIS dataset y agregando hipótesis ASR (creadas utilizando “Google Text to Speech” API), se logra unos resultados de F1 score del 87.13 y un error del 5.04% para el reconocimiento de entidades y detección de intenciones, respectivamente. En [3] Wang, se propone un modelo híbrido de red neuronal convolucional y red de memoria a largo y corto plazo (CNN-BLSTM) para la codificación y una red neuronal recurrente basada en la atención para la decodificación, con el uso del ATIS dataset se logra unos resultados de F1 score del 97.76% y una precisión del 97.17% para el reconocimiento de entidades y detección de intenciones, respectivamente. Por ejemplo, en [6] Xu y Sarikaya proponen el uso de Redes Neuronales Convolucionales basadas en CRF Triangular (TriCRF) para la detección conjunta. Con esta técnica, los autores logran obtener 5.91% en el valor de error de intención (reducción de 1%) y 95.42% en la métrica de F1 (incremento de 1%), respecto a otros métodos. Adicionalmente, en el presente trabajo no solo se emplea el dataset ATIS sino también en otros dominios (comunicaciones, calendarios, alarmas y notas), donde el modelo obtuvo valores de error de intención entre 5.5% y 7..1% y de la métrica F1 entre 86% y 90%. Otro ejemplo es [5], donde Guo et al. proponen el uso de Redes Neuronales Recursivas (RecNN) para la detección conjunta. En el trabajo se evalúan los resultados del modelo en comparación con otros modelos (p. ej. TriCRF), así como el impacto en utilizar el algoritmo de Viterbi para la optimización del modelo. Respecto a otros modelos, se obtienen valores de exactitud de 95.4% para detección de intención y 93.22% para reconocimiento de entidades. Al incluir el algoritmo de Viterbi, se mejora en 0.4% la exactitud del reconocimiento de entidades. Cabe resaltar que además de emplear el dataset de ATIS, también se emplea la data de

Cortana (asistente personal de Microsoft) en los dominios de calendario, comunicaciones y clima. Los resultados obtenidos cuando se usaba la data de los 3 dominios en conjunto muestran una exactitud de 92.28% y 87.86% para detección de intención y reconocimiento de entidades, respectivamente.

III. Diseño del Experimento

En este trabajo se explora el uso de un modelo Secuencia a Secuencia ("seq2seq") [8], también conocido como Codificador - Decodificador (*encoder-decoder*), con mecanismo de Atención [9] para detectar intenciones y reconocer intenciones simultáneamente. El modelo propuesto es una adaptación del modelo Codificador - Decodificador con Alineamiento y Atención descrito en [1]. Como parte del estudio se propone realizar los siguientes experimentos:

- Comparar el rendimiento del modelo propuesto con otros modelos sobre el conjunto de datos ATIS [10]
- Comparar la influencia del padding al inicio y al final de la secuencia
- Probar el modelo propuesto con un conjunto de datos manualmente etiquetados en el dominio de consultas sobre cursos de capacitación
- Analizar el rendimiento de una variante del modelo sólamente en el reconocimiento de intenciones para el conjunto de datos propio.

En esta sección se describen los conjuntos de datos a utilizar y la metodología a seguir durante el desarrollo de los experimentos.

3.1 Descripción del conjunto de Datos

ATIS

El conjunto de datos ATIS (*Airline Travel Information System*) [10] es una referencia ampliamente utilizada para el entrenamiento y prueba de modelos con la arquitectura Secuencia a Secuencia. Este conjunto de datos se ha conformado a partir de grabaciones de voz de personas haciendo reservas de vuelo. El conjunto de datos utilizado se conforma de 4978 oraciones de entrenamiento y 893 oraciones de prueba. Cada oración (secuencia de entrada) posee una intención, que se puede entender como una clase, y una secuencia de etiquetas correspondientes a las entidades reconocidas (secuencia de salida). Este conjunto de datos tiene 943 palabras, 129 entidades y 26 intenciones. ATIS ha sido conformado a partir de textos en inglés. En la Tabla I se presenta un ejemplo de ATIS.

Tabla I: Ejemplo del conjunto de datos ATIS.

Sentence	first	class	fares	from	boston	to	denver
Slots	B-class_type	I-class_type	O	O	B-fromloc	O	B-toloc
Intent	airfare						

Conjunto de Datos Propio

Se viene conformando un conjunto de datos manualmente etiquetados a partir de textos de correos electrónicos enviados al área de informes del Instituto Nacional de Investigación y Capacitación de Telecomunicaciones (INICTEL-UNI), donde se consulta por información de los cursos de capacitación impartidos. Se han etiquetado hasta la fecha un total de 1200 correos electrónicos, lo que representa 6306 oraciones. Para cada oración se ha obtenido una intención y una secuencia de etiquetas de las entidades del dominio. Este conjunto de datos posee 4806 palabras, 36 entidades y 21 intenciones. Sin embargo, debido a que los números de ejemplos por cada intención se encuentran desbalanceados, se aplicará una selección de 7 intenciones para probar el modelo propuesto.

En la Tabla II se muestra un ejemplo de este conjunto de datos.

Tabla II: Ejemplo del conjunto de datos propio.

Oraciones	Secuencia de Entidades	Intenciones
Informacion del curso de SISTEMAS DE video vigilancia CCTV porfa	<o> <o> <o> <o> B-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> I-<nombre_curso> <o>	informacion_curso

3.2 Metodología

Preprocesamiento de Datos

Para ambos conjuntos de datos a emplear es necesario realizar un preprocesamiento de los textos. Este preprocesamiento consiste en los siguientes pasos:

- Añadir los tokens de inicio de oración (BOS) y fin de oración (EOS) a las secuencias de entrada y de salida
- Seleccionar el número de palabras del diccionario y reemplazar las restantes por el token desconocido (UNK)
- Tokenizar las secuencias de entrada y de salida, es decir, separar las secuencias en tokens y obtener diccionarios que permitan el mapeo de una palabra a un número
- Desfasar las secuencias para ingresarlas adecuadamente al codificador y decodificador.

Adicionalmente, para el conjunto de datos propio es necesario una selección previa de los ejemplos correspondientes a las 7 intenciones de forma que el conjunto de datos se encuentre más equilibrado.

Métricas de Evaluación

Tanto como para el caso de la detección de intenciones como para el reconocimiento de entidades, los elementos a predecir (entidades e intenciones) presentan una frecuencia heterogénea por lo que la mejor métrica es el F1-score.

Modelo Propuesto

El modelo propuesto consta de un codificador basado en una sola capa LSTM bidireccional y decodificadores basados en capas LSTM unidireccionales. Así mismo, se ha considerado un alineamiento de las salidas del codificador (como secuencia, es decir salidas por cada palabra de la secuencia de entrada) a las entradas respectivas del decodificador. Adicionalmente, se ha empleado el mecanismo de Atención de Bahdanau [11]. En la Figura 1 se muestra un diagrama de la arquitectura seguida para el modelo propuesto.

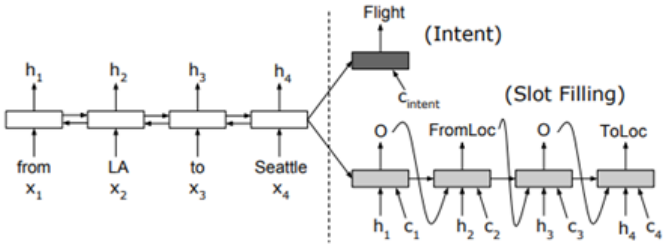

Figura 1: Arquitectura seguida para el modelo propuesto.

Ajuste de hiperparámetros

Para ajustar los hiperparámetros del modelo será necesario hacer un tunning manual basado en la evolución de los resultados. Como se indica al inicio de esta sección, se realizarán diferentes experimentos para los cuales se irán ajustando los hiperparámetros. El mejor modelo obtenido en los primeros experimentos servirán como punto de partida para continuar la experimentación siguiente. Los hiperparámetros principales serán: el tamaño de los embeddings, la dimensión latente, el tamaño del batch y el learning rate.

Marco de la metodología

Este trabajo se desarrolló en el marco que se muestra en la Figura 1, que implica seis fases claves [7].

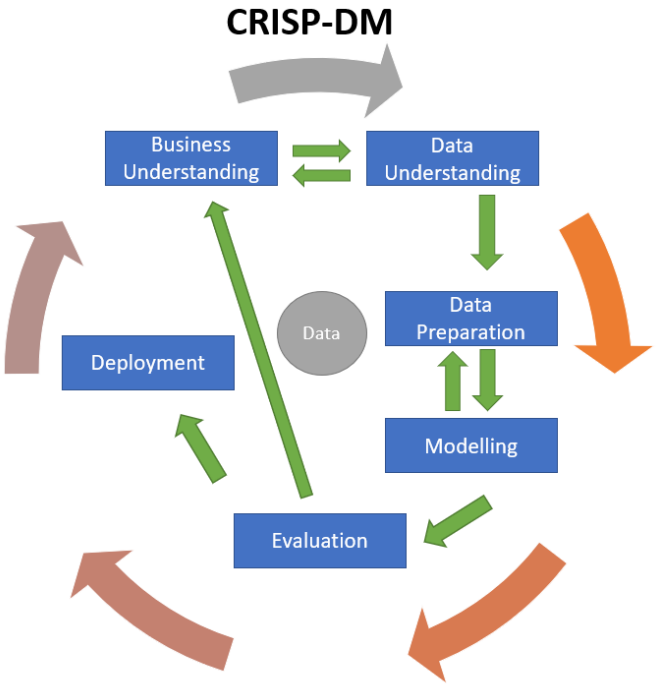

Figura 1. Metodología del desarrollo del modelo a utilizar

Comprensión del negocio. La automatización de tareas, la toma de decisiones y ejecución de acciones mediante diálogos, son factores críticos en las organizaciones; por ello las soluciones como los chatbots agregan valor a una organización, dado que permiten mejorar la atención al usuario y, por tanto, satisfacer su necesidad.
Comprensión de los datos. Recolección inicial de los datos, identificación de problemas de calidad y conocimiento preliminar sobre los datos, con la finalidad de formular hipótesis.
Preparación de los datos. Pre procesamiento de los datos para construir el conjunto final de datos.
Modelado. En esta fase, se seleccionan y aplican las técnicas de modelado, y se calibran los parámetros para obtener métricas de validación óptimas.
Evaluación. Con los modelos construidos se comparan los resultados con los objetivos de negocio.
Despliegue. Creación del modelo final y puesta en producción para que el usuario pueda usarlo.

REFERENCIAS

[1] Bing Lu, Ian Lane, «Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling», 17th Annual Conference of the International Speech Communication Association, Volume 08-12-September-2016, Pages 685 - 689, 2016.
[2] Schumann R., Angkititrakul P.,«Incorporating ASR Errors with Attention-Based, Jointly Trained RNN for Intent Detection and Slot Filling», IEEE International

Conference on Acoustics, Speech, and Signal Processing, Volume 2018-April, Pages 6059 - 6063, 2018.

[3] Wang, Y., Tang, L., He, T., «Attention-based CNN-BLSTM networks for joint intent detection and slot filling», 17th China National Conference on Computational Linguistics, CCL 2018 and 6th International Symposium on Natural Language Processing Based on Naturally Annotated Big Data, Volume 11221 LNAI, Pages 250 - 261, 2018

[4] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, “Using recurrent neural networks for slot filling in spoken language understanding,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 530–539, 2015.

[5] D. Guo, G. Tur, W. Yih and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks," 2014 IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 554-559, doi: 10.1109/SLT.2014.7078634.

[6] P. Xu and R. Sarikaya, "Convolutional neural network based triangular CRF for joint intent detection and slot filling," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 78-83, doi: 10.1109/ASRU.2013.6707709.

[7] R. Wirth and J. Hipp, "CRISP -DM: Towards a standard process model for data mining", Proc. 4th Intl. Conference on Practical Applications of Knowledge Discovery and Data mining, pp. 29 -39, 2000

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[9] Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In Advances in neural information processing systems (pp. 2204-2212).

[10] Hemphill, C. T., Godfrey, J. J., & Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990.

[11] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.