Ejercicio práctico: Predicción de puntuación. Parte 3. Apartado 7

Julián María Galindo Álvarez

En este apartado vamos a investigar sobre otros enfoques y bibliografía relacionada con la tarea de predicción de puntuación abordada a lo largo del trabajo. En concreto vamos a centrarnos en dos artículos:

- Self-Attention Based Model For Punctuation Prediction Using Word And Speech Embeddings (Jiangyan Yi, Jianhua Tao)[9]
- Adversarial Transfer Learning for Punctuation Restoration (Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, Cunhang Fan) [10]

El primero de los trabajos propone utilizar un modelo basado en el mecanismo de autoatención para predecir los signos de puntuación de las secuencias de palabras, al igual que el modelo T-BRNN-pre [7] que hemos estudiado en el apartado 6. El segundo aborda la tarea proponiendo transferencia de aprendizaje adversaria a partir de un encoder bidireccional preentrenado basado en transformers (BERT).

Cabe mencionar que ambos artículos son de los mismos autores, siendo el primero de 2019 y el segundo de 2020.

1. Self-Attention Based Model For Punctuation Prediction Using Word And Speech Embeddings

Este trabajo propone utilizar un modelo basado en la autoatención para predecir los signos de puntuación de las secuencias de palabras. El modelo se entrena utilizando embeddings de palabras y del habla que se obtienen del preentrenamiento de Word2Vec [5] y Speech2Vec [3], respectivamente. Así, el modelo puede utilizar cualquier tipo de datos textuales y del habla. Los experimentos se realizan con los conjuntos de datos IWSLT2011 [1] en inglés

y los resultados muestran que el modelo supera al modelo anterior del estado del arte (BLSTM-CRF [11]) en hasta un 7,8 % de puntuación F1 global. Los resultados también muestran que se obtiene una mejora del rendimiento de hasta un 4,7 % en la puntuación F1 global con respecto al mejor modelo ensemble anterior (Teacher-Ensemble [11]).

1.1. Arquitectura

El codificador consiste en una pila de N capas idénticas, como se muestra en la izquierda de la Fig.1. Cada capa tiene dos subcapas: la primera es un mecanismo de autoatención multicabeza y la segunda es una red totalmente conectada.

El decodificador también está compuesto por una pila de N capas idénticas, como se muestra en la parte derecha de la Fig.1. Además se añade una subcapa de mecanismo de atención multicabeza.

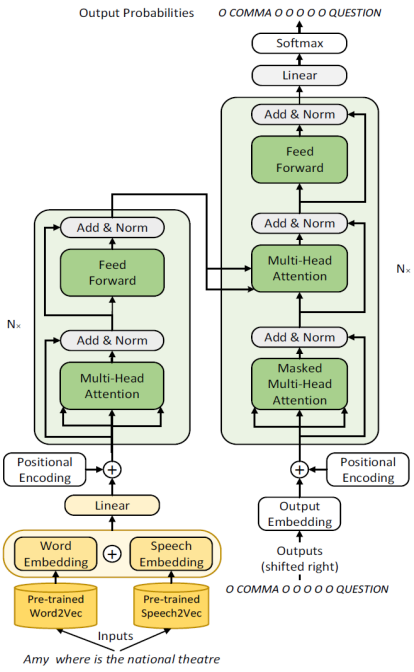

Figura 1: Arquitectura del modelo basado en la autoatención para predicción de la puntuación.

Al igual que en otros modelos de secuencias, los embeddings aprendidos se utilizan para convertir los tokens de salida en vectores. Este trabajo usa modelos preentrenados Word2Vec [5] y Speech2Vec [3] como embeddings de entrada, como se muestra en la parte inferior izquierda de la Fig.1.

Las entradas del modelo son secuencias de palabras, por ejemplo, "Amy where is the national theatre". Las salidas son signos de puntuación, como "O COMMA O O O O QUESTION".

Los parámetros que han usado para evaluar el modelo son: el codificador está compuesto por una pila de $N = 6$ capas idénticas. El decodificador también está compuesto por una pila de $N = 6$ capas idénticas. Hay $h = 8$ capas de atención paralela (cabezas).

1.2. Experimentos

los autores realizan experimentos con conjuntos de datos IWSLT [1] en inglés que contienen charlas TED, como el usado en nuestro trabajo. El conjunto de entrenamiento contiene alrededor de 2,1 millones de palabras y 144.000 frases. El conjunto de validación tiene unas 296.000 palabras y 21.000 frases. El conjunto de pruebas contiene unas 13.000 palabras y 860 frases. Los conjuntos de datos contienen tres tipos de signos de puntuación (COMMA, PERIOD y QUESTION) y un signo de no puntuación.

En Tabla 1 se compara el rendimiento del modelo con el estado del arte. Entre ellos se encuentra el modelo estudiado en el apartado 6, el T-BRNN-pre [7].

Model	COMMA			PERIOD			QUESTION			Overall		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
T-LSTM [6]	49,6	41,4	45,1	60,2	53,4	56,6	57,1	43,5	49,4	55,0	47,2	50,8
T-BRNN-pre [7]	65,5	47,1	54,8	73,3	72,5	72,9	70,7	63,0	66,7	70,0	59,7	64,4
BLSTM-CRF [11]	58,9	59,1	59,0	68,9	72,1	70,5	71,8	60,6	65,7	66,5	63,9	65,1
Teacher-Ensemble [11]	66,2	59,9	62,9	75,1	73,7	74,4	72,3	63,8	67,8	71,2	65,8	68,4
Self-attention-word-speech	**67,4**	**61,1**	**64,1**	**82,5**	**77,4**	**79,9**	**80,1**	**70,2**	**74,8**	**76,7**	**69,6**	**72,9**

Tabla 1: Resultados y comparativa del modelo Self-attention-word-speech.

Al comparar el modelo de Self-attention-word-speech con el anterior modelo BLSTM-CRF [11] del estado del arte, la puntuación F1 global mejora en un 7,8 % en el conjunto de prueba y también supera al mejor modelo ensemble Teacher-Ensemble [11] en un 4,5 %.

2. Adversarial Transfer Learning for Punctuation Restoration

Los embeddings de palabras y las etiquetas de parte del habla (POS) son útiles para las tareas de predicción de la puntuación. Sin embargo, existen dos principales inconvenientes. El primero es que los embeddings de palabras están preentrenados por modelos lingüísticos unidireccionales. Así, los embeddings sólo contienen información de contexto de izquierda a derecha. El otro es que las etiquetas POS las proporciona un etiquetador externo. Por lo tanto, el coste de computación se incrementa y la predicción incorrecta de las etiquetas puede afectar al rendimiento de la predicción de los signos de puntuación durante la decodificación. Los autores proponen el transferencia de aprendizaje adversarial para resolver estos problemas. Se utiliza un modelo de codificador bidireccional preentrenado a partir de transformadores (BERT) para inicializar un modelo de puntuación. Así, los parámetros del modelo transferido llevan representaciones tanto de izquierda a derecha como de derecha a izquierda. Además, usan una tarea adicional de etiquetado POS para ayudar al entrenamiento de la tarea de predicción de la puntuación.

Los experimentos se realizan con los conjuntos de datos de IWSLT2011 mencionado anteriormente. Los resultados muestran que los modelos de predicción de puntuación entrenados con parámetros transferidos del modelo BERT preentrenado obtienen ganancias de rendimiento significativas sobre los modelos con inicialización aleatoria, hasta un 9,4 % de puntuación F1 global en el conjunto de pruebas. Los resultados también demuestran que los modelos de predicción de la puntuación obtienen una mejora adicional del rendimiento con el conocimiento invariante de la tarea de etiquetado POS. El mejor modelo de los autores (BERT-BLSTM-CRF) supera al anterior modelo de última generación (DRNN-LWMA-pre [4]) en hasta un 9,2 % de puntuación F1 absoluta en el conjunto de pruebas.

2.1. Arquitectura

En este trabajo los autores tratan de transferir los parámetros de un modelo BERT preentrenado para inicializar un modelo de predicción de puntuación, como se muestra en la Fig. 2, inspirados en los prometedores resultados del modelo de codificador bidireccional preentrenado a partir de transformadores (BERT) en muchas tareas de procesamiento del lenguaje natural (PLN). El modelo BERT se entrena fusionando el contexto de las direcciones izquierda y derecha. A diferencia de los embeddings de palabras,

los parámetros transferidos contienen representaciones tanto de izquierda a derecha como de derecha a izquierda.

El modelo consta de capas BERT y BLSTM-CRF Fig. 2. Las capas BERT proceden de un modelo BERT preentrenado. Las capas BLSTM-CRF están motivadas por el trabajo [8].

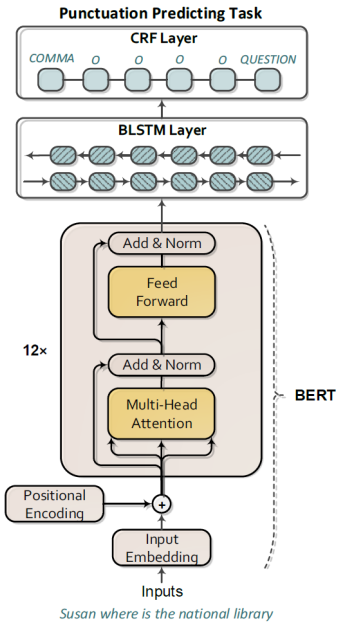

Figura 2: La arquitectura del modelo BERT-BLSTM-CRF. Las capas BERT son inicializadas por un modelo de representación del lenguaje preentrenado. Las BLSTM-CRF se inicializan de forma aleatoria.

En el modelo adversarial BERT-BLSTM-CRF Fig. 3, las capas de tareas compartidas proceden del modelo BERT preentrenado, que tiene una pila de 12 capas idénticas. Los clasificadores de tareas específicas se utilizan para una tarea de predicción de puntuación y una tarea de etiquetado POS, respectivamente. Ambos están formados por capas BLSTM-CRF. Se introduce una capa de inversión de gradiente para garantizar que las distribuciones de características en todas las tareas sean lo más indistinguibles posible para el discriminador de tareas. Las salidas del discriminador de tareas son las etiquetas de las tareas: PUN y POS (PUN indica la tarea de predicción de la puntuación y POS se refiere a la tarea de etiquetado POS).

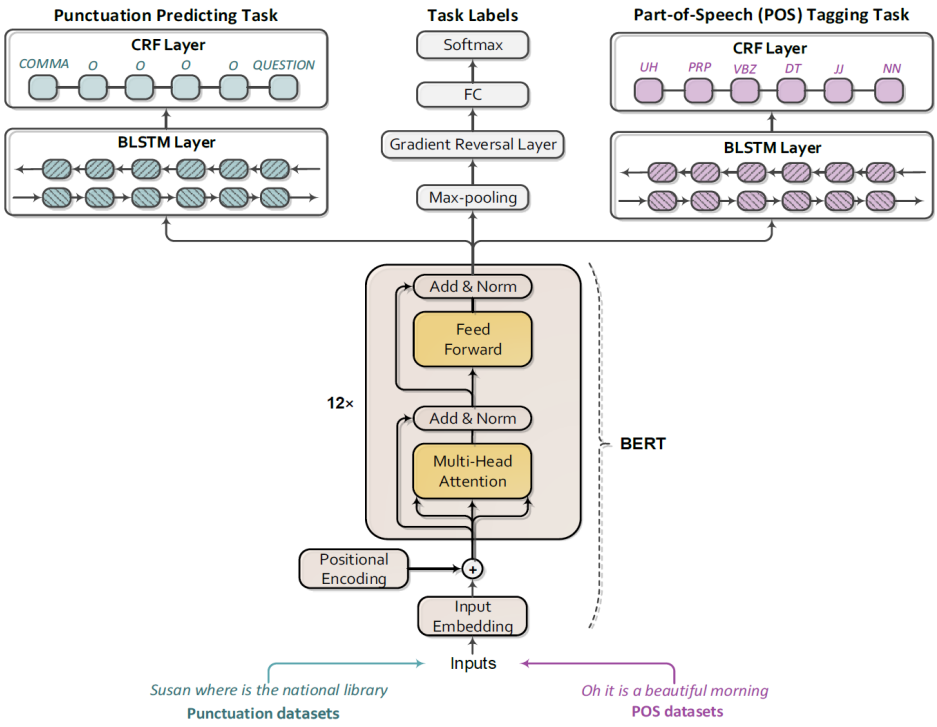

Figura 3: La arquitectura del modelo adversarial.

2.2. Experimentos

Los experimentos se realizan sobre el IWSLT [1], el mismo conjunto ya visto con frases de las TED Talks.

El modelo BERT-BLSTM-CRF consigue mejores resultados que el anterior modelo de vanguardia DRNN-LWMA-pre [4] en hasta un 9,2 % puntuación F1 absoluta en el conjunto de pruebas. Además se ve como supera en varios puntos tanto al modelo visto en el apartado 6 (T-BRNN-pre [7]) como el primer modelo que hemos visto en este apartado, de los mismos autores Self-attention [9].

Model	COMMA			PERIOD			QUESTION			Overall		
F_1		P	R	F_1	P	R	F_1	P	R	F_1	P	R
CRF Best [8]	–	–	–	–	–	–	–	–	–	49,8	58,0	53,5
DNN-A [2]	48,6	42,4	45,3	59,7	68,3	63,7	–	–	–	54,8	53,6	54,2
CNN-2A [2]	48,1	44,5	46,2	57,6	69,0	62,8	–	–	–	53,4	55,0	54,2
T-LSTM [6]	49,6	41,4	45,1	60,2	53,4	56,6	57,1	43,5	49,4	55,0	47,2	50,8
T-BRNN-pre [7]	65,5	47,1	54,8	73,3	72,5	72,9	70,7	63,0	66,7	70,0	59,7	64,4
BLSTM-CRF [11] s	58,9	59,1	59,0	68,9	72,1	70,5	71,8	60,6	65,7	66,5	63,9	65,1
Teacher-Ensemble [11]	66,2	59,9	62,9	75,1	73,7	74,4	72,3	63,8	67,8	71,2	65,8	68,4
DRNN-LWMA-pre [4]	62,9	60,8	61,9	77,3	73,7	75,5	69,6	69,6	69,6	69,9	67,2	68,6
Self-attention [9]	67,4	61,1	64,1	82,5	77,4	79,9	80,1	70,2	74,8	76,7	69,6	72,9
Mejor modelo BERT-BLSTM-CRF	**76,2**	**71,2**	**73,6**	**87,3**	**81,1**	**84,1**	**79,1**	**72,7**	**75,8**	**80,9**	**75,0**	**77,8**

Tabla 2: Resultados y comparativa con el estado del arte.

3. Conclusión

El modelo basado en la autoatención puede aprender características léxicas y acústicas utilizando cualquier tipo de datos de texto sin el correspondiente audio y datos de habla sin el correspondiente texto. Los resultados experimentales en los conjuntos de datos de IWSLT2011 en inglés demuestran que el método propuesto es eficaz, superando el estado del arte en ese momento.

El modelo basado en aprendizaje de transferencia adversarial para mejorar el rendimiento de las tareas de predicción de puntuación permite representaciones bidireccionales que se transfieren desde un modelo BERT preentrenado a los modelos de predicción de puntuación. Los experimentos realizados sobre el mismo conjunto demuestran que el modelo supera a los modelos anteriores más avanzados, incluido el basado en autatención.

Referencias

[1] Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658, 2016.

[2] Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. Punctuation prediction for unsegmented transcript based on word vector. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 654–658, 2016.

[3] Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*, 2018.

[4] Seokhwan Kim. Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7280–7284. IEEE, 2019.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] Ottokar Tilk and Tanel Alumäe. Lstm for punctuation restoration in speech transcripts. In *Sixteenth annual conference of the international speech communication association*, 2015.

[7] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech*, pages 3047–3051, 2016.

[8] Nicola Ueffing, Maximilian Bisani, and Paul Vozila. Improved models for automatic punctuation prediction for spoken and written text. In *Interspeech*, pages 3097–3101, 2013.

[9] Jiangyan Yi and Jianhua Tao. Self-attention based model for punctuation prediction using word and speech embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7270–7274. IEEE, 2019.

[10] Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, and Cunhang Fan. Adversarial transfer learning for punctuation restoration. *arXiv preprint arXiv:2004.00248*, 2020.

[11] Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Ya Li, et al. Distilling knowledge from an ensemble of models for punctuation prediction. In *Interspeech*, pages 2779–2783, 2017.