Enfoques eficaces para la traducción automática neuronal basada en la atención

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang,hyhieu,manning}@stanford.edu

Abstract

Últimamente se ha utilizado un mecanismo de atención para mejorar la traducción automática neuronal (NMT) al centrarse selectivamente en partes de la oración fuente durante la traducción. Sin embargo, se ha trabajado poco en la exploración de arquitecturas útiles para NMT basada en la atención. Este artículo examina dos clases simples y efectivas de mecanismos de atención: un enfoque global que siempre atiende a todas las palabras fuente y uno local que solo analiza un subconjunto de palabras fuente a la vez. Demostramos la efectividad de ambos enfoques en las tareas de traducción WMT entre inglés y alemán en ambas direcciones. Con atención local conseguimos una importante ganancia de 5,0 puntos BLEU respecto a sistemas no atencionales que ya incorporan técnicas conocidas como el abandono. Nuestro modelo de conjunto que utiliza diferentes arquitecturas de atención produce un nuevo resultado de última generación en la tarea de traducción del inglés al alemán del WMT'15 con 25,9 puntos BLEU, una mejora de 1,0 puntos BLEU con respecto al mejor sistema existente respaldado por NMT y reranker de n-gramos [1].

1 Introducción

La traducción automática neuronal (NMT) logró rendimientos de vanguardia en tareas de traducción a gran escala, como del inglés al francés (Luong et al., 2015) y del inglés al alemán (Jean et al., 2015). NMT es atractivo porque requiere un conocimiento mínimo del dominio y es conceptualmente simple. El modelo de (Luong et al., 2015) lee todas las palabras fuente hasta llegar al símbolo de final de oración <eos>. Luego comienza a emitir una palabra objetivo a la vez, como se ilustra en la figura 1. NMT suele ser una gran red neuronal entrenada de un extremo a otro y tiene la capacidad de generalizar bien a secuencias de palabras muy largas. Esto significa que el modelo no tiene que almacenar explícitamente gigantescas tablas de frases y modelos de lenguaje como en el caso de la MT estándar; por lo tanto, NMT ocupa una pequeña cantidad de memoria. Por último, implementar decodificadores NMT es fácil a diferencia de los decodificadores altamente complejos en MT estándar (Koehn et al., 2003).

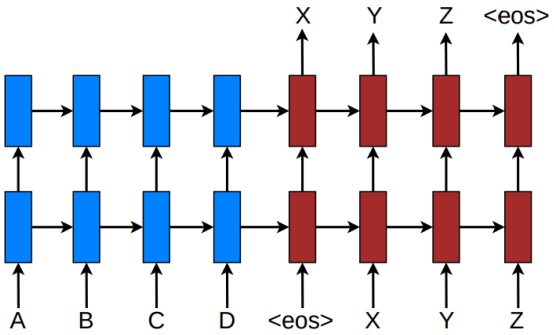

Figure 1: Traducción automática neuronal: una arquitectura recurrente de apilamiento para traducir una secuencia de origen A B C D en una secuencia de destino X Y Z. Aquí, <eos> marca el final de una oración.

Paralelamente, el concepto de "atención" ha ganado popularidad recientemente en el entrenamiento de redes neuronales, permitiendo a los modelos aprender alineamientos entre diferentes modalidades, por ejemplo, entre objetos de imagen y acciones de agentes en el problema de control dinámico (Mnih et al., 2014), entre marcos de voz y texto en la tarea de reconocimiento de voz (?), o entre las características visuales de una imagen y su descripción de texto en la tarea de generación de leyendas de imagen (Xu et al., 2016). En el contexto del NMT, (Bahdanau et al., 2016) ha aplicado con éxito dicho mecanismo de atención para traducir y alinear palabras de forma conjunta. Hasta donde sabemos, no ha habido ningún otro trabajo que explore el uso de arquitecturas basadas en la atención para NMT.

En este trabajo, diseñamos, teniendo en cuenta la simplicidad y la eficacia, dos tipos novedosos de modelos basados en la atención: un enfoque global en el que se atienden todas las palabras fuente y uno local en el que sólo se considera un subconjunto de palabras fuente a la vez. El primer enfoque se

[1] Todo nuestro código y modelos están disponibles públicamente en http://nlp.stanford.edu/projects/nmt

parece al modelo de (Bahdanau et al., 2016) pero es arquitectónicamente más simple. Este último puede verse como una combinación interesante entre los modelos de atención dura y blanda propuestos en (Xu et al., 2016): es computacionalmente menos costoso que el modelo global o la atención blanda; al mismo tiempo, a diferencia de la atención dura, la atención local es diferenciable en casi todas partes, lo que facilita su implementación y entrenamiento [2]. Además, también examinamos varias funciones de alineación para nuestros modelos basados en la atención. Experimentalmente, demostramos que ambos enfoques son efectivos en las tareas de traducción WMT entre inglés y alemán en ambas direcciones. Nuestros modelos atencionales producen un aumento de hasta 5.0 BLEU sobre los sistemas no atencionales que ya incorporan técnicas conocidas como la deserción. Para la traducción del inglés al alemán, logramos nuevos resultados de última generación (SOTA) tanto para WMT'14 como para WMT'15, superando en más de 1,0 el rendimiento de los sistemas SOTA anteriores, respaldados por modelos NMT y reordenadores LM de n-gramas. AZUL. Realizamos análisis exhaustivos para evaluar nuestros modelos en términos de aprendizaje, la capacidad de manejar oraciones largas, opciones de arquitecturas de atención, calidad de alineación y resultados de traducción.

2 Traducción automática neuronal

Un sistema de traducción automática neuronal es una red neuronal que modela directamente la probabilidad condicional $p(y|x)$ de traducir una oración fuente, $x1, ..., x_n$, a una oración objetivo, $y1, ..., y_m$ [3]. Una forma básica de NMT consta de dos componentes: (a) un codificador que calcula una representación s para cada oración fuente y (b) un decodificador que genera una palabra objetivo a la vez y, por lo tanto, descompone la probabilidad condicional como:

$$logp(y|x) = \sum_{j=1}^{m} logp(y_j|y_{<j}, s) \qquad (1)$$

Una opción natural para modelar dicha descomposición en el decodificador es utilizar una arquitectura de red neuronal recurrente (RNN), en la que la mayoría de los trabajos recientes de NMT, como (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Kalchbrenner and Blunsom, 2013; Luong et al., 2015; Jean et al., 2015) tienen en común. Sin embargo, difieren en términos de f qué arquitecturas RNN se utilizan para el decodificador y cómo el codificador calcula la representación de la oración fuente. (Kalchbrenner and Blunsom, 2013) utilizaron un RNN con la unidad oculta estándar para el decodificador y una red neuronal convolucional para codificar la representación de la oración fuente. Por otro lado, tanto (Sutskever et al., 2014) y (Luong et al., 2015) apilaron varias capas de un RNN con una unidad oculta de memoria a corto plazo (LSTM) tanto para el codificador como para el decodificador. (Cho et al., 2014), (Bahdanau et al., 2016), y (Jean et al., 2015) adoptaron una versión diferente del RNN con una unidad oculta inspirada en LSTM, la unidad recurrente cerrada (GRU), para ambos componentes [4]. Con más detalle, se puede parametrizar la probabilidad de decodificar cada palabra y_j como:

$$p(y_j|y_{<j}, s) = softmax(g(h_j)) \qquad (2)$$

Siendo g la función de transformación que genera un vector del tamaño de un vocabulario [5]. Aquí, h_j es la unidad oculta RNN, calculada de manera abstracta como:

$$h_j = f(h_{j-1}, s), \qquad (3)$$

Donde f calcula el estado oculto actual dado el estado oculto anterior y puede ser una unidad RNN básica, una GRU o una unidad LSTM. En (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014), la representación de origen s solo se usa una vez para inicializar el estado oculto del decodificador. Por otro lado, en (Bahdanau et al., 2016; Jean et al., 2015) y este trabajo, de hecho, implica un conjunto de estados ocultos de la fuente que se consultan durante todo el proceso de traducción. Este enfoque se denomina mecanismo de atención, que analizaremos a continuación. En este trabajo, a continuación (Sutskever et al., 2014; Luong et al., 2015), utilizamos la arquitectura LSTM

[2] Hay un trabajo reciente de (Gregor et al., 2015), que es muy similar a nuestra atención local y se aplica a la tarea de generación de imágenes. Sin embargo, como detallamos más adelante, nuestro modelo es mucho más simple y puede lograr un buen rendimiento para NMT.

[3] Se supone que todas las oraciones terminan con una ficha especial de "fin de oración" $< eos >$

[4] Todos utilizaron una única capa RNN, excepto los dos últimos trabajos que utilizaron un RNN bidireccional para el codificador.

[5] Se pueden proporcionar a g otras entradas, como la palabra actualmente predicha yj como en (Bahdanau et al., 2016)

de apilamiento para nuestros sistemas NMT, como se ilustra en la figura 1. Usamos la unidad LSTM definida en (Zaremba et al., 2015). Nuestro objetivo formativo se formula de la siguiente manera:

$$J_t = \sum_{(x,y)\in\mathbb{D}} -log\rho(y|x) \quad (4)$$

siendo $\mathbb{D}$ nuestro corpus de entrenamiento paralelo

3 modelos basados en la atención

Nuestros diversos modelos basados en la atención se clasifican en dos grandes categorías: global y local. Estas clases difieren en términos de si la "atención" se pone en todas las posiciones fuente o sólo en unas pocas posiciones fuente. Ilustramos estos dos tipos de modelos en las Figuras 2 y 3 respectivamente. Estos dos tipos de modelos tienen en común el hecho de que en cada paso de tiempo t en la fase de decodificación, ambos enfoques primero toman como entrada el estado oculto h_t en la capa superior de un LSTM de apilamiento. El objetivo es entonces derivar un vector de contexto c_t que capture información relevante del lado fuente para ayudar a predecir la palabra objetivo actual y_t. Si bien estos modelos difieren en cómo se deriva el vector de contexto c_t, comparten los mismos pasos posteriores. Específicamente, dado el estado oculto objetivo h_t y el vector de contexto del lado fuente c_t, empleamos una capa de concatenación simple para combinar la información de ambos vectores para producir un estado oculto de atención de la siguiente manera:

$$\bar{h}_t = tanh(W_c[c_t; h_t]) \quad (5)$$

Luego, el vector de atención $\bar{h}_t$ pasa a través de la capa softmax para producir la distribución predictiva formulada como:

$$p(y_j|y_{<j}, s) = softmax(W_s\bar{h}_t) \quad (6)$$

Ahora detallamos cómo cada tipo de modelo calcula el vector de contexto del lado fuente c_t.

3.1 Atención Global

La idea de un modelo de atención global es considerar todos los estados ocultos del codificador al derivar el vector de contexto c_t. En este tipo de modelo, un vector de alineación de longitud variable a_t, cuyo tamaño es igual al número de pasos de tiempo en el lado de origen, se deriva comparando el estado oculto objetivo actual h_t con cada

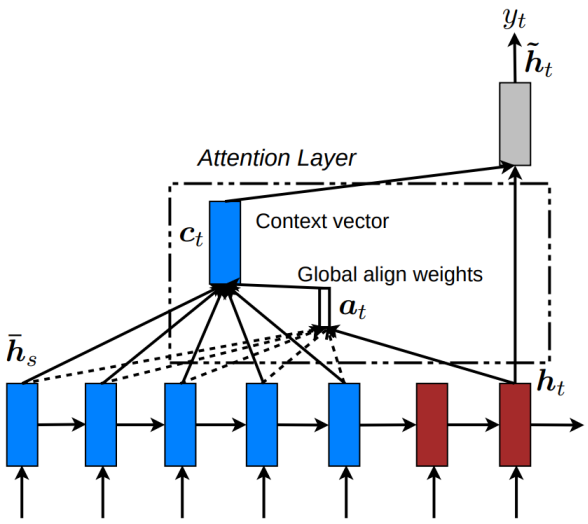

Figure 2: Modelo de atención global: en cada paso de tiempo t, el modelo infiere un vector de peso de alineación de longitud variable basado en el estado objetivo actual h_t y todos los estados fuente $\bar{h}_s$. Luego se calcula un vector de contexto global c_t como el promedio ponderado, según a_t, sobre todo el estado fuente.

estado oculto de origen $\bar{h}_t$:

$$a_t(s) = align(h_t, \bar{h}_s) = \frac{exp(score(h_t, \bar{h}_s))}{\sum_{s^t} exp(score(h_t, \bar{h}_{s^t}))} \quad (7)$$

Aquí, la puntuación se refiere a una función basada en contenido para la cual consideramos tres alternativas diferentes:

$$score(h_t, \bar{h}_s) = \begin{cases} h_t^\top, \bar{h}_s & \text{dot} \\ h_t^\top W_a \bar{h}_s & \text{general} \\ v_a^\top tanh(W_a[h_t; \bar{h}_s]) & \text{concat} \end{cases}$$

Además, en nuestros primeros intentos de construir modelos basados en la atención, utilizamos una función basada en la ubicación en la que las puntuaciones de alineación se calculan únicamente a partir del estado oculto objetivo h_t de la siguiente manera:

$$a_t = softmax(W_a h_t) \quad \text{locación} \quad (8)$$

Dado el vector de alineación como pesos, el vector de contexto ct se calcula como el promedio ponderado de todos los estados ocultos de origen [6].
Comparación con (Bahdanau et al., 2016): si bien nuestro enfoque de atención global es similar en

[6]Ecuacion 8 implica que todos los vectores de alineación en tienen la misma longitud. Para oraciones cortas, solo usamos la parte superior de at y para oraciones largas, ignoramos las palabras cerca de e.

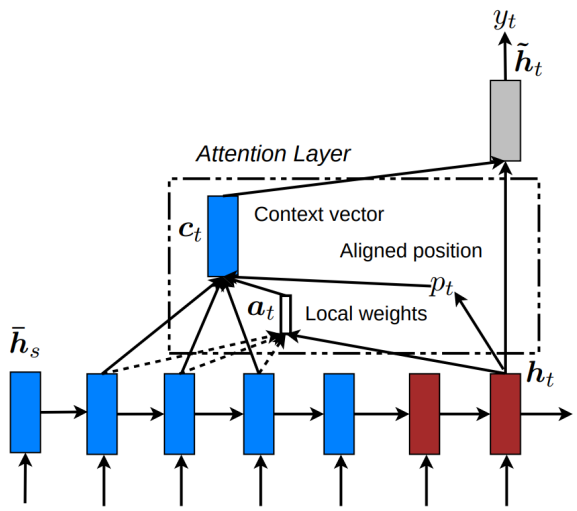

Figure 3: Modelo de atención local: el modelo primero predice una única posición alineada para la palabra objetivo actual. Luego se utiliza una ventana centrada alrededor de la posición de origen p_t para calcular un vector de contexto c_t, un promedio ponderado de los estados ocultos de origen en la ventana. Los pesos en se infieren del estado objetivo actual h_t y de los estados fuente $\bar{h}_s$ en la ventana.

espíritu al modelo propuesto por (Bahdanau et al., 2016), existen varias diferencias clave que reflejan cómo hemos simplificado y generalizado el modelo original. Primero, simplemente usamos estados ocultos en las capas superiores de LSTM tanto en el codificador como en el decodificador, como se ilustra en la Figura 2. (Bahdanau et al., 2016), por otro lado, utilizan la concatenación de los estados ocultos de origen hacia adelante y hacia atrás en el codificador bidireccional y los estados ocultos de destino en su decodificador unidireccional no apilable. En segundo lugar, nuestra ruta de cálculo es más sencilla; vamos de $h_t \rightarrow a_t \rightarrow c_t \rightarrow \bar{h}_t$ y luego hacemos una predicción como se detalla en la ecuación 5, ecuación 6 y Figura 2. Por otro lado, en cualquier momento t, (Bahdanau et al., 2016) construyen a partir del estado oculto anterior $h_{t-1} \rightarrow a_t \rightarrow c_t \rightarrow h_t$, que, a su vez, pasa por una capa de salida profunda y una capa de salida máxima antes de hacer predicciones [7]. Por último, (Bahdanau et al., 2016) solo experimentaron con una función de alineación, el producto $concat$; mientras que más adelante mostraremos que las otras alternativas son mejor.

[7]Nos referiremos nuevamente a esta diferencia en la Sección 3.3.

3.2 Atención Local

La atención global tiene el inconveniente de que tiene que atender a todas las palabras del lado de origen para cada palabra de destino, lo cual es costoso y puede hacer potencialmente poco práctico traducir secuencias más largas, por ejemplo, párrafos o documentos. Para abordar esta deficiencia, proponemos un mecanismo de atención local que elige centrarse solo en un pequeño subconjunto de las posiciones de origen por palabra de destino. Este modelo se inspira en la compensación entre los modelos de atención suave y dura propuestos por (Xu et al., 2016) para abordar la tarea de generación de leyendas de imágenes. En su trabajo, la atención suave se refiere al enfoque de atención global en el que se colocan pesos "suavemente" sobre todos los parches de la imagen de origen. La atención intensa, por otra parte, selecciona una parte de la imagen para atenderla a la vez. Si bien es menos costoso en el momento de la inferencia, el modelo de atención dura no es diferenciable y requiere técnicas más complicadas, como la reducción de la varianza o el aprendizaje por refuerzo, para entrenar. Nuestro mecanismo de atención local se centra selectivamente en una pequeña ventana de contexto y es diferenciable. Este enfoque tiene la ventaja de evitar los costosos cálculos incurridos en la atención suave y, al mismo tiempo, es más fácil de entrenar que el enfoque de atención dura. En detalles concretos, el modelo primero genera una posición alineada p_t para cada palabra objetivo en el momento t. Luego, el vector de contexto c_t se deriva como un promedio ponderado sobre el conjunto de estados ocultos de origen dentro de la ventana $[p_t - D, p_t + D]$; D se selecciona empíricamente [8]. A diferencia del enfoque global, el vector de alineación local en ahora es de dimensión fija, es decir, $\epsilon\mathbb{R}^{2D+1}$. Consideramos dos variantes del modelo como se muestra a continuación. Alineación monótona (local-m): simplemente configuramos $p_t = t$ asumiendo que las secuencias fuente y objetivo están alineadas de manera más o menos monótona. El vector de alineación en se define según la ecuación 7 [9]. Alineación predictiva (local-p): en lugar de asumir alineaciones monótonas, nuestro modelo predice una posición alineada de la siguiente manera:

[8]Si la ventana cruza los límites de la oración, simplemente ignoramos la parte exterior y consideramos las palabras en la ventana.

[9]local-m es igual que el modelo global excepto que el vector a_t tiene una longitud fija y es más corto.

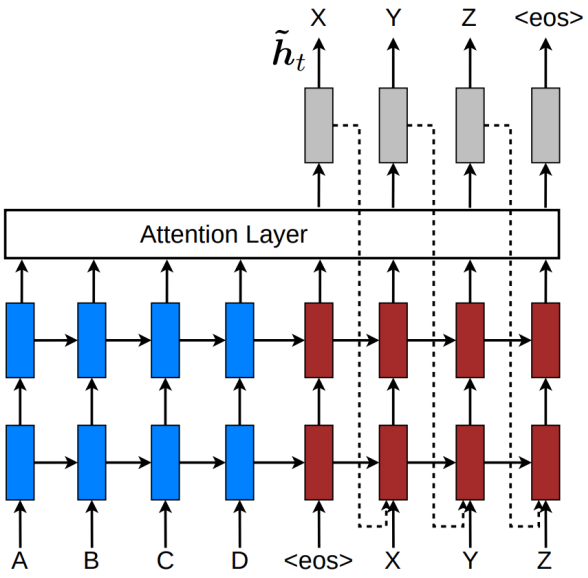

Figure 4: Enfoque de alimentación de entradas: los vectores de atención $\bar{h}_t$ se alimentan como entradas para los siguientes pasos de tiempo para informar al modelo sobre decisiones de alineación pasadas.

$$p_t = S \cdot sigmoid(v_p^\top tanh(W_p h_t)) \quad (9)$$

W_p y v_p son los parámetros del modelo que se aprenderán para predecir posiciones. S es la longitud de la oración fuente. Como resultado de sigmoide, $p_t \epsilon [0, S]$. Para favorecer los puntos de alineación cerca de p_t, colocamos una distribución gaussiana centrada alrededor de pt. Específicamente, nuestros pesos de alineación ahora se definen como:

$$a_t(s) = align(h_t, \bar{h}_s) = (-\frac{(s - p_t)^2}{2\sigma^2}) \quad (10)$$

Usamos la misma función de alineación que en la ecuación 7 y la desviación estándar se establece empíricamente como $\sigma = \frac{D}{2}$. Tenga en cuenta que p_t es un número real; mientras que s es un número entero dentro de la ventana centrada en p_t [10].

En comparación con (Gregor et al., 2015), han propuesto un mecanismo de atención selectiva, muy similar a nuestra atención local, para la tarea de generación de imágenes. Su enfoque permite al modelo seleccionar un parche de imagen de diferente ubicación y zoom. En cambio, utilizamos el mismo "zoom" para todas las posiciones de destino, lo que simplifica enormemente la formulación y aun así logra un buen rendimiento.

3.3 Enfoque de alimentación de insumos

En nuestros enfoques globales y locales propuestos, las decisiones de atención se toman de forma independiente, lo cual no es óptimo. Mientras que, en la traducción automática estándar, a menudo se mantiene un conjunto de cobertura durante el proceso de traducción para realizar un seguimiento de qué palabras fuente se han traducido. Del mismo modo, en los NMT atencionales, las decisiones de alineación deben tomarse de manera conjunta teniendo en cuenta la información de alineación pasada. Para abordar esto, proponemos un enfoque de alimentación de entradas en el que los vectores de atención $\bar{h}_t$ se concatenan con entradas en los siguientes pasos de tiempo, como se ilustra en la Figura 4 [11]. Los efectos de tener tales conexiones son dobles: (a) esperamos hacer que el modelo plenamente conscientes de las opciones de alineación anteriores y (b) creamos una red muy profunda que se extiende tanto horizontal como verticalmente. Comparación con otros trabajos de (Bahdanau et al., 2016) utilizan vectores de contexto, similares a nuestro c_t, para construir estados ocultos posteriores, que también pueden lograr el efecto de "cobertura". Sin embargo, no se ha realizado ningún análisis sobre si dichas conexiones son útiles como se hace en este trabajo. Además, nuestro enfoque es más general; Como se ilustra en la Figura 4, se puede aplicar a arquitecturas recurrentes de apilamiento generales, incluidos los modelos no atencionales. (Xu et al., 2016) proponen un enfoque doblemente atencional con una restricción adicional agregada al objetivo de entrenamiento para garantizar que el modelo preste igual atención a todas las partes de la imagen durante el proceso de generación de subtítulos. Esta restricción también puede ser útil para capturar el efecto del conjunto de cobertura en NMT que mencionamos anteriormente. Sin embargo, optamos por utilizar el enfoque de alimentación de entradas, ya que proporciona flexibilidad al modelo para decidir sobre las restricciones de atención que considere adecuadas.

[10] local-p es similar al modelo local-m excepto que calculamos dinámicamente p_t y usamos una distribución gaussiana truncada para modificar los pesos de alineación originales $align(h_t, \bar{h}_s)$ como se muestra en la ecuación 10. Al utilizar p_t para derivar, podemos calcular gradientes de backprop para W_p y v_p. Este modelo es diferenciable en casi todas partes.

[11] Si n es el número de celdas LSTM, el tamaño de entrada de la primera capa LSTM es $2n$; los de las capas posteriores son n.

System	Ppl	BLEU
Winning WMT'14 system – phrase-based + large LM (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + ensemble 8 models (Jean et al., 2015)		21.6
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+1.3)
Base + reverse + dropout	8.1	14.0 (+1.4)
Base + reverse + dropout + global attention (location)	7.3	16.8 (+2.8)
Base + reverse + dropout + global attention (location) + feed input	6.4	18.1 (+1.3)
Base + reverse + dropout + local-p attention (general) + feed input	5.9	19.0 (+0.9)
Base + reverse + dropout + local-p attention (general) + feed input + unk replace		20.9 (+1.9)
Ensemble 8 models + unk replace		23.0 (+2.1)

Table 1: Resultados WMT'14 inglés-alemán: se muestran las perplejidades (ppl) y las puntuaciones BLEU tokenizadas de varios sistemas en newstest2014. Destacamos el mejor sistema en negrita y damos mejoras progresivas en cursiva entre sistemas consecutivos. local-p se refiere a la atención local con alineamientos predictivos. Indicamos para cada modelo de atención la función de puntuación de alineación utilizada entre paréntesis.

4 experimentos

Evaluamos la efectividad de nuestros modelos en las tareas de traducción WMT entre inglés y alemán en ambas direcciones. newstest2013 (3000 oraciones) se utiliza como conjunto de desarrollo para seleccionar nuestros hiperparámetros. Los resultados de traducción se informan en BLEU (Papineni et al., 2002), que distingue entre mayúsculas y minúsculas, en newstest2014 (2737 oraciones) y newstest2015 (2169 sen). tensiones). A continuación (Luong et al., 2015), informamos la calidad de la traducción utilizando dos tipos de BLEU: (a) BLEU tokenizado [12] para que sea comparable con el trabajo NMT existente y (b) NIST [13] BLEU para que sea comparable con los resultados de WMT.

4.1 Detalles de la capacitación

Todos nuestros modelos se entrenan con datos de entrenamiento de WMT'14 que constan de 4,5 millones de pares de oraciones (116 millones de palabras en inglés, 110 millones de palabras en alemán). De manera similar a (Jean et al., 2015), limitamos nuestro vocabulario a las 50.000 palabras más frecuentes para ambos idiomas. Las palabras que no están en estos vocabularios preseleccionados se convierten en un token universal. Al entrenar nuestros sistemas NMT, siguiendo (Bahdanau et al., 2016; Jean et al., 2015), filtramos pares de oraciones cuyas longitudes superan las 50 palabras y mezclamos minilotes a medida que avanzamos. Nuestros modelos LSTM apilables tienen 4 capas, cada una con 1000 celdas e incrustaciones de 1000 dimensiones. Seguimos (Sutskever et al., 2014; Luong et al., 2015) en el entrenamiento de NMT con configuraciones similares: (a) nuestros parámetros se inicializan uniformemente en $[-0.1, 0.1]$, (b) entrenamos durante 10 épocas usando ing SGD simple, (c) se emplea un programa de tasa de aprendizaje simple: comenzamos con una tasa de aprendizaje de 1; después de 5 épocas, comenzamos a reducir a la mitad la tasa de aprendizaje en cada época, (d) el tamaño de nuestro mini-lote es 128 y (e) el gradiente normalizado se reescala cada vez que su norma excede 5. Además, también utilizamos el abandono con probabilidad de 0,2 para nuestros LSTM como lo sugiere (Zaremba et al., 2015). Para los modelos de abandono, entrenamos durante 12 épocas y comenzamos a reducir a la mitad la tasa de aprendizaje después de 8 épocas. Para los modelos de atención local, establecemos empíricamente el tamaño de la ventana $D = 10$. Nuestro código está implementado en MATLAB. Cuando se ejecuta en un dispositivo con una sola GPU Tesla K40, logramos una velocidad de 1.000 palabras objetivo por segundo. Se necesitan entre 7 y 10 días para entrenar completamente un modelo.

[12] Todos los textos se tokenizan con tokenizer.perl y las puntuaciones BLEU se calculan con multi-bleu.perl.

[13] Con el script mteval-v13a según las pautas de WMT.

System	BLEU
Top – NMT + 5-gram rerank (Montreal)	24.9
Our ensemble 8 models + unk replace	25.9

Table 2: Resultados WMT'15 inglés-alemán: puntuaciones NIST BLEU de la entrada ganadora en WMT'15 y nuestra mejor en newstest2015.

4.2 Resultados inglés-alemán

Comparamos nuestros sistemas NMT en la tarea inglés-alemán con varios otros sistemas. Estos incluyen el sistema ganador en WMT'14 (Buck et al., 2014), un sistema basado en frases cuyos modelos de lenguaje se entrenaron en un enorme texto monolingüe, el corpus Common Crawl. Para los sistemas NMT de extremo a extremo, hasta donde sabemos, (Jean et al., 2015) es el único trabajo que experimenta con este par de idiomas y actualmente con el sistema SOTA. Solo presentamos resultados para algunos de nuestros modelos de atención y luego analizaremos el resto en la Sección 5.

Como se muestra en la Tabla 1, logramos mejoras progresivas cuando (a) invertimos la oración fuente, +1.3 BLEU, como se propone en (Sutskever et al., 2014) y (b) utilizando abandono escolar, +1,4 BLEU. Además de eso, (c) el enfoque de atención global proporciona un impulso significativo de +2,8 BLEU, lo que hace que nuestro modelo sea ligeramente mejor que el sistema de atención base de (Bahdanau et al., 2016) (fila RNNSearch). Cuando (d) utilizamos el enfoque de alimentación de entrada, obtenemos otra ganancia notable de +1,3 BLEU y superamos a su sistema. El modelo de atención local con alineamientos predictivos (fila local-p) demuestra ser aún mejor, lo que nos brinda una mejora adicional de +0,9 BLEU además del modelo de atención global. Es interesante observar la tendencia reportada previamente en (Luong et al., 2015) de que la perplejidad se correlaciona fuertemente con la calidad de la traducción. En total, logramos una ganancia significativa de 5,0 puntos BLEU sobre la línea de base no atencional, que ya incluye técnicas conocidas como la inversión de fuente y la deserción.

La técnica de reemplazo desconocido propuesta en (Luong et al., 2015; Jean et al., 2015) produce otra buena ganancia de +1,9 BLEU, lo que demuestra que nuestros modelos de atención aprenden alineaciones útiles para trabajos desconocidos. Finalmente, al reunir 8 modelos diferentes de diversas configuraciones, por ejemplo, usando diferentes

System	Ppl.	BLEU
WMT'15 systems		
SOTA – phrase-based (Edinburgh)		29.2
NMT + 5-gram rerank (MILA)		27.6
Our NMT systems		
Base (reverse)	14.3	16.9
+ global (location)	12.7	19.1 (+2.2)
+ global (location) + feed	10.9	20.1 (+1.0)
+ global (dot) + drop + feed		22.8 (+2.7)
+ global (dot) + drop + feed + unk	9.7	24.9 (+2.1)

Table 3: Resultados WMT'15 alemán-inglés: rendimiento de varios sistemas (similar a la Tabla 1). El sistema base ya incluye inversión de fuente a la que agregamos atención global, abandono, alimentación de entrada y reemplazo de tinta.

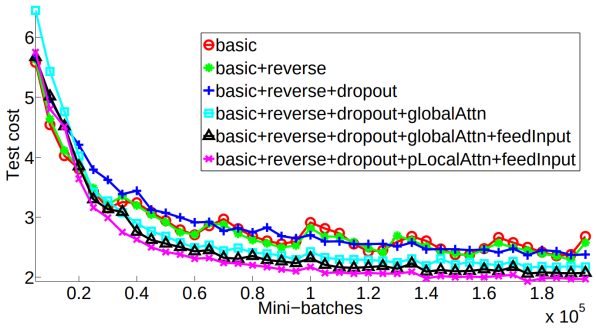

Figure 5: Curvas de aprendizaje: coste de la prueba (con perplejidad) en newstest2014 para NMT inglés-alemán a medida que avanza la formación.

enfoques de atención, con y sin abandono, etc., pudimos lograr un nuevo resultado SOTA de 23,0 BLEU, superando al existente.

Últimos resultados en WMT'15: a pesar de que nuestros modelos se entrenaron en WMT'14 con un poco menos de datos, los probamos en newstest2015 para demostrar que se pueden generalizar bien a diferentes conjuntos de pruebas. Como se muestra en la Tabla 2, nuestro mejor sistema establece un nuevo rendimiento SOTA de 25,9 BLEU, superando al mejor sistema existente respaldado por NMT y un reranker LM de 5-grams en +1,0 BLEU.

4.3 Resultados alemán-inglés

Llevamos a cabo una serie de experimentos similares para la tarea de traducción del WMT'15 del alemán al inglés. Si bien nuestros sistemas aún no han igualado el rendimiento del sistema SOTA, mostramos la efectividad de nuestros enfoques con ganancias grandes y progresivas en términos de

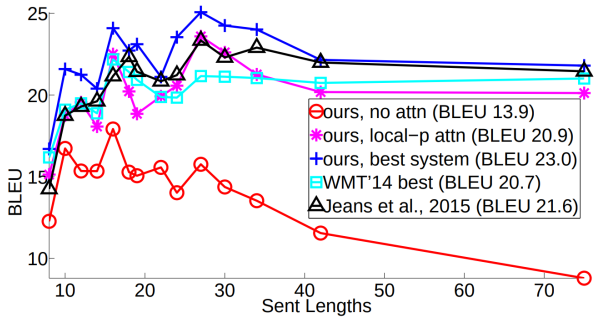

Figure 6: Análisis de longitud: cualidades de traducción de diferentes sistemas a medida que las oraciones se vuelven más largas.

System	Ppl	BLEU	
		Before	After unk
global (location)	6.4	18.1	19.3 (+1.2)
global (dot)	6.1	18.6	20.5 (+1.9)
global (general)	6.1	17.3	19.1 (+1.8)
local-m (dot)	>7.0	x	x
local-m (general)	6.2	18.6	20.4 (+1.8)
local-p (dot)	6.6	18.0	19.6 (+1.9)
local-p (general)	5.9	19	20.9 (+1.9)

Table 4: Arquitecturas atencionales: actuaciones de diferentes modelos atencionales. Entrenamos dos modelos locales-m (punto); ambos tienen personas ppl $>$ 7,0.

BLEU, como se ilustra en la Tabla 3. El mecanismo de atención nos brinda una ganancia de +2,2 BLEU y además de Con esto, obtenemos otro impulso de hasta +1,0 BLEU gracias al enfoque de alimentación de insumos. Usando una mejor función de alineación, el producto escalar basado en contenido, junto con la deserción, produce otra ganancia de +2,7 BLEU. Por último, al aplicar la técnica de sustitución de palabras desconocidas, obtenemos +2,1 BLEU adicionales, lo que demuestra la utilidad de la atención a la hora de alinear palabras raras.

5 Análisis

Realizamos análisis exhaustivos para comprender mejor nuestros modelos en términos de aprendizaje, capacidad de ha Manejar oraciones largas, opciones de arquitecturas de atención y calidad de alineación. Todos los resultados reportados aquí están en newstest2014 en inglés y alemán.

5.1 Curvas de aprendizaje

Comparamos modelos construidos uno encima del otro como se enumera en la Tabla 1. Es agradable

Method	AER
global (location)	0.39
local-m (general)	0.34
local-p (general)	0.36
ensemble	0.34
Berkeley Aligner	0.32

Table 5: Alignment Error Rate (AER) for different methods.

observar en la Figura 5 una clara separación entre los modelos atencionales y no atencionales. El enfoque de alimentación de insumos y el modelo de atención local también demuestran su capacidad para reducir los costos de las pruebas. El modelo no atencional con abandono (la curva + azul) aprende más lentamente que otros modelos sin abandono, pero a medida que pasa el tiempo, se vuelve más robusto en términos de minimizar errores de prueba.

5.2 Efectos de traducir oraciones largas

Seguimos (Bahdanau et al., 2016) para agrupar oraciones de longitud similar y calcular una puntuación BLEU por grupo. La Figura 6 muestra que nuestros modelos atencionales son más efectivos que el no atencional en el manejo de oraciones largas: la calidad no se degrada a medida que las oraciones se vuelven más largas. Nuestro mejor modelo (el azul + curva) supera a todos los demás sistemas en cucharones de todas las longitudes.

5.3 Elecciones de arquitecturas atencionales

Examinamos diferentes modelos de atención (global, local-m, local-p) y diferentes funciones de alineación (ubicación, punto, general, concat) como se describe en la Sección 3. Debido a recursos limitados, no podemos ejecutar todas las combinaciones posibles. Sin embargo, los resultados del Cuadro 4 nos dan una idea sobre las diferentes opciones. La función basada en la ubicación no aprende buenas alineaciones: el modelo global (ubicación) solo puede obtener una pequeña ganancia cuando realiza el reemplazo de palabras desconocidas en comparación con el uso de otras funciones de alineación [14]. Para las funciones basadas en

[14]Existe una diferencia sutil en cómo recuperamos alineaciones para las diferentes funciones de alineación. En el paso de tiempo t en el que recibimos y_{t-1} como entrada y luego calculamos h_t, a_t, c_t y $\bar{h}_t$ antes de predecir y_t, el vector de alineación a_t se usa como pesos de alineación para (a) la palabra predicha yt en la ubicación funciones de alineación basadas en contenido y (b) la palabra de entrada y_{t-1} en las funciones basadas en contenido.

English-German translations	
src	Orlando Bloom and Miranda Kerr still love each other
ref	Orlando Bloom und Miranda Kerr lieben sich noch immer
best	Orlando Bloom und Miranda Kerr lieben einander noch immer .
base	Orlando Bloom und Lucas Miranda lieben einander noch immer .
src	”We’re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security”, said Roger Dow , CEO of the U.S. Travel Association.
ref	”Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht”, sagte Roger Dow , CEO der U.S. Travel Association .
best	”Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist ”, sagte Roger Dow , CEO der US - die .
base	”Wir freuen uns uber die $< unk >$, dass ein $< unk > < unk >$ mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit”, sagte Roger Cameron, CEO der US - $< unk >$.
German-English translations	
src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben.
ref	However , in an interview , Bloom has said that he and Kerr still love each other.
best	In an interview , however , Bloom said that he and Kerr still love.
base	However , in an interview , Bloom said that he and Tina were still $< unk >$.
src	Wegen der von Berlin und der Europaischen Zentralbank verh¨angten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Wahrung gen¨otigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure imposed by the European Central Bank and the Federal Central Bank with the strict austerity imposed on the national economy in the face of the single currency, many people believe that the European project has gone too far .

Table 6: Traducciones de muestra: para cada ejemplo, mostramos la fuente (src), la traducción humana (ref), la traducción de nuestro mejor modelo (mejor) y la traducción de un modelo no atencional (base). Ponemos en cursiva algunos segmentos de traducción correctos y resaltamos algunos incorrectos en negrita.

contenido, nuestra implementación *concat* no produce buenos rendimientos y más Se debe realizar un análisis para comprender la razón [15]. Es interesante observar que el punto funciona bien para la atención global y lo general es mejor para la atención local. Entre los diferentes modelos, el modelo de atención local con alineamientos predictivos (localp) es el mejor, tanto en términos de perplejidades como de BLEU.

[15]Con *concat*, las perplejidades logradas por diferentes modelos son 6.7 (global), 7.1 (local-m) y 7.1 (local-p). Tantas perplejidades podrían deberse a que simplificamos la matriz W_a para establecer la parte que corresponde a $\bar{h}_s$ como identidad

5.4 Calidad de alineación

Un subproducto de los modelos de atención son las alineaciones de palabras. Si bien (Bahdanau et al., 2016) visualizaron alineaciones para algunas oraciones de muestra y observaron mejoras en la calidad de la traducción como una indicación de un modelo de atención funcional, ningún trabajo ha evaluado las alineaciones aprendidas en su conjunto. Por el contrario, nos propusimos evaluar la calidad de la alineación utilizando la métrica de tasa de error de alineación (AER). Dados los datos de alineación de oro proporcionados por RWTH para 508 oraciones Europarl inglés-alemán, ”forzamos” a decodificar nuestros mod-

elos de atención para producir traducciones que coincidan con las referencias. Extraemos solo alineaciones uno a uno seleccionando la palabra de origen con el peso de alineación más alto por palabra de destino. Sin embargo, como se muestra en la Tabla 6, pudimos lograr puntuaciones AER comparables a las alineaciones uno a muchos obtenidas por el alineador Berkeley (Liang et al., 2006) [16]. También encontramos que las alineaciones producidas por los modelos de atención local lograr TAE inferiores a los del global. El AER obtenido por el conjunto, si bien es bueno, no es mejor que el AER local, lo que sugiere la conocida observación de que el AER y las puntuaciones de traducción no están bien correlacionados (Fraser and Marcu, 2007). Mostramos algunas visualizaciones de alineación en el Apéndice A.

5.5 Traducciones de muestra

En la Tabla ¡6 mostramos ejemplos de traducciones en ambas direcciones. Resulta atractivo observar el efecto de los modelos atencionales en la traducción correcta de nombres como "Miranda Kerr" y "Roger Dow". Los modelos no atencionales, si bien producen nombres sensatos desde la perspectiva del modelo de lenguaje, carecen de conexiones directas desde el lado fuente para realizar traducciones correctas. También observamos un caso interesante en el segundo ejemplo, que requiere traducir la frase doblemente negada "no incompatible". El modelo atencional produce correctamente "nicht. . . unvereinbar"; mientras que el modelo no atencional genera "nicht vereinbar", que significa "no compatible" [17]. El modelo atencional también demuestra su superioridad en la traducción de oraciones largas como en el último ejemplo.

6 Conclusión

En este artículo, proponemos dos mecanismos de atención simples y efectivos para la traducción automática neuronal: el enfoque global que siempre mira todas las posiciones de origen y el local que solo atiende a un subconjunto de posiciones de origen a la vez. Probamos la eficacia de nuestros modelos en las tareas de traducción WMT entre inglés y alemán en ambas direcciones. Nuestra atención local produce grandes ganancias de hasta 5,0 BLEU sobre los modelos no atencionales que ya incorporan técnicas conocidas como la deserción escolar. Para la dirección de traducción del inglés al alemán, nuestro modelo conjunto ha establecido nuevos resultados de vanguardia tanto para WMT'14 como para WMT'15, superando en rendimiento a los mejores sistemas existentes, respaldados por modelos NMT y reordenadores LM de n-gramas, por más superior a 1,0 BLEU. Hemos comparado varias funciones de alineación y hemos arrojado luz sobre qué funciones son mejores para cada modelo de atención. Nuestro análisis muestra que el modo NMT basado en la atención Los ls son superiores a los que no prestan atención en muchos casos, por ejemplo, al traducir nombres y manejar oraciones largas.

[16] Concatenamos los 508 pares de oraciones con 1 millón de pares de oraciones de WMT y ejecutamos el alineador Berkeley.

[17] La referencia utiliza una traducción más elegante de "incompatible", que es "im Widerspruch zu etwas stehen". Sin embargo, ambos modelos no lograron traducir la "experiencia del pasajero".

7 Reconocimiento

Agradecemos el apoyo de una donación de Bloomberg L.P. y el apoyo de NVIDIA Corporation con la donación de GPU Tesla K40. Agradecemos a Andrew Ng y su grupo, así como a Stanford Research Computing, por permitirnos utilizar sus recursos informáticos. Agradecemos a Russell Stewart por sus útiles debates sobre los modelos. Por último, agradecemos a Quoc Le, Ilya Sutskever, Oriol Vinyals, Richard Socher, Michael Kayser, Jiwei Li, Panupong Pasupat, Kelvin Guu, los miembros del Stanford NLP Group y los revisores anónimos por sus valiosos comentarios y opiniones.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3579–3584, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007. Squibs and discussions: Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126, Sofia, Bulgaria. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA. Association for Computational Linguistics.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2015. Recurrent neural network regularization.

Una visualización de alineación

Visualizamos los pesos de alineación producidos por nuestros diferentes modelos de atención en la Figura 7. La visualización del modelo de atención local es mucho más nítida que la del global. Este contraste coincide con nuestra expectativa de que la atención local esté diseñada para centrarse únicamente en un subconjunto de palabras cada vez. Además, dado que traducimos del inglés al alemán e invertimos la oración original en inglés, el blanco avanza hacia las palabras ”realidad” y ”.” en el modelo de atención global revela un patrón de acceso interesante: tiende a hacer referencia al comienzo de la secuencia fuente. En comparación con las visualizaciones de alineación en (Bahdanau et al., 2016), nuestros patrones de alineación no son tan nítidos como los de ellos. Tal diferencia podría deberse posiblemente al hecho de que traducir del inglés al alemán es más difícil que traducir al francés como se hace en (Bahdanau et al., 2016), lo cual es un punto interesante a examinar en trabajos futuros.

... traducido por Honorio Apaza

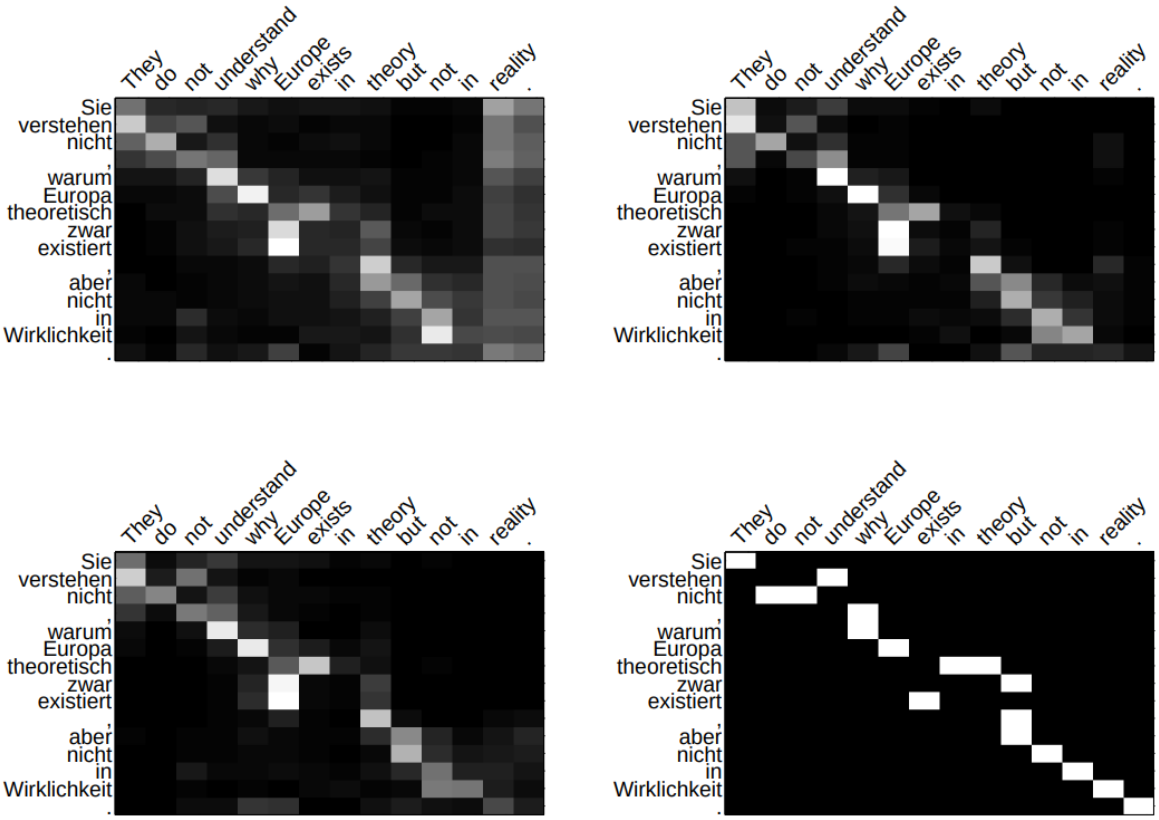

Figure 7: Visualizaciones de alineación: se muestran imágenes de los pesos de atención aprendidos por varios modelos: (arriba a la izquierda) global, (arriba a la derecha) local-m y (abajo a la izquierda) local-p. Las alineaciones doradas se muestran en la esquina inferior derecha..