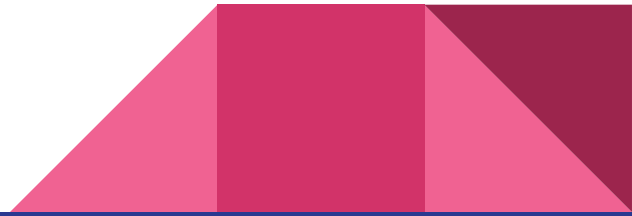
Learning to respond with DNN

for retrieval-based human-computer conversation system

IIC3633 - 2018.2
Henry Blair González

Sistemas de Conversación: Contexto histórico

- Sistemas orientados a tareas:
  - Inputs del sistema predecibles y restringidos
  - Dominios acotados
  - Outputs del sistema también muy restringidos, asociados al manejo de la tarea específica.



¿Cómo Ampliar el dominio del sistema?

Objetivo: Dominio Abierto, flexibilidad en inputs y outputs. **Conversar** con el compu.

- Deep Learning (DNN)
  - Enormes cantidades de datos (Foros, Redes sociales, Blogs, etc)
  - Enormes cantidades de computo

¿Cómo modelar las tareas? Consultas

- Tuplas (publicación, respuesta) -> pequeño documento
  - Muchas respuestas, muchas tuplas con la misma publicación.

- Dada una consulta(q0) se le somete a análisis de contenido (tf.idf)
  - Candidato(respuesta con la publicación asociada)

¿Cómo modelar las tareas? Consultas

- Reformulación contextual
  - revisar declaraciones previas a la consulta e incorporarlas en el análisis de contenido
  - analizar relevancia
    - Dado un contexto con N frases, hay 2^N formas de concatenarlas con la consulta original
    - Estrategias
      - Sin contexto
      - Todo el contexo
      - Add-One
      - Drop-Out
      - Todas las anteriores

¿Cómo modelar las tareas? Consultas

- Score y Rank usando DNN (LSTM)
  - Se captura semántica de la consulta, la tupla de respuesta candidata y el contexto
  - Dado el clasico match de consulta con documento, se hace de manera análoga
    - match consulta con respuesta: score de relación entre respuesta y consulta (Mayor = Mejor)
      - f: { (q, r) } -> R
    - match consulta con publicación: análogo al anterior
      - g:{ (q,p) } -> (0,1)
    - match consulta con contexto: establece qué tan correlacionada está una consulta concatenada con el contexto con la consulta original (Mayor = Mejor)
      - h{ (q_i,q_k) } -> (0,1)

Captura de semántica

- Score y Rank usando DNN (LSTM)
  - Word Embeddings: Las palabras se mapean a vectores de baja dimensión que estiman significado
  - LSTM Bidireccional: El siguiente proceso se hace en ambas direcciones.
    - Crea h_t de 4 salidas: i_t, f_t, o_t, l_t
    - sea x_t el word embedding de una palabra en la posición t de una frase S:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ l_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot \begin{bmatrix} h_{t-1} \\ e_t \end{bmatrix}$$

$$\tilde{h}_t = f_t \cdot \tilde{h}_{t-1} + i_t \cdot l_t$$

$$h_t^s = o_t \cdot \tilde{h}_t \qquad \sigma(\cdot) = \frac{1}{1+e^{-\cdot}}$$

el h_t mostrado con ~ es una variable auxiliar

Más capas de la DNN

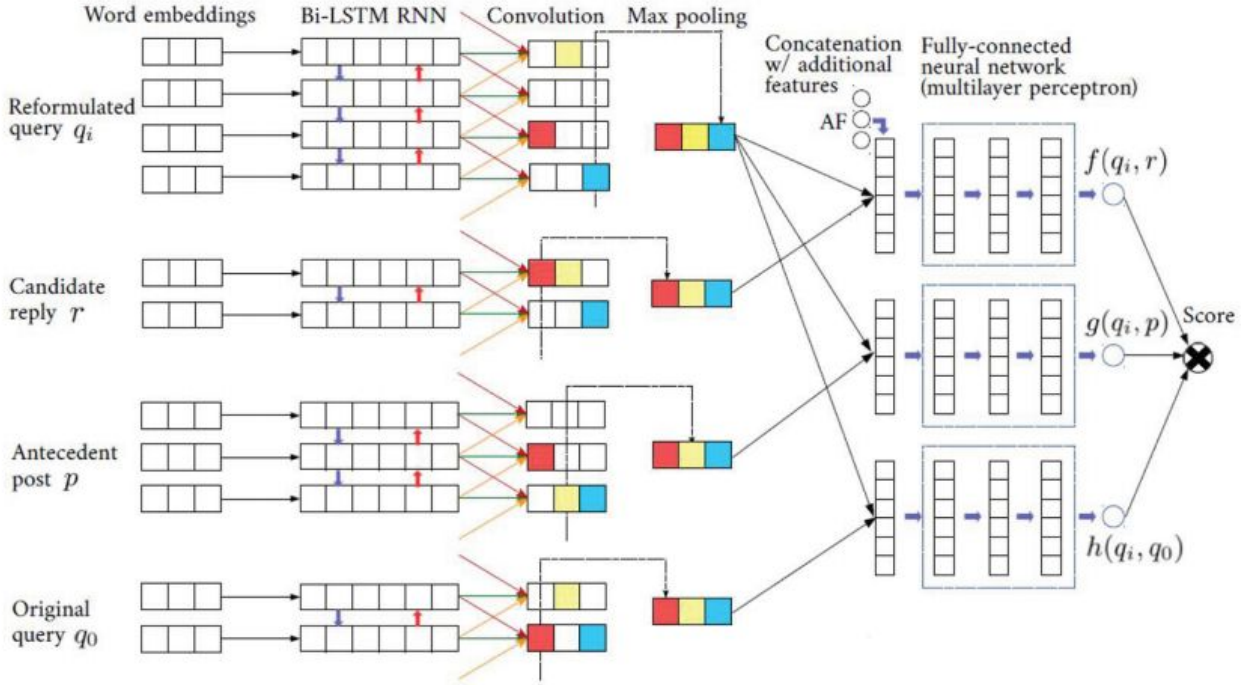
- Red neuronal convolucional
  - Para una ventana de m vectores h se convolucionan con un filtro de m componentes y se suma un parametro b(bias o sesgo)
  - Cada posición del vector obtenido corresponde a una palabra
  - Puede que haya más de un filtro para capturar más features
  - Los distintos filtros no comparten los parámetros de sesgo
- Red neuronal MLP
  - extrae más features del vector resultante de la CNN
- Cálculo de f,g, y h.

3 valores, mejor tener sólo 1

Se unifica en un sólo score con la siguiente función

$$\mathcal{F}(q_0, r) = \sum_{i=0}^{|\mathcal{Q}|} \left(h(q_0, q_i) \sum_{p} \left(f(q_i, r) \cdot g(q_i, p) \right) \right)$$

La DNN


Word embeddings
Bi-LSTM RNN
Convolution
Max pooling
Concatenation w/ additional features
AF
Fully-connected neural network (multilayer perceptron)
Reformulated query q_i
Candidate reply r
Antecedent post p
Original query q_0
$f(q_i, r)$
$g(q_i, p)$
$h(q_i, q_0)$
Score

Entrenamiento del modelo

Dada Una respuesta positiva y una negativa, con sus parámetros de ponderación y bias en el set Omega

$$\underset{\Omega}{\text{minimize}} \sum_{q_0, r^+} \max\left\{0, \Delta + \mathcal{F}(q_0, r^+) - \mathcal{F}(q_0, r^-)\right\} + \lambda \|\Omega\|_2^2$$

SetUp Experimental

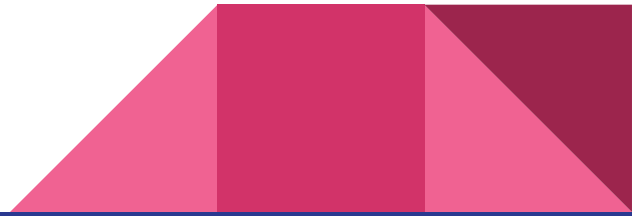
- Hiperparámetros
  - WordEmbeddings de 128 dimensiones inicializados de manera aleatoria
- DataSet
  - Idioma contemplado: Chino
  - Cerca de 180000 frases distintas con más de 2 ocurrencias
- Capas
  - LSTM Bidireccional de 128 hidden units
  - CNN de 256 componentes y una ventana de largo 3

SetUp Experimental

- Métricas de evaluación
  - p@1
  - MAP
  - nDCG
  - MRR
- Baselines de generación de respuestas
  - SMT
  - LSTM-RNN
  - NRM

SetUp Experimental

- Baselines de recuperación de respuestas
  - Random
  - Okapi BM25
  - DeepMatch
  - CNN



Resultados

Model	p@1	MAP	nDCG@5	nDCG@10	nDCG@20	MRR
SMT (Ritter et al., [26])	0.363					
LSTM-RNN (Sutskever et al., [32])	0.441					
NRM (Shang et al., [29])	0.465					
Random Match	0.266	0.246	0.247	0.289	0.353	0.083
Okapi BM25	0.272	0.253	0.337	0.302	0.368	0.169
DeepMatch (Lu and Li, [17])	0.457	0.317	0.419	0.454	0.508	0.275
LSTM-RNN (Palangi et al., [25])	0.338	0.283	0.330	0.371	0.431	0.228
ARC (Hu et al., [7])	0.394	0.294	0.397	0.421	0.477	0.232
DeepMatch w/ context adaption	0.603	0.378	0.555	0.584	0.628	**0.349**
LSTM-RNN w/ context adaption	0.362	0.296	0.354	0.395	0.453	0.237
ARC w/ context adaption	0.400	0.309	0.383	0.422	0.480	0.319
Deep Learning-to-Respond (DL2R)	**0.731***	**0.416***	**0.663***	**0.682***	**0.717***	0.333

Resultados

	p@1	MAP	nDCG@5	nDCG@10	nDCG@20	MRR
No Context	0.522	0.340	0.476	0.509	0.559	0.296
Whole Context	0.698	0.404	0.635	0.657	0.696	0.327
Add-One	0.716	0.411	0.650	0.670	0.706	0.322
Drop-Out	0.720	0.413	0.656	0.675	0.711	0.328
Combined	0.731	0.416	0.663	0.682	0.717	0.333

Resultados

	p@1	MAP	nDCG@5	nDCG@10	nDCG@20	MRR
Query-Reply w/o Query-Context	0.522	0.340	0.476	0.509	0.559	0.296
Query-Posting w/o Query-Context	0.510	0.302	0.404	0.425	0.489	0.285
Query-Reply w/ Query-Context	0.596	0.366	0.528	0.561	0.603	0.327
Query-Posting w/ Query-Context	0.563	0.362	0.483	0.516	0.568	0.316
Full Combination	0.731	0.416	0.663	0.682	0.717	0.333

Conclusión

A través de Deep Learning el sistema funciona a través de un pseudo sistema de recomendación de respuestas a sus preguntas y el modelo hace un ranking de las respuestas más pertinentes desde un análisis de contenido.

El uso de combinaciones de inclusión de contexto muestra que lo mejor es incluir todos los criterios

El uso de más medidas que correlacionen el contenido de la consulta con el contenido de la respuesta para incluir en el score final mejora el desempeño de manera significativa.

Gracias!