Sugerencia de Respuestas Cortas para Conversaciones de Chat

Daniela Bosch, Jonathan David Mutal

Noviembre 2017

Objetivo

¿Qué?

- Predecir `respuestas cortas` de acuerdo al contexto.

¿Para qué?

- Facilitar al usuario una respuesta `inmediata`.

Etapas a seguir

- Recolección de conversaciones de chats en español.
- Preprocesamiento de los chats
- Separar el contexto de las respuestas cortas para texto de entrenamiento.
- Evaluar con diferentes clasificadores.

Extracción de corpus

Problema:

- No hay corpus en la web.
- Muchos chats son privados (privacidad).

Extracción de corpus

Problema:

- No hay corpus en la web.
- Muchos chats son privados (privacidad).

Solución:

- No sólo recolectar chats privados si no grupos de diferentes IM (whatsapp, telegram, etc..).
- Mezclar con algún corpus parecido (sms, etc..).

Extracción de corpus

Problema:

- No hay corpus en la web.
- Muchos chats son privados (privacidad).

Solución:

- No sólo recolectar chats privados si no grupos de diferentes IM (whatsapp, telegram, etc..).
- Mezclar con algún corpus parecido (sms, etc..).

Alrededor de 10 personas nos han brindado sus chats...

IM	Tamaño	Vocab
Whatsapp	12MB	59,630
Facebook	15MB	non
Telegram	4MB	non

Table 1: Tamaño de los datos de entrenamiento

Datos de entrenamientos variados

Problemas:

- Mezcla de dominios (conversaciones grupales y personales).
- Palabras mal escritas.
- Palabras personalizadas.
- Chats con multimedia que forman parte de la conversación (imagenes, audios, videos, gif).
- Emoticones.

Datos de entrenamiento variados

Posibles soluciones:

- Realizar un pre-procesamiento.

Datos de entrenamiento variados

Posibles soluciones:

- Realizar un pre-procesamiento.
  - Eliminar STOP WORDS.
  - Normalizar algunas palabras. Ejemplo: sisi por si. okiis por ok.
  - Mapear cada emoticon con una palabra.

Datos de entrenamiento variados

Posibles soluciones:

- Realizar un pre-procesamiento.
  - Eliminar STOP WORDS.
  - Normalizar algunas palabras. Ejemplo: sisi por si. okiis por ok.
  - Mapear cada emoticon con una palabra.
- Ignorar multimedia (por ahora).
- Dividir el texto de entrenamiento en diferentes dominios.

Llego el gran problema...

¿Como usamos estos datos para hacer el bot?

Llego el gran problema...

¿Como usamos estos datos para hacer el bot?

Usar clasificadores (supervisado):

- Definir una entrada (el contexto).
- Definir las clases (respuestas cortas).

Pero.. ¿Qué es una respuesta corta?

- Definimos una respuesta corta como cualquier mensaje compuesto por N tokens.

Pero.. ¿Qué es una respuesta corta?

- Definimos una respuesta corta como cualquier mensaje compuesto por N tokens.

Y.. ¿Qué es el contexto de una respuesta corta?

Tenemos varias alternativas

- M turnos anteriores, es decir M mensajes anteriores con largo mayor igual a N.
- Turnos anteriores hasta encontrar una respuesta del mismo usuario.
- Turnos anteriores de acuerdo al tiempo de respuesta.

Ejemplo con 3 turnos anteriores y respuesta corta de 1

Chat

A: Hola B, como estas?
B: Bien y vos?
A: Super, hacemos algo hoy?
B: Dale

Conjunto de entrenamiento

Por lo que un dato de entrenamiento sería:
X: hola B, como estas? bien y vos? super, hacemos algo hoy?
Y: Dale

Caracterizando el contexto

Una vez definido el contexto y las clases. ¿Como caracterizamos el contexto?

Caracterizando el contexto

Una vez definido el contexto y las clases. ¿Como caracterizamos el contexto?

- Bolsa de palabras:
  - Con palabras. Problema con nuevos ejemplos. ¿Que pasa con una palabra nueva? ¿Que pasa si hay errores de ortografía?

Caracterizando el contexto

Una vez definido el contexto y las clases. ¿Como caracterizamos el contexto?

- Bolsa de palabras:
  - Con palabras. Problema con nuevos ejemplos. ¿Que pasa con una palabra nueva? ¿Que pasa si hay errores de ortografía?
  - Con N-gramas. Seguimos teniendo el mismo problema

Caracterizando el contexto

Una vez definido el contexto y las clases. ¿Como caracterizamos el contexto?

- Bolsa de palabras:
  - Con palabras. Problema con nuevos ejemplos. ¿Que pasa con una palabra nueva? ¿Que pasa si hay errores de ortografía?
  - Con N-gramas. Seguimos teniendo el mismo problema
  - Con subwords. Un poquito mejor.

Caracterizando el contexto

Una vez definido el contexto y las clases. ¿Como caracterizamos el contexto?

- Bolsa de palabras:
  - Con palabras. Problema con nuevos ejemplos. ¿Que pasa con una palabra nueva? ¿Que pasa si hay errores de ortografía?
  - Con N-gramas. Seguimos teniendo el mismo problema
  - Con subwords. Un poquito mejor.
- Word embeddings de diferentes dominios:
  - SBWCE
  - Corpus del chat
  - Con corpus de twitter

Clases

Problema

Gran cantidad de clases (respuestas cortas). Imposible predecir con tan poco corpus.

Solución

Reducir a 3 clases. ¿Como hacerlo?

- Clustering.
- Respuestas más vistas en el corpus.
- A ojo.

Clasificadores

Algunos clasificadores a utilizar:

- SVM
- Logistic Regression
- Decision Trees
- Naïve Bayes

Otras soluciones - Trabajo futuro

- Definir previamente las respuestas cortas y etiquetar el corpus.
- Predecir varias respuestas cortas
- Usar un modelo de lenguaje.
- Usar como contexto todos los turnos anteriores pesados por proximidad
- Hacer modelos neuronales seq2seq.

Colaboración

Nosotros no pedimos monedas... solo sus chats íntimos.

Agradecemos cualquier colaboración

{jonathanmutal95, danielarbosch}@gmail.com

Gracias por su atención!