Database-driven Chatbot

Eduardo Neves
Universidade de Coimbra
email@example.com

Abstract

With the rising implementation of chatbots in today's technologies (!!!) their implementations has risen in various eneterprises. One example is the inclusion of many chatbots in websites. With many applications

1 Introdução

A Chatbot is

2 Dados e Abordagem

The data for a typical Chatbot is based on real or made-up conversations between two parts. In reality, "any given tet with utility" can be used to build the algorithm. Para este trabalho foi usado um 'dataset' baseado em em legendas de filmes para português.

2.1 Conjunto de dados

O pacote de dados usado foi tirado do projeto "OPUS ... the open parallel corpus", uma coleção de testos traduzidos da internet baseada em produtos 'open source'. Este conjunto estava inserido na ala 'Open Subtitles', retirada opensubtitles.org(!!!!!).

Como o foco do projeto é um algoritmo em portguês, retirou-se o dataset "pt", que coleciona mais de 500000 legendas de filmes até o ano de 2017. Separados por anos, alguns filmes contam com várias versões do mesmo filme. O seu tratamento é explicado na secção seguinte.

3 Experimentação e Metodologia

3.1 Tratamento dos dados

Com o conjunto de dados inicial com redundâncias e pouco estruturado, procurou-se correr um pequena organização à informação. A organização do fluxo de entrada para treino é feita aquando a inicialização do processo.

Redução de redundâncias

Como referido, algumas versões do mesmo filme são dispostas e organizadas em conjunto. Ao remover este entrave, pode-se remover uma camada na diretoria e agrupar apenas por ano. Apesar desta diferenciação não ser necessária poderá ser útil em comparações ortográficas ou até numa maior confiança de traduções mais recentes. Para selecionar o ficheiro mais relevante apenas se retirou o ficheiro com menor volume de espaço no disco para 'performance'.

Alteração do formato

O *dataset* original continha apenas ficheiros em .xml, com muitas "especificações!!!!!" desnecessárias a este trabalho. Um exemplo é a separação por palavras das falas de cada personagem. Simplificou-se então para um formato .json, onde cada entrada é a fala de uma personagem.

Redução de amostragem

Para além da redução na alteração do formato, reduziu-se também a quantidade de dados usados para o programa. Utilizou-se apenas filmes a partir de 2000, o que resultou em 37851 filmes, com uma média de 455 falas de personagens.

3.2 Modelo

lstm

4 Conclusions