shagunsodhani / **Evaluating Prerequisite Qualities for Learning End-to-end Dialog Systems.md**

Created 6 years ago

Summary of "Evaluating Prerequisite Qualities for Learning End-to-end Dialog Systems" paper

`Evaluating Prerequisite Qualities for Learning End-to-end Dialog Systems.md`

Evaluating Prerequisite Qualities for Learning End-to-end Dialog Systems

Introduction

- The paper presents a suite of benchmark tasks to evaluate end-to-end dialogue systems such that performing well on the tasks is a necessary (but not sufficient) condition for a fully functional dialogue agent.
- Link to the paper

Dataset

- Created using large-scale real-world sources - OMDB (Open Movie Database), MovieLens and Reddit.
- Consists of ~75K movie entities and ~3.5M training examples.

Tasks

QA Task

- Answering Factoid Questions without relation to the previous dialogue.
- KB(Knowledge Base) created using OMDB and stored as triplets of the form (Entity, Relation, Entity).

- Question (in Natural Language Form) generated by creating templates using SimpleQuestions
- Instead of giving out just 1 response, the system ranks all the answers in order of their relevance.

Recommendation Task

- Providing personalised responses to the user via recommendation instead of providing universal facts as in case 1.
- MovieLens dataset with a *user x item* matrix of ratings.
- Statements (for any user) are generated by sampling highly ranked movies by the user and forming a statement about these movies using natural language templates.
- Like the previous case, a list of ranked responses is generated.

QA + Recommendation Task

- Maintaining short dialogues involving both factoid and personalised content.
- Dataset consists of short conversations of 3 exchanges (3 from each participant).

Reddit Discussion Task

- Identify most likely response is discussions on Reddit.
- Data processed to flatten the potential conversation so that it appears to be a two participant conversation.

Joint Task

- Combines all the previous tasks into one single task to test all the skills at once.

Models Tested

- **Memory Networks** - Comprises of a memory component that includes both long term memory and short term context.
- **Supervised Embedding Models** - Sum the word embeddings of the input and the target independently and compare them with a similarity metric.
- **Recurrent Language Models** - RNN, LSTM, SeqToSeq
- **Question Answering Systems** - Systems that answer natural language questions by converting them into search queries over a KB.

- **SVD(Singular Value Decomposition)** - Standard benchmark for recommendation.
- **Information Retrieval Models** - Given a message, find the most similar message in the training dataset and report its output or find a most similar response to input directly.

Result

QA Task

- QA System > Memory Networks > Supervised Embeddings > LSTM

Recommendation Task

- Supervised Embeddings > Memory Networks > LSTM > SVD

Task Involving Dialog History

- QA + Recommendation Task and Reddit Discussion Task
- Memory Networks > Supervised Embeddings > LSTM

Joint Task

- Supervised word embeddings perform very poorly even when using a large number of dimensions (2000 dimensions).
- Memory Networks perform better than embedding models as they can utilise the local context and the long-term memory. But they do not perform as well on standalone QA tasks.