Corpus-based analysis on cross-domain experiments in classification-and-ranking generation

ABSTRACT

Problem statement: Overgeneration-and-ranking architecture works well in written language where sentence is the basic unit. However, in spoken language where utterance is the basic unit, the disadvantage becomes critical as spoken language also render intentions, hence short strings may be of equivalent impact. Approach: In classification-and-ranking, response was deliberately chosen from dialogue corpus rather than wholly generated, such that it allows short ungrammatical utterances as long as they satisfy the intended meaning of input utterance. Because the architecture is intention-based, it adopted an open-domain knowledge representation, whereby response utterances were semantically represented using some ontology general enough for future reuse in another domain. Results: This study presented corpus-based analysis on cross-domain experimentation using different type of corpus to validate the consistency of the response classifier that delimits the searching space for ranking. The open-domain quality for classification-an-ranking architecture was tested on two mixed-initiative, transaction dialogue corpus in theater reservation and emergency planning. Results showed consistent distribution accuracies in both classification and ranking experiment, indicating that the approach is viable for cross-domain implementations. Conclusion: The ability of a response generation system to directly learn response utterances from the domain corpus suggested the possibility to build a dialogue system by feeding the learning module with a target corpus and the system learned the response behavior directly from the training corpus.

Keyword: Corpus-based; Open-domain; Natural language generation; Dialogue systems