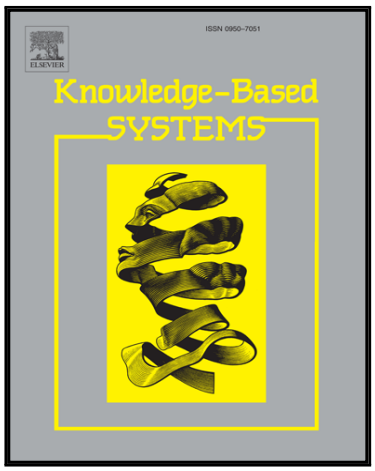

Response Selection from Unstructured Documents for Human-Computer Conversation Systems

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li

PII: S0950-7051(17)30564-6
DOI: 10.1016/j.knosys.2017.11.033
Reference: KNOSYS 4128

To appear in: *Knowledge-Based Systems*

Received date: 19 June 2017
Revised date: 23 November 2017
Accepted date: 25 November 2017

Please cite this article as: Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, Response Selection from Unstructured Documents for Human-Computer Conversation Systems, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.11.033

Response Selection from Unstructured Documents for Human-Computer Conversation Systems

Zhao Yan[a], Nan Duan[b], Junwei Bao[c], Peng Chen[d], Ming Zhou[b], Zhoujun Li[a,*]

[a]*Beihang University, Beijing, China*
[b]*Microsoft Research Asia, Beijing, China*
[c]*Harbin Institute of Technology, Harbin, China*
[d]*Microsoft Xiaoice Team, Beijing, China*

Abstract

This paper studies response selection for human-computer conversation systems. Existing retrieval-based human-computer conversation systems are intended to reply to user utterances based on existing utterance-response pairs. However, collecting sufficient utterance-response pairs is intractable in practical situations, especially for many specific domains. We introduce DocChat a novel information retrieval approach for human-computer conversation systems that can use unstructured documents rather than semi-structured utterance-response pairs, to react to user utterances. The key of DocChat is a learning to rank model with features designed at various levels of granularity which is proposed to quantify the relevance between utterances and responses directly. We conduct comprehensive experiments on both sentence selection and real human-computer conversation scenarios. Empirical studies of sentence selection datasets shows reasonable improvements and the strong adaptability of our model. We compare DocChat with Xiaoice, a famous open domain chitchat engine in China. Side-by-side evaluation shows that DocChat is a good complement for human-computer conversation systems using utterance-response pairs as the primary source of responses. Furthermore, we release a large scale open-domain dataset

*Corresponding author
Email addresses: yanzhao@buaa.edu.cn (Zhao Yan), nanduan@microsoft.com (Nan Duan), baojunwei001@gmail.com (Junwei Bao), peche@microsoft.com (Peng Chen), mingzhou@microsoft.com (Ming Zhou), lizj@buaa.edu.cn (Zhoujun Li)