A Robust Natural Language Question Answering System for Customer Helpdesk Applications

Lahiru Thilina Samarakoon

(108010U)

Thesis submitted in partial fulfilment of the requirements for the degree Master of Philosophy

Department of Electrical Engineering

University of Moratuwa
Sri Lanka

June 2012

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree of Diploma in any other University of Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the MPhil thesis under my supervision.

Signature of the supervisor: Date:

Signature of the supervisor: Date:

ABSTRACT

This thesis describes a restricted-domain question answering system which can be used in automating a customer helpdesk of a commercial organization. Even though there has been an increasing interest in data-driven methods over the past decade to achieve more natural human-machine interactions, such methods require a large amount of manually labeled representative data on how user converses with a machine. However, this is a requirement that is difficult to be satisfied in the early phase of system development. In addition, the systems should be maintainable by a domain expert who is less technically skilled when compared to a computer engineer. The knowledge based approach that is presented here is aimed at maximally making use of the user experience available with the customer service representatives (CSRs) in the organization and presents how true representative data can be collected. The approach takes into account the syntactic, lexical, and morphological variations, as well as a way of synonym transduction that is allowed to vary over the system's knowledge base. The query understanding method, which is based on a statistical classifier, a ranking algorithm based on Vector Space Model (VSM) and a pattern writing process, takes into account the intent, context, and content components of natural language meaning as well as the word order. A genetic algorithm-based method is presented for finding the domain specific ranking parameters. An evaluation of the approach is presented by deploying a system in a real-world enterprise helpdesk environment in the telecommunication domain. The evaluation shows that the system is able to answer user questions with an accuracy of 94.4%. Furthermore, maintenance of the deployed system is carried out by CSRs successfully.

Keywords:
Question Answering, Automated Customer Helpdesk, Vector Space Model.

ACKNOWLEDGEMENTS

I would like to thank Exetel Pty Ltd, Australia for sponsoring the project. Special thank goes to the Sri Lanka Institute of Information Technology (SLIIT) for providing the facilities of their research center. I also thank the University of Moratuwa for their support.

Special thanks to my supervisor, Professor Sisil Kumarawadu, for his constant encouragement and support, and the rest of my committee, Dr. Ranga Rodrigo, Dr. Asanka Rodrigo, Dr Malitha Wijesundara, Dr. Koliya Pulasinghe and Dr. Harsha Abeykone, for their assistance to guide this work.

Most importantly, I would like to thank my family for their love and support of this endeavor.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES