中国科学技术大学

学士学位论文

中文知识关系抽取的研究与实现

姓　　名：	曾锃煜
院　　系：	计算机科学与技术系
学　　号：	PB13203234
导　　师：	陈欢欢 教授
完成时间：	二〇一七年六月

University of Science and Technology of China

A dissertation for bachelor's degree

Research and Realization of Chinese Knowledge Relation Extraction

Author :	Zengyu Zeng
Department :	No.11 Department
Student ID :	PB13203234
Supervisor :	Prof. Huanhuan Chen
Finished Time :	June, 2017

致　谢

在中国科技大学完成本科学业的四年里，我所从事的学习和研究工作，都是在导师以及系里其他老师和同学的指导和帮助下进行的。在完成论文之际，请容许我对他们表达诚挚的谢意。

感谢班主任王海龙老师多年的关怀。感谢蒋凡等老师，他们本科及研究生阶段的指导给我研究生阶段的研究工作打下了基础。

感谢张练钢等师兄师姐们的指点和照顾；感谢李卓华等几位同班同学，与你们的讨论使我受益良多；感谢王译锋等师弟师妹，我们在实验室共同学习共同生活，一起走过了这段愉快而难忘的岁月。

感谢科大，感谢一路走过来的兄弟姐妹们，在最宝贵年华里，是你们伴随着我的成长。

最后，感谢我家人一贯的鼓励和支持，你们是我追求学业的坚强后盾。

曾锃煜

2017 年 4 月 10 日

目 录

表格索引

插图索引

算法索引

摘 要

互联网不断发展，其中的信息也随着时间日渐增多，传统的返回检索方式开始无法满足获取所需信息和知识资源的全面性和以高效率完成。实体的知识关系抽取，可以从自然语言（中文文本）中抽取实体，并将实体之间的关系结构化，提高了用户可获取信息的全面性和获取的效率。

信息提取（IE）系统寻求从自然语言中提取语义关系文本，但大多数系统使用监督学习关系特定的例子，因此受到训练数据可用性的限制。开放式信息提取系统例如 TextRunner，在另一方面，致力于处理没有限制数量的从互联网获取的实体关系。

传统上，信息提取专注于精确、狭义的、预先指定的要求。例如从一些会议通告里提取时间和地点。而转移到另一个领域里，则需要用户对实体关系命名并手工制定新的提取规则或对新的训练集例子进行手工标注。这样的人力工作量随着目标实体关系的数量线性增加。

开放式关系抽取（Open Relation Extraction，ORE）是实体关系抽取的一种，它克服了传统信息提取（IE）的缺陷，即传统的信息获取技术对每种关系模式各自训练了他们的提取器。

有很多系统流行于英文的 ORE，例如 OLLIE，ReVerb 和 Exemplar 等。然而，对于其他语言的 ORE 则基本没有相关研究的报告。本毕业设计采用了基于语法分析的系统 ZORE（Zh ORE）来对简体中文文本进行关系和语义模式的抽取。ZORE 从自动解析的依赖树里定义了候选的关系，然后将实体的关系和语义模式不断地通过一种新的双重传播算法。

本文内容包括了对于所采取的实体关系抽取系统（ZORE）的介绍及其实现，以及关于 ZORE 所需组件的介绍，并将其应用在实际工程中。

关键词： 开放式关系抽取 ZORE 双重传播算法

ABSTRACT

The continuous development of the Internet, where the information is also increasing with the time, the traditional return search method began to meet the need to obtain the required information and knowledge resources, comprehensive and efficient to complete. The knowledge of the entity can be extracted from the natural language (Chinese text), and the relationship between the entities is structured, which improves the comprehensiveness and efficiency of the information available to the user.

The information extraction (IE) system seeks to extract semantic relation text from natural language, but most systems use specific examples of supervised learning relationships, which are therefore limited by the availability of training data. Open information extraction systems such as TextRunner, on the other hand, are committed to dealing with unrestricted quantities of physical relationships obtained from the Internet. Traditionally, information extraction focuses on precise, narrow, pre-specified requirements. Such as extracting time and place from some of the meeting announcements. And moved to another domain, the user needs to name the entity relationship and manually create new extraction rules or manually annotate the new training set examples. This amount of human work increases linearly with the number of target entities.

Open Relation Extraction (ORE) is a kind of entity relationship extraction, which overcomes the shortcomings of traditional information extraction (IE), that is, the traditional information acquisition technology for each relationship model of their training of their extractors.

There are many systems popular in English ORE, such as OLLIE, ReVerb and Exemplar. However, ORE for other languages is basically no relevant research report. The graduation design uses a system based on grammar analysis ZORE (Zh ORE) to the simplified Chinese text and semantic model of the relationship between the extraction. Zore defines the candidate relationship from the dependent tree of the automatic resolution, and then continually passes the entity's relationship and semantic pattern through a new dual-propagation algorithm.

This article covers the introduction and implementation of the Entity Relationship Extraction System (ZORE), as well as the introduction of the components required for ZORE, and applies it to the actual project.

Keywords: Open relation extratction, ZORE, Double propagation algorithm

参考文献

[1] 翻译小组 C. lshort 中文版 3.20. 2003.
[2] Qiu L, Zhang Y. ZORE: A Syntax-based System for Chinese Open Relation Extraction. EMNLP, 2014, 104(5):8913–8921.
[3] Qiu L, Zhang Y. ZORE: A Syntax-based System for Chinese Open Relation Extraction. EMNLP, 2014..
[4] 王元卓，贾岩涛，等. 基于开放网络知识的信息检索与数据挖掘. 计算机研究与发展, 2014, 52(2):456–474.