中国科学技术大学

学士学位论文

中国科学技术大学

中文知识关系抽取的研究与实现

作者姓名： 曾锃煜

学科专业：计算机科学与技术

导师姓名： 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's Name: Zengyu Zeng

Speciality: Computer Science and Tech.

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

图目录

表目录

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

With the continuous development of the Internet, with the passage of time, more and more information, the traditional return search method began to meet the need to obtain the required information and knowledge resources needs, fully and effectively completed. The knowledge of the entity can be extracted from the natural language (Chinese text), the structure of the relationship between the entities, and improve the user's available information is comprehensive and efficient.

The information extraction (IE) system seeks to extract semantic relations text from natural language, but most systems use specific examples of supervised learning relationships, thus limiting the availability of training data. On the other hand, an open information extraction system such as TextRunner is dedicated to handling unrestricted physical relationships obtained from the Internet.

Traditionally, information extraction focuses on precise, narrow and pre-defined requirements. Such as extracting time and place from certain meeting notifications. And move to another domain, the user needs to name the entity relationship and manually create a new extraction rule or manually annotate the new training set example. The human workload increases linearly with the number of target entities.

Open relational extraction (ORE) is an entity relationship extraction that overcomes the shortcomings of traditional information extraction (IE), the traditional information acquisition techniques for each relational model to develop their extractors.

English ORE has many popular systems, such as OLLIE, ReVerb and Exemplar. However, ORE in other languages is basically no relevant research report. The graduation design uses a system based on parsing. ZORE (Zh ORE) simplifies the relationship between Chinese text and semantic models. Zore defines the candidate relationship from the automatic resolution dependency tree, and then passes the entity's relationship and semantic pattern continually through the new dual-propagation algorithm.

This article introduces the introduction and implementation of the Entity Relationship Extraction System (ZORE), and introduces the components required by ZORE and applies it to the actual project.

Key Words: Open relation extratction, ZORE, Double propagation algorithm

第 1 章　简介

1.1　模板简介

测试脚注[1]。

1.1.1　模板介绍 1

1.1.1.1　模板测试

1.1.2　模板介绍 2

1.2　系统要求

1.3　问题反馈

测试脚注[2]

[1]分别编号

[2]脚注 2

第 2 章　图片

本章展示图片相关用法。

2.1　示例

图 2.1　测试图片

2.2　带图注的图

图 2.2　带图注的图片

注：the solid lines represent the time histogram of the spontaneous activities of an old monkey cell(gray) and a young monkey cell (black). The bin-width is 1

第 3 章　表格

3.1　A Simple Table

表 3.1　这里是表的标题

a	b
c	d

注：这里是表的注释

3.2　长表格

表 3.2　长表格演示

名称	说明	备注
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC

续下页

表 3.2　长表格演示（续）

名称	说明	备注
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC
AAAAAAAAAAAA	BBBBBBBBBBB	CCCCCCCCCCCCCC

第 4 章　数学

4.1　定理、引理和证明

定义 4.1　If the integral of function f is measurable and non-negative, we define its (extended) **Lebesgue integral** by

$$\int f = \sup_g \int g, \tag{4.1}$$

where the supremum is taken over all measurable functions g such that $0 \leq g \leq f$, and where g is bounded and supported on a set of finite measure.

例 4.1　Simple examples of functions on $\mathbb{R}^d$ that are integrable (or non-integrable) are given by

$$f_a(x) = \begin{cases} |x|^{-a} & \text{if } |x| \leq 1, \\ 0 & \text{if } x > 1. \end{cases} \tag{4.2}$$

$$F_a(x) = \frac{1}{1+|x|^a}, \qquad \text{all } x \in \mathbb{R}^d. \tag{4.3}$$

Then f_a is integrable exactly when $a < d$, while F_a is integrable exactly when $a > d$.

引理 4.1 (Fatou)　Suppose $\{f_n\}$ is a sequence of measurable functions with $f_n \geq 0$. If $\lim_{n\to\infty} f_n(x) = f(x)$ for a.e. x, then

$$\int f \leq \liminf_{n\to\infty} \int f_n. \tag{4.4}$$

注　We do not exclude the cases $\int f = \infty$, or $\liminf_{n\to\infty} f_n = \infty$.

推论 4.2　Suppose f is a non-negative measurable function, and $\{f_n\}$ a sequence of non-negative measurable functions with $f_n(x) \leq f(x)$ and $f_n(x) \to f(x)$ for almost every x. Then

$$\lim_{n\to\infty} \int f_n = \int f. \tag{4.5}$$

命题 4.3　Suppose f is integrable on $\mathbb{R}^d$. Then for every $\epsilon > 0$:

i. There exists a set of finite measure B (a ball, for example) such that

$$\int_{B^c} |f| < \epsilon. \tag{4.6}$$

ii. There is a $\delta > 0$ such that

$$\int_E |f| < \epsilon \qquad \text{whenever } m(E) < \delta. \tag{4.7}$$

定理 4.4 Suppose $\{f_n\}$ is a sequence of measurable functions such that $f_n(x) \to f(x)$ a.e. x, as n tends to infinity. If $|f_n(x)| \leq g(x)$, where g is integrable, then

$$\int |f_n - f| \to 0 \qquad \text{as } n \to \infty, \tag{4.8}$$

and consequently

$$\int f_n \to \int f \qquad \text{as } n \to \infty. \tag{4.9}$$

证明 Trivial. □

4.2 自定义

Axiom of choice Suppose E is a set and E_α is a collection of non-empty subsets of E. Then there is a function $\alpha \mapsto x_\alpha$ (a "choice function") such that

$$x_\alpha \in E_\alpha, \qquad \text{for all } \alpha. \tag{4.10}$$

Observation 1 Suppose a partially ordered set P has the property that every chain has an upper bound in P. Then the set P contains at least one maximal element.

A concise proof Obvious. □

第 5 章　算法环境

模板中使用 `algorithm2e` 宏包实现算法环境。关于该宏包的具体用法，请阅读宏包的官方文档。

Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while *not at end of this document* **do**
    read current;
    if *understand* **then**
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end

算法 5.1: 算法示例 1

```
input : A bitmap Im of size w × l
output: A partition of the bitmap
special treatment of the first line;
for i ← 2 to l do
    special treatment of the first element of line i;
    for j ← 2 to w do
        left ← FindCompress(Im[i, j − 1]);
        up ← FindCompress(Im[i − 1, ]);
        this ← FindCompress(Im[i, j]);
        if left compatible with this then // O(left, this)==1
            if left < this then Union(left,this);
            else Union(this,left);
        end
        if up compatible with this then          // O(up, this)==1
            if up < this then Union(up,this);
            // this is put under up to keep tree as
               flat as possible
            else Union(this,up);
            // this linked to up
        end
    end
    foreach element e of the line i do FindCompress(p);
end
```

算法 5.2: 算法示例 2

第 6 章　代码环境

模板中使用 listings 宏包实现代码环境。详细用法见宏包的官方说明文档。

以下是代码示例，可以在文中任意位置引用代码 6.1 。

代码 6.1　示例代码

```
1 #include <stdio.h>
2
3 int main( )
4 {
5    printf("hello, world\n");
6    return 0;
7 }
```

第 7 章　交叉引用

图 2.1 位于第 10 页，其标题为测试图片。

表 3.2 位于第 12 页，其标题为长表格演示。

代码 6.1 位于第 18 页，其标题为示例代码。

算法 5.1 位于第 16 页，其标题为算法环境。

第 8 章　引用文献标注

8.1　著者-出版年制标注法

```
\citestyle{ustcauthoryear}
```

`\cite{ZORE}`	⇒	Qiu et al. (2014)
`\citet{knuth86a}`	⇒	Knuth (1986)
`\citet[chap.~2]{knuth86a}`	⇒	Knuth (1986, chap. 2)
`\citep{knuth86a}`	⇒	(Knuth, 1986)
`\citep[chap.~2]{knuth86a}`	⇒	(Knuth, 1986, chap. 2)
`\citep[see][]{knuth86a}`	⇒	(see Knuth, 1986)
`\citep[see][chap.~2]{knuth86a}`	⇒	(see Knuth, 1986, chap. 2)
`\citet*{knuth86a}`	⇒	Knuth (1986)
`\citep*{knuth86a}`	⇒	(Knuth, 1986)
`\citet{knuth86a,tlc2}`	⇒	Knuth (1986); Mittelbach et al. (2004)
`\citep{knuth86a,tlc2}`	⇒	(Knuth, 1986; Mittelbach et al., 2004)
`\cite{knuth86a,knuth84}`	⇒	Knuth (1984, 1986)
`\citet{knuth86a,knuth84}`	⇒	Knuth (1984, 1986)
`\citep{knuth86a,knuth84}`	⇒	(Knuth, 1984, 1986)

8.2　顺序编码制标注法

```
\citestyle{ustcnumerical}
```

`\cite{knuth86a}` ⇒ [2]

`\citet{knuth86a}` ⇒ Knuth[2]

`\citet[chap.~2]{knuth86a}` ⇒ Knuth[2, chap. 2]

`\citep{knuth86a}` ⇒ [2]

`\citep[chap.~2]{knuth86a}` ⇒ [2] chap. 2

`\citep[see][]{knuth86a}` ⇒ see[2]

`\citep[see][chap.~2]{knuth86a}` ⇒ see[2] chap. 2

`\citet*{knuth86a}` ⇒ Knuth[2]

`\citep*{knuth86a}` ⇒ [2]

`\citet{knuth86a,tlc2}` ⇒ Knuth[2], Mittelbach et al.[3]

`\citep{knuth86a,tlc2}` ⇒ [2,3]

`\cite{knuth86a,knuth84}` ⇒ [1,2]

`\citet{knuth86a,knuth84}` ⇒ Knuth[1,2]

`\citep{knuth86a,knuth84}` ⇒ [1,2]

`\cite{knuth86a,knuth84,tlc2}` ⇒ [1–3]

8.3 其他形式的标注

`\citealt{tlc2}` ⇒ Mittelbach et al. 3

`\citealt*{tlc2}` ⇒ Mittelbach, Goossens, Braams, and Carlisle 3

`\citealp{tlc2}` ⇒ 3

`\citealp*{tlc2}` ⇒ 3

`\citealp{tlc2,knuth86a}` ⇒ 2,3

`\citealp[pg.~32]{tlc2}` ⇒ 3 pg. 32

`\citenum{tlc2}` ⇒ 3

`\citetext{priv.\ comm.}` ⇒ [priv. comm.]

`\citeauthor{tlc2}` ⇒ Mittelbach et al.

`\citeauthor*{tlc2}` ⇒ Mittelbach, Goossens, Braams, and Carlisle

`\citeyear{tlc2}` ⇒ 2004

`\citeyearpar{tlc2}` ⇒ 2004

参考文献

Knuth D E. May 1984. Literate programming[J]. *The Computer Journal*. 27(2):97–111.

Knuth D E. 1986. Computers and Typesetting: A The TEXbook[M]. Reading, MA, USA: Addison-Wesley.

Mittelbach F, Goossens M, Braams J, et al. 2004. The LATEX Companion[M]. 2nd ed. Reading, MA, USA: Addison-Wesley.

Qiu L, Zhang Y. 2014. Zore: A syntax-based system for chinese open relation extraction[J]. *EMNLP*.

附录 A　论文规范