Holographic Lexical Chain and Its Application in Chinese Text Summarization

Shengluan Hou[1,2(✉)], Yu Huang[1,2], Chaoqun Fei[1,2], Shuhan Zhang[1,2], and Ruqian Lu[1,3]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
houshengluanl989@163.com
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Key Lab of MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Abstract. Lexical chain has been widely used in many NLP areas. However, when using it for Web text summarization, especially for domain-specific text summarization, we got low accuracy results. The main reason is that traditional lexical chains only take nouns into consideration while information of other grammatical parts is missing. We introduce lexical chains of predicates and adjectives (adverbs) respectively. These three types of lexical chains together are called holographic lexical chains (HLCs), which capture most of the information included in the text. A specifically designed construction method for HLC is presented. We applied HLC method to Chinese text summarization and used machine learning methods whose features are adapted to the new method. In a comparative study of Chinese foreign trade texts, we got summarization results with accuracy of 86.88%. Our HLC construction method obtained improvements of 7.02% in accuracy than the known best methods in Chinese text summarization.

Keywords: Holographic lexical chain · Text summarization · Machine learning · Lexical cohesion

1 Introduction

Lexical cohesion is a classical tool for analyzing the content of natural language text. Lexical chain is a list of "about the same thing" words, which exploit the lexical cohesion among related words and contributes the continuity of text meaning [13, 19]. Lexical chain has been widely used in many natural language applications, such as text summarization [1–4, 10], machine translation [23], discourse quality measurement [20] and some other areas [5, 17, 22].

The influx of large amount of web documents in the web age results in a great deal of attention on automatic text summarization [14, 15]. Lexical chain is a good tool for text summarization because of its easy computation and high efficiency. There are many efforts of lexical chain based text summarization [2, 4, 10, 19]. Generally, these lexical chain based text summarization methods can be decomposed into two

L. Chen et al. (Eds.): APWeb-WAIM 2017, Part I, LNCS 10366, pp. 266–281, 2017.
DOI: 10.1007/978-3-319-63579-8_21

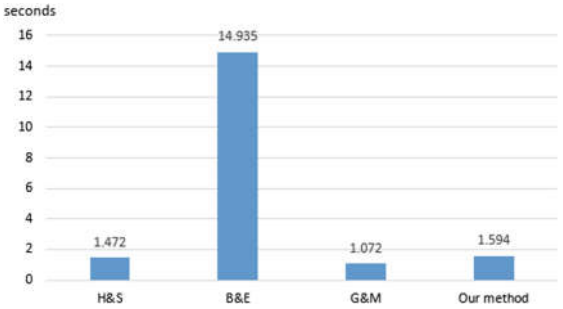

Fig. 6. Time cost of different methods

5 Conclusion

In this paper, we presented our approach of holographic lexical chains (HLCs), which contain three kinds of lexical chains: noun lexical chains, predicate lexical chains and adjective (adverb) lexical chains. HLCs capture most of the information included in the text. A specifically designed HLC construction method and scoring criterion are also presented in this paper. We applied HLC technique to Chinese text summarization. In the summary sentence selection step, we used two machine learning methods: linear regression and support vector machine whose features are adapted to HLC and other text features. Comparative experiments on Chinese foreign trade text summarization demonstrate that our holographic lexical chain construction method outperforms other methods in Chinese text summarization.

One of the future works will be further improving efficiency and accuracy of the HLC technique. A possible way is to integrate HLC with some other features, such as those of discourse structure. Another important task is further applying HLC to other NLP applications.

Acknowledgement. This work was supported by National Key Research and Development Program of China under grant 2016YFB1000902, National Natural Science Foundation of China (No. 61232015, 61472412, 61621003), Beijing Science and Technology Project: Machine Learning based Stomatology and Tsinghua-Tencent-AMSS Joint Project: WWW Knowledge Structure and its Application.

References

1. Alam, H., Kumar, A., Nakamura, M., et al.: Structured and unstructured document summarization: design of a commercial summarizer using lexical chains. In: ICDAR, vol. 3, pp. 1147 (2003)
2. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. Adv. Autom. Text Summar. 111–121 (1999)
3. Brügmann, S., Bouayad-Agha, N., Burga, A., et al.: Towards content-oriented patent document processing: intelligent patent analysis and summarization. World Patent Inf. **40**, 30–42 (2015)

文章编号：1003-0077（2017）00-0000-00

文本摘要常用数据集和方法研究综述

侯圣峦[1,2] 张书涵[1,2] 费超群[1,2]

（1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院大学，北京 100049）

摘要：文本摘要成为人们从互联网上海量文本信息中便捷获取知识的重要手段。现有方法都是在特定数据集上进行训练和效果评价，数据集包括一些公用数据集和作者自建数据集。已有综述文献对现有方法进行全面细致的总结，但大都是对方法进行总结，而缺少对数据集的详细描述。从调研数据集的角度出发，对文本摘要常用数据集及在该数据集上的经典和最新方法进行综述。对公用数据集的综述包括数据来源、语言及获取方式等，对自建数据集的总结包括数据规模、获取和标注方式等。对于每一种公用数据集，给出了文本摘要问题的形式化定义。同时，对经典和最新方法在特定数据集上的实验效果进行了分析。最后，总结了已有常用数据集和方法的现状，并指出一些存在的问题。

关键词：文本摘要；自然语言处理；机器学习；人工智能

中图分类号：TP391 **文献标识码：**A

Commonly Used Datasets and Methods of Text Summarization: A Survey

HOU Shengluan[1,2], ZHANG Shuhan[1,2], and FEI Chaoqun[1,2]

(1. Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract： Text Summarization has become an essential way of knowledge acquisition from mass text documents on the Internet. Existing methods were trained and evaluated on specific datasets, which are commonly used datasets or manually constructed datasets. The survey literatures on text summarization mostly summarize methods, which lack of reviews on the experimental datasets. This work commenced the survey of evaluation datasets, and then summarized the commonly used datasets and manually constructed datasets, companion with their corresponding approaches. The review on commonly used datasets includes their sources, language and the way of access, etc. The data scale, annotation methods are included in the review on manually constructed datasets. For each commonly used dataset, the formal definition of text summarization problem is given. Meanwhile, we analyzed the experimental results of classical and latest text summarization methods on one specific dataset. We concluded the present situation of existing datasets and methods at last, from which we pointed out some problems.

Key words: Text Summarization; Natural Language Processing; Machine Learning; Artificial Intelligence

0 引言

文本摘要任务旨在从一篇或多篇相同主题的文本中抽取能够反映主题的精简压缩版本[1-2]，可

收稿日期：；**定稿日期**：
基金项目：国家重点研发计划（2016YFB1000902）；国家自然科学基金(61232015, 61472412, 61621003)

录　用　通　知

侯圣峦同志:

您好！您于 2018 年 6 月 28 日交来的稿件，题名为**文本摘要常用数据集和方法研究综述**（稿件编号：2018-0071），经过专家们的评审后，已被录用。由于每期刊登的篇幅数有限，故还不能将具体刊登的期号通知您，望能得到您的谅解。

特此通知。

《中文信息学报》编辑部

2018 年 10 月 15 日

文章编号：1003-0077（2017）00-0000-00

面向中文的修辞结构关系分类体系及无歧义标注方法

侯圣峦[1,2] 费超群[1,2] 张书涵[1,2]

（1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院大学，北京 100190）

摘要：修辞结构理论是一种重要的篇章结构理论，其核心是修辞结构关系。基于修辞结构理论，结合中文文本特点，提出面向中文的层次化修辞结构关系分类体系及多元定义。同时，针对标注者遇到的歧义问题，提出了无歧义标注方法。设计并实现了基于 Java 图形界面的标注工具 RSTTagger，以句子的主谓结构关键词构成的元组作为基本标注单位，自底向上逐级标注最终标注成一颗完整的篇章结构树。选取 160 篇中文外贸领域语料进行标注，为验证标注结果的一致性，不同标注者同时标注其中 50 篇，标注一致性达到 76.63%。该标注框架可以应用到其他领域语料标注中，已标注语料可以作为篇章结构理论研究的基础语料库。

关键词：自然语言处理；修辞结构理论；修辞结构关系；篇章结构分析

中图分类号：TP391 **文献标识码：**A

Chinese-oriented Rhetorical Structure Relation Taxonomy and Unambiguous Annotation Method

HOU Shengluan[1,2], FEI Chaoqun[1,2], and ZHANG Shuhan[1,2]

(1. Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Rhetorical Structure Theory(RST) is one of the most prevalent discourse structure theories. Core of RST is the Rhetorical Structure Relation(RSR), which holds between textual units of various sizes. Based on English-oriented RST and the characteristics of Chinese natural language text, hierarchical taxonomy and multiple definitions of Chinese-oriented RSR were presented in this paper. Moreover, an unambiguous annotated method was proposed to deal with the problem of ambiguity. A Java-GUI based tagging tool called RSTTagger was designed and implemented, which is a bottom-up tagger, whose elementary tagging unit is a subject-predicate structure and tagging result is a full discourse structure tree. To validate our proposed tagging framework, we selected 160 Chinese foreign trade text as the tagging corpus, from which 50 texts were randomly selected to be tagged by different annotators. We got annotator agreement with score 76.63%. This framework can be extended to other domain text annotation. The 160 finished corpus can also be the research corpus of discourse parsing.

Key words: Natural Language Processing; Rhetorical Structure Theory; Rhetorical Structure Relation; Discourse Parsing

收稿日期：；定稿日期：

基金项目：国家重点研发计划（2016YFB1000902）；国家自然科学基金(61232015, 61472412, 61621003)

录　用　通　知

侯圣峦同志:

您好！您于 2018 年 7 月 16 日交来的稿件，题名为**面向中文的修辞结构关系分类体系及无歧义标注方法**（稿件编号：2018-0082），经过专家们的评审后，已被录用。由于每期刊登的篇幅数有限，故还不能将具体刊登的期号通知您，望能得到您的谅解。

特此通知。

《中文信息学报》编辑部

2018 年 9 月 17 日