情报工程 ISSN 2095-915X
第2卷 第1期 43-52 2016年2月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.1 43-52 Feb 2016

doi:10.3772/j.issn.2095-915x.2016.01.006

中文语义角色标注在情报分析领域的应用研究

孟令恩[1]，何彦青[2]，李颖[2]

（1 北京智课网科技有限公司；2 中国科学技术信息研究所 北京 100038）

摘要：语义角色标注（Semantic Role Labeling，简称SRL）作为自然语言处理的基础研究，其理论发展比较成熟，而应用层面的探究相对滞后。本文着眼于SRL在情报分析领域的应用，归纳了过去15年间的相关研究，分析了应用中的问题。最后，提出了中文SRL未来可应用的场景。

关键词：SRL，情报分析，自动问答，信息抽取，信息分类，情感分析，特征工程

Application Study of the Chinese SRL in Intelligence Analysis

MENG LingEn[1], HE YanQing[2], LI Ying[2]

(1. Smart Study Technology Co.Ltd, Beijing 2. Institute of Scientific and Technical Information of China, Beijing, 100038)

Abstract: As a basic research in natural language processing, the theoretical development of SRL is relatively mature, while its application research is lagging behind. This paper focuses on the application of SRL in information analysis;summarizes the relevant researches of SRL application for the past 15 years; and, analyzes its problems. Finally, we present the application scenarios of the Chinese SRL in the future.

Keywords: Semantic role labeling, intelligence analysis, automatic question answering, information extraction, information classification, sentiment analysis, feature engineering

基金支持：本文为国家自然科学基金项目“面向专利文本的统计机器翻译语境分析 (ID：61303052)“研究成果的一部分。
作者简介：孟令恩，硕士。研究方向：自然语言处理，数据挖掘，联系方式：menglingen@163.com；何彦青，博士。主要研究方向为机器翻译、自然语言处理，联系方式：heyq@istic.ac.cn；李颖，博士，本文通讯作者。主要从事科技信息资源系统、数字图书馆系统、数字出版系统，以及情报工程理论与应用等研究。联系方式：liying@istic.ac.cn。

引言

大数据和云计算给情报分析带来了更多的机遇与挑战。情报分析起步于检索，发展到以抽取和挖掘为主流研究方向。当前的情报分析领域，非结构化数据的比例越来越大，数据处理要求的精细度越来越高，对智能化方法和工具的依赖越来越强，分析结果的可视化需求也越来越迫切。在此背景之下，基于大数据理论与方法的数据分析、数据挖掘、数据监管、舆情监测等成为情报学研究的热点和重点[1]。自然语言处理作为人工智能的重要分支，在情报学研究，特别是情报分析中的作用越来越重要。其中，自然语言处理的语义学研究，尤其是语义角色标注SRL成为重要的课题，将其应用于情报分析是理论结合实践的体现，可充分诠释自然语言处理的深远意义。

本研究搜集整理了2001−2015年15年间中文SRL在情报分析中应用的相关论文，进行了详实的总结归纳、分析问题，提出了可能的应用前景。

本文内容结构如下：引言之后，第1部分对SRL进行了分类总结，主要从语料库建设、SRL模型的有效特征、SRL概率模型三方面展开描述；第2部分详细归纳了SRL在情报分析中的应用；第3部分分析了SRL在情报分析中的应用现状和当前的问题；第4部分提出了SRL在情报分析中可能的应用场景；最后为结束语。

1 SRL概述

SRL源于语言学家对句子的各个成分(也称论元，argument)与谓词(predicate)之间语义关系的研究[2]，信息工程学和计算语言学领域对SRL的英文定义分别如下：

SRL detecting basic event structures such as who did what to whom, when and where.

SRL identify the arguments of a given verb and assign them semantic labels describing the roles they play in the predicate (i.e., identify predicate argument structures).

由以上定义可知，SRL的概念内涵是：以句子作为原始的分析对象，以句子中的谓词为核心，分析句子中其他成分与谓词之间的关系。

目前，SRL以有监督学习为主，且实用效果较好。为此，本文将SRL概述大致分为三方面进行归纳：语料、特征工程和分类模型的选择。而CNN，RNN，DNN等深度学习自选特征的研究不在本探究范围内。

1.1 SRL语料

常用的相关英文语料有：框架网FrameNet[3]，英文命题库PropBank[4]和英文名词命题库NomBank三种。中文SRL语料库，主要有中文命题库(Chinese PropBank)和中文名词命题库(Chinese Nombank)，两者的标注规范分别对应英文命题库和英文名词命题库；山西大学构建的Chinese FrameNet是基于框架语义理论，参照了FrameNet的形式；台湾中研院陈凤仪构建了中文句结构树资料库(Sinica Treebank)；北京大学袁毓林教授组织建设了中文网库[5]。另外还有特定领域的语料建设成果，比如沈阳航空航天大学专利领域的语料等，本研究团队亦曾人工建设了小规模的科技领域的树库。

在SRL语料构建中，人工标注的过程极为繁琐，需要在对每一句话进行深刻理解的基础上，依次找到语句中所有角色的边界，保证不重、不漏、精细。最后，按照所定的规则，逐一赋予不同角色的正确标签，是高成本的任务。正因为如此，目前的SRL语料库规模都不够大。

1.2 SRL特征工程

特征工程是有监督学习的重要研究课题，称其为最关键环节也不为过。特征工程直接影响最

后的标注效果，从 CoNLL04(the Conference on Natural Language Learning 2004) 有关 SRL 的评测开始，学者在特征工程方面的探究热度持续不减，包括验证各种有效特征，进行有效特征的组合，引入辅助特征等。表 1 列举了基于短语结构句法的常用有效特征。

表 1 SRL 常用有效特征举例

特征说明	特征说明
当前子树离谓词间隔多少子树	当前路径中心词
当前子树短语类型	当前父结点的中心词
短语在谓词前 / 后出现次数	当前路径的中心词的词性
左邻兄弟短语	路径父结点中心词的词性
左邻兄弟中心词词性	当前路径在谓词前还是在谓词后
右邻兄弟短语	到谓词和当前子树成分的总路径
右邻兄弟短语中心词	谓词和当前子树成分的总路径的左路径
右邻兄弟短语中心词的词性	当前谓词的推出规则
父亲短语类型	当前路径成分的第一个单词
当前谓词	当前路径成分的最后一个单词

表 1 中的一些组合关系以及一些特殊类型的辅助特征，如：双谓词、多谓词句型等特征证实为有效稳定的特征，在此不再一一列举，可参照相关文献 [6-10]。应用时有必要区分基于短语句法分析树和基于依存句法分析树的有效特征。

1.3 SRL 概率模型

概率模型在 SRL 中充当精准分类的作用，依据不同机理形成的分类器多种多样，效果不同。目前，大多数 SRL 工作集中使用有监督的学习方法，最常用的有基于最大熵分类器的 SRL[11]，基于支持向量机的 SRL[12] 等，Surdeanu 等 [13] 人使用的决策树模型与之相比，效果明显不佳，甚至相差了 9 个百分点。

在 SRL 最近一次国际测评 CoNLL2009 shared task 的英文开放测试中，车万翔基于依存关系的 SRL 系统以 83.44% 的精确度和 77.07% 的召回率获得最高分。此后，从英文移植到中文，并经过改进之后，白龙飞 [14] 基于依存关系的中文 SRL 效果达到了 85.82%。在随后几年的探究之后，中文 SRL 的效果被证实为完全可以作为语义分析的一种工具，从而可应用于情报分析中，通过利用其语义特点使情报分析更智能化。

2 SRL 在情报分析中的应用

有关 SRL 的基础理论和应用研究，国内在 2008、2009 年两年之间达到了研究高峰，这两年产文高的主要原因是 CoNLL08 与 CoNLL09 两次评测会议，基于依存句法的 SRL 使得 SRL 性能提高到了较为理想的状态，其后多年 SRL 依然是研究热点。但 SRL 发展至今，其基础理论方面的研究文献远多于其应用文献，各大数据库检索结果显示，不论发文数量还是引用数量，基础理论的数量都远远多于应用。表 2 给出了 15 年间中文 SRL 应用领域的细分结果：

表 2 中文 SRL 在情报分析中应用细分表

SRL 应用领域	细分领域	代表论文
自动问答	问句分析	谭伟 [15]，吕德新 [16]
	问句检索	谭伟 [15]，安强强 [17]
	答案抽取	安强强 [17]，刘宁峰 [18]
情感分析	语义倾向判断	张岩 [19]，戴霖 [20]，施寒潇 [21]
	评价对象抽取	鞠久鹏 [22]，王荣洋 [23]，李景玉 [24]
信息抽取	事件抽取	于江德 [25]，吴刚 [26][27]
	实体关系识别	张奇 [28]，江超男 [29]
	命名实体识别	江超男 [29]
	主题提取	孟令恩 [30]，张帆 [31]
信息分类	文本分类	薛伟 [32]
	言语意图分类	王可 [33]
语句相似度计算	句相似度	张祎挺 [34]

表 2 中的应用领域均为自然语言处理的重要研究方向，也体现了大数据情报分析的重要应用场景，值得关注。下节具体介绍每一应用领域。

2.1 SRL 在自动问答系统中的应用

自动问答是利用计算机系统对用户提问问题进行处理，并在相关资源中找到合理答案的过程。问答系统大致由问句分析、问句检索、答案抽取三步构成。SRL 的应用常见于问句分析。

2.1.1 问句分析与问句检索

国内最早的应用见于谭伟的论文，该研究将输入的问句进行语义角色标注，之后通过标注好的语义角色生成自定义语义框架，最后利用语义框架测量查询语句与候选语句之间的语义相关性，从而确定最终答案。在当时较好地解决了基于关键词匹配的问答系统的缺陷。其中，语义框架由语义角色组成，主要有 ARG0–ARG5、ARGM–LOC、ARGM–TMP 和 ARGM–ADV*。图 1 中的涂色框代表了 SRL 的应用。

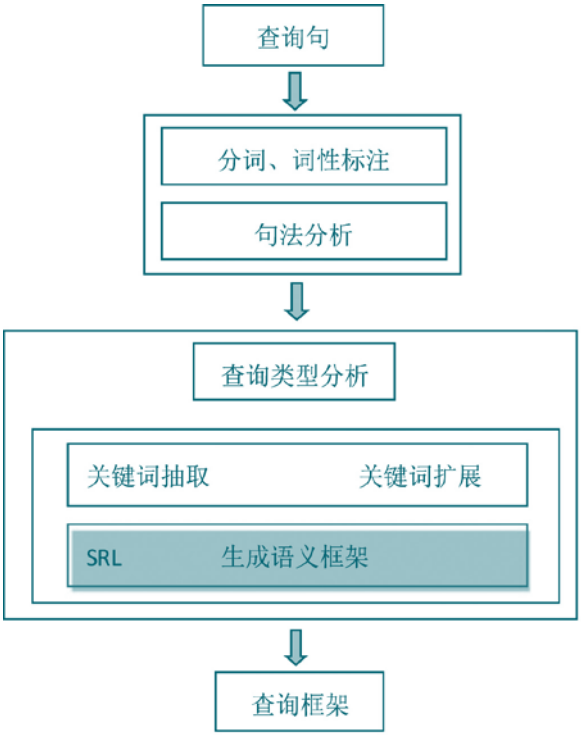

图 1 查询句分析流程图

* 语义角色的含义如下：核心论元 ARGN(N 为 0~5) 和附加论元 ARGM–X。ARG0 通常表示动作的施事 (及物动词的主语)，ARG1 通常表示动作的影响 (及物动词的直接宾语) 等，ARG2~5 根据谓语动词不同会有不同的语义含义。ARGM–LOC(地点)、ARGM–TMP(时间) 和 ARGM–ADV(副词)。

疑问句问点的查找。文章根据实际需要，最终选取了 28 个语义角色作为标记，采用了最大熵模型，选取了三方面的特征：有关事件特征，问点特征，事件与问点的关系特征，合计 9 种单特征，以及这些单特征的组合特征，系统在开放测试的环境下取得了问点 SRL92.3% 的正确率。

2.1.2 答案提取

安强强引入 SRL，指出中文 SRL 效果不如英文，为此进行了改进。首先，通过学习语义搭配关系，将其用于 SRL 的后处理，提升了 SRL 近 3 个百分点；之后，基于 SRL 进行了答案抽取，答案抽取过程类似于语义角色比对，示例参见图 2。刘宁峰将 SRL 应用于答案提取阶段，使用了简化的 SRL，更接近 SAO(Subject-Action-Object) 结构，通过 SAO 结构的对比，分析了句子的相似度，并基于此选择正确的答案。张晓孪也将 SRL 引入问答系统，但论文没有论及具体的应用方法。

问句：谁提出了进化论？
[ARGO 谁][TARGET 提出] 了 [ARGI 进化论]?
候选答句 1：法国生物学家拉马克在其《动物哲学》一书中，最早提出进化论。
[ARGO 法国生物学家拉马克][ARGM-LOC 在其《动物哲学》书中]，[ARGM—ADV 最早][TARGET 提出][ARGI 进化论]。
候选答句 2：科学家提出达尔文进化论 6 大热点区域。
[ARGO 科学家][TARGET 提出][ARGI 达尔文进化论 6 大热点区域]。
问句的 TARGET 为“提出”，ARGI 为“进化论”。
候选答句 I 的 TARGET 为“提出”，ARGI 为“进化论”，匹配成功。
候选答句 2 的 TARGET 为“提出”，ARGI 的中心词为“区域”，匹配失败。
候选答案 I 匹配成功，问句疑问词的语义角色为 ARGO，候选答案 1 的 ARGO 为“法国生物学家拉马克”，从而可得答案为“法国生物学家拉马克”。

图 2 SRL 在问答系统中应用示例

SRL 在问答系统中的应用较为广泛，因为其标注的论元本身隐含着答案，也可以作为直接答案来回答问题。问句分析与答案抽取都可以直接用到其中的信息，也可以利用其间接生成语义框架来帮助系统的问答。另外，SRL 也可以作为问答系统的辅助功能，从某一方面来检测问答的精确度等。但前提是 SRL 的精度要有保证。

2.2 SRL 在情感分析中的应用

情感分析是利用技术手段了解用户对产品的评价、网民对某一事件的看法、了解百姓对某商品、政策等的态度等。本文将其分为语义倾向判断和评价对象抽取等任务。

2.2.1 语义倾向判断

语义倾向是一个词与其本义的偏离距离，目前此距离的衡量主要用两个指标：偏离方向 (正负) 与偏离强度 (正负程度)。语义倾向性判断常用的方法有：基于词共现的统计检验方法，基于机器学习算法进行分类，另有不太常用的模糊集方法等，详见参考文献 [19-21]。

张岩将语义角色引入句子来判断句子倾向，这是倾向性判断的创新手法。该研究先将句子进行语义角色标注，之后提取谓词周围角色的语义信息作为句子结构信息，这些信息将会在将句子转化成向量的过程中作为重要的语义信息加入，整个向量转化过程中巧妙利用了语义信息。

戴霖通过对语义角色标注后的角色分析来判别句子的倾向，如：[ARG0 中国民众][Predicate 有][ARG1 五味杂陈的困惑]，通过提取其中具有感情色彩的角色，判断中国民众对此新闻的倾向。

施寒潇在细粒度情感分析研究中，通过使用SRL方法抽取语义特征进行了对象属性的评价，主要将谓词以及ARG0、ARG1和ARGM-ADV角色带有的语义信息作为特征加入模型的训练，经过十折交叉验证，加入语义角色特征的系统精度能达到最高值。

2.2.2 评价对象抽取

SRL在情感分析的另一基础性工作，即评价对象抽取中应用比较广泛。

鞠久鹏利用SRL提取了语义角色特征，并加入到评价对象抽取的任务中，将SRL的谓词分成be动词和实义动词，将形容词分为表语形容词和形容词修饰名词两种。如果句子中含有情感词，就用语义角色标记加情感词词性作为特征，否则只用语义角色标记做特征。

王荣洋借鉴了SRL的思想，提出了基于浅层语义分析框架的评价对象抽取方法，将句中的情感表述视为语义角色抽取，将评价对象问题转换成了SRL问题。

李景玉利用SRL信息作为特征应用于中文微博候选评价对象的筛选中。主要的特征提取自ARG0、ARG1和谓词。

与在问答系统中的应用相比较，将SRL应用到情感分析中难度有所增加。简单地说，可以直接将标注好的语义角色提取出来进行应用，但是更多借鉴了SRL的方法来重新标注需要标注的“情感词”。它们分别属于内容的直接使用和方法的间接借鉴。

2.3 SRL在信息抽取中的应用

信息抽取是指从文本中抽取出特定事件等有用信息。本研究将信息抽取细分成多个子任务：事件抽取、共指关系、关系抽取、命名实体识别和主题抽取等。

2.3.1 事件抽取

于江德将条件随机场模型用于SRL，之后将其用于事件抽取的语句SRL。大致分为两部分：第一阶段针对SRL的谓词，通过其进行时间的触发，第二阶段是从事件语句中抽取事件的要素，即进行SRL。

吴刚基于SRL理论提出了基于论元结构层次的事件抽取模式。文章将时间模板看作是句子语义的抽象，模板中的槽是事件的参与者，这样就将事件元素与SRL中的论元进行了对应，是一种新视角的信息抽取方法。

刘封基于PA结构(Predicate-Argument Structure)的语义信息抽取，在引入PA结构后，语义信息抽取就被分成了两步：由句法结构得到PA结构和将PA结构映射到语义结构。

何中市等人基于SRL方法提出了一种中文事件识别方法。主要研究工作是围绕语义角色进行了特征的构建，提出了与语义角色相关的四种概念定义。在上述基础上提取到特征，再利用条件随机场概率模型进行事件抽取实验。从实验结果可知，加入语义角色特征的系统性能提高了近4个百分点。

谢宝陵等人将SRL引入到军事战书的关键信息抽取，用SRL的方法做了军标信息(例如：兵种、武器、级别等)的抽取，方法采用J48决策树的角色分类模型，通过分析语句中被标注为触发词(武器项、兵种等)的内容，与注释表进行匹配。

2.3.2 关系抽取

张奇在未定义类型的实体关系抽取中引入了SRL工具。并分为两个阶段处理：第一阶段是SRL中谓词的识别，第二阶段利用FrameNet的语义框架进行了谓语与实体关系的映射。Berkeley大学开发的FrameNet以语义框架为基础，正是这

种特性使得第二阶段任务可实施。

廉营基于 SRL 对中文微博的人物关系进行了抽取研究。主要在 SRL 的基础上构造了有效的特征集，如：人名实体的语句成分特征 (如：人名同是主语的子句设定为并列关系，将所有的实体关系用数字进行表示) 等一系列的特征，用 CRF 进行了关系的分类实验。

2.3.3 命名实体识别

江超男将 SRL 思想引入了命名实体的识别任务中，把人名、机构名映射成为语义角色中的某种角色，通过维特比算法对句子进行了语义角色的自动标注，针对标注好的句子用规则匹配的方式抽取出了相应的角色作为识别出的实体名字，并在此基础上抽取了事件和关系的三元组。

2.3.4 主题抽取

孟令恩等人将 SRL 技术引入到专利文献中，将专利文献的长句简化为短句，语义角色标注短句，通过归纳总结语义框架，对应抽取到了专利摘要中的主题和所属领域。此方法对英文进行了实验，但该方法泛化力强，中文可用相同的流程进行操作。

张帆等人借助领域本体 , 在对句子进行 SRL 的基础上 , 采用一种语义标注、依存句法分析以及领域本体属性类相结合的方法 , 对创新点句内部的核心主题词以及主题词对应属性实例进行识别。其后主要工作为 : 探讨如何进一步利用领域知识提高属性识别的效果并对现有的属性实例进行整合 , 以及将领域中分散的创新点句通过其属性汇聚起来的方法。

2.4 SRL 在信息分类中的应用

信息分类是计算机对大量的文本文档按照既定的标准 (比如：主题内容) 实现大量信息的自动归类。本文按照应用范围分为：文本分类和言语分类。

(1) 文本分类

薛伟研究认为 SRL 有助于改善文本分类性能，文章借助了 FrameNet 中对谓词有详尽标注的便利条件，方便地提取了既定的语义角色及其对应含义。由于 FrameNet 词语收集有限，文章也利用其他语料库进行了扩展工作。

(2) 言语分类

王可在言语意图分类中引入了 SRL 进行句子的结构分析。其方法是把言语意图句分析成一些更细微的结构，然后再组合起来。其分析方法是将范畴语法和框架语义结合起来，分别对待句子的句法层和语义层，较好的解决了语法 – 语义不能分层的问题，句法、语义接口问题一直是语义理论家的争论焦点。

2.5 SRL 用于语句相似度计算

张祎挺利用 SRL 计算句子相似度，通过挖掘语义角色框架中蕴含的语义信息，将其应用到句子相似度的计算上，实验结果证实了语义角色框架可提高句子相似性的识别。同时，还存在着一些不足，比如在 SRL 部分，没有处理含有两个以上主动词的句子的情况，在计算词语相似度时，将知网中不存在的词的相似度计算做了简化处理。

在相似度计算方面，SRL 应用比较广泛。例如在自动问答系统中，很多部分都需要用到句子相似度的计算，自动问答系统的问题库中需要计算问题与问题之间的相似度，在答案抽取模块中需要计算问题和候选文本中句子的相似度，在答案抽取模块中的多文档自动文摘也要用句子相似度来对句子进行聚类；在基于实例的机器翻译中利用句子相似原理进行源语言检索；在信息过滤技术中利用句子模糊匹配来过滤敏感信息等。

3 SRL 应用现状及其问题

纵观中文 SRL 的 15 年应用研究可以看出，SRL 在情报分析中的应用文章十分有限，而近两年 SRL 的研究热度已然减退。事实上，SRL 的应用并没有像大部分文章中所述那样的“应用方面很广”、“得到了大量应用”。SRL 在情报分析中应用最多的是信息抽取。

SRL 作为自然语言处理的重要手法，在语义分析中起到了承接句法分析，启发深层语义分析的作用。但是，基于上文的归纳可以说， SRL 还并没有实现真正意义上的推广，原因是作为一种语义分析工具应用至情报分析过程中还是诸多问题，具体如下：

(1)SRL 对短句稳定有效，在长句中的表现有待提高

对于句子短、结构简单的语句，SRL 可以正确地进行标注，但一旦句子变长，因为受到句法分析能力的限制，当句法分析不准确时，从句法树中获得的特征会形成“错误累积”，这种不准确会传播到语义角色标注过程中，从而影响语义角色标注效果。本文作者在专利长句的信息抽取实验中，通过对一部分带有从句的长句进行了语义角色标注的实验后发现：对于大部分的专利长句，语义角色标注并不能完全正确的识别其中的角色，其识别的角色要么是带有从句的“长”角色，要么是找不到正确的边界。将 SRL 应用到专利信息抽取中时，一旦有专利长句，准确率就会大大下降，专利术语多的现象又会导致句子结构复杂，这也充分说明了提高长句性能的必要性和迫切性。

(2)SRL 的指代消解问题

如果仅仅利用 SRL 的技术而不搭配指代消解技术，那么有些语句即便角色标注正确，也可能不会产生实际效果。比如：在问答系统的答案匹配过程中，用户输入“谁发明了留声机？”的问题，那么如果知识库中有这样的一段话：“誉满全球的发明大王－－托马斯·阿尔瓦·爱迪生一生发明无数，他发明了留声机等无数杰作。”如果进行语义角色标注时，“他发明了留声机等无数杰作”这句话会被标注为：[ARG0 他] [Predicate 发明] 了 [ARG1 留声机等无数杰作]。在进行答案提取时，没有指代消解技术让系统了解：[ARG0 他] = 托马斯·阿尔瓦·爱迪生这一事实的话，就不可能提取正确答案。可以说指代消解和 SRL 不可分割。很多应用都需要指代消解技术的支持。

(3) 语义角色定义的不完备与不统一

SRL 的语料库不止一种导致了语义角色标注的标签不一致，还有手工标注失误等原因。就 FrameNet 而言，其语义框架覆盖有限，导致在大数据量上应用时常出现找不到对应语义框架的问题。旨在实施应用的研究者不得不引入辅助的语义工具，比如 WordNet、同义词词林等进行语义的映射和扩充。

(4) 语料库的单一性

现在具有规模的 SRL 语料库大都建立在新闻经济领域之上，如果用此领域训练好的模型去标注与之行文风格相差较大的领域文本，会出现准确率下降问题。比如：本研究采用新闻语料训练而成的模型，对科技文献进行语义角色标注时，如果不采取针对科技文献特征的相应处理，测试发现标注正确率下降明显。

(5) 情报学者的自然语言处理技术应用能力不足

情报研究者不仅要能使用自然语言处理工具，还要深入理解这些技术如何解决具体的应用问题，并非易事。

4 SRL 应用展望

4.1 SRL 理论上待解决问题

(1) 提高长语句标注正确率

如果长句的 SRL 的效果不好，可将长句通过一系列手段拆分成短句。作者做过长句拆分短句的实验，以数百个带有引导词的长句为实验数据，通过句法分析的特征进行自动拆分，准确率能达到 80% 以上。在自然语言处理顶级大会 ACL2015 会议中，来自斯坦福的 NLP 教授对英文长句划分短句作了详细报告，有望未来也能在中文处理上做进一步的验证，促进 SRL 在情报分析中的应用。

(2) 面向某一特定领域的 SRL

面向某一特定领域进行 SRL，能根据领域需求发掘更多的有效特征，提出有针对性的处理手法，领域移植性问题是自然语言处理中的根本性问题，针对某一特定领域的 SRL，虽然缩小和限制了领域，却是提高 SRL 效果的可行方法。

4.2 SRL 情报分析应用场景

(1) 文字的编辑校对

文字的编辑校对是对用词、语法等的检查、校对和编排，是自然语言处理应用的一个方向，在情报分析的数据监测中也有重要的应用。如果引入 SRL，就可以检查到语义的错误。比如有一句话：明子看到腿上流出的鲜血后才明白过来，原来他被狗咬了。如果进行语义角色的标注，那么 [ARG0 狗] [Predicate 咬] [ARG1 小明] 是正确的含义，如果语义角色在 ARG0 与 ARG1 之间颠倒了，说明该句子可能有语义错误，需要进行编辑校对。

(2) 自动文摘

SRL 在句式精简化方面的作用显而易见，如果将其中的语义角色按照主次进行划分，去掉次要角色，只保留句式的主要角色，这样就能得到一篇经过 SRL 处理过的简单的文摘。

(3) 机器翻译

Wu[35] 对统计机器翻译结果实行了 SRL 指导的重排序策略，Liu[36] 设计了语义角色重排序特征和删除特征，通过词对齐映射生成目标端的语义角色，帮助句子做重排序。目前 SRL 在机器翻译中的应用还较少，例如词对齐、短语抽取等。

结束语

大数据时代信息具有容量大、多源、密度低、维数高等特点，而数据处理要求响应快和智能化。数据量在剧增，“噪声”信息也在剧增，然而用户的要求越来越趋向于多元化、个性化和精准化。“用户为中心”成为这个时代自然语言处理的痛点，这一切都离不开高并发、智能化的信息处理能力。与此同时，情报分析的方式也正由传统宏观统计与计量的方式纵向发展为微观细颗粒的挖掘方式，宏观的计量方式优点在于能从整体上把握研究领域的框架体系，但对于框架中具体内容的处理能力远远不足。“大数据”处理能力在提升，但能精准挖掘分析文本内容的研究成果并不多，很大一部分的研究仍旧停在中宏观层面。SRL 作为语义学的重要基础技术，以其细颗粒，精细化的标注方式在浅层语义分析中起到了承接句法分析，启发深层语义分析的作用，在其基础研究得到一定发展后，情报人员应该将其作为一种实用化的工具，在各类情报分析中深度应用。

参考文献

[1] 曾建勋 , 魏来 . 大数据时代的情报学变革 [J]. 情报学报 , 2015,34(1):37-44.

[2]Levin B. English verb classes and alternations: a preliminary investigation [M]. Chicago: University of Chicago Press, 1993.

[3]Baker C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Project [C] // Annual Meeting of the Association for Computational, 1998:86-90.

[4]PALMER M, GILDEA D, KINGSBURY P. The proposition bank: an annotated corpus of semantic roles [J]. Computational Linguist, 2005,31(1):71-106.

[5] 袁毓林 . 语义角色的精细等级及其在信息处理中的应用 [J]. 中文信息学报 , 2007, 21(4):10-20.

[6] 刘怀军 , 车万翔 , 刘挺 . 中文 SRL 的特征工程 [J]. 中文信息学报 .2007,21(1):79-84.

[7] 丁金涛 , 王红玲 , 周国栋 , 等 . 语义角色标注中特征优化组合研究 [J]. 计算机应用与软件 ,2009,26(5):17-21.

[8] 李世奇 , 赵铁军 , 李晗静 , 等 . 基于特征组合的中文语义角色标注 [J]. 软件学报 ,2011,22(2):222-232.

[9] 王步康 , 王红玲 , 袁晓虹 , 等 . 基于依存句法分析的中文语义角色标注 [J]. 中文信息学报 ,2010,24(1):25-29.

[10] 李军辉，王红玲 . SRL 中句法特征的研究 [J]. 中文信息学报 ,2009(6):11-18.

[11] 刘挺 , 车万翔 , 李生 . 基于最大熵分类器的 SRL[J]. 软件学报 ,2007(3):565-572.

[12]Pradhan S, Hacioglu K, Krugler V, et al. Support Vector Learning for Semantic Argument Classification [J]. Machine Learning Journal, 2005,60(1): pp 11-39.

[13]Surdeanu M, Màrquez l, Carreras X, et al. Combination strategies for semantic role labeling [J]. Intelligence Research,2007,29:105-151.

[14] 白龙飞 . 基于依存树的中文语义角色标注技术研究 [D]. 沈阳 : 东北大学 , 2013.

[15] 谭伟 . 面向网络的中文问答系统相关技术的研究与系统初步实现 [D]. 北京 : 清华大学 ,2005.

[16] 吕德新 . 中文自动问答系统中问题理解技术的研究 [D]. 沈阳 : 沈阳航空工业学院 ,2006.

[17] 安强强 . 基于 SRL 的中文问答系统研究 [D]. 西安 : 西北大学 ,2009.

[18] 刘宁锋 , 史晓东 . 中文问答系统中答案抽取的研究 [J]. 电脑知识与技术 ,2011,7(12): 2865-2868.

[19] 张岩 . 基于语义角色的句子语义倾向判断 [D]. 北京 : 北京邮电大学 ,2008.

[20] 戴霖 . 网络舆情信息挖掘关键技术研究与应用 [D]. 浙江 : 浙江工商大学 ,2011.

[21] 施寒潇 . 细粒度情感分析研究 [D]. 苏州 : 苏州大学 ,2013.

[22] 鞠久朋 . 评价对象抽取研究 [D]. 苏州 : 苏州大学 ,2011.

[23] 王荣洋 . 评价对象抽取关键技术研究 [D]. 苏州 : 苏州大学 ,2012.

[24] 李景玉 , 张仰森 , 蒋玉茹 . 基于多特征融合的中文微博评价对象抽取方法 [J]. 计算机应用研究 ,2016, 33(2): 378-383.

[25] 于江德 , 樊孝忠 , 庞文博 . 事件信息抽取中 SRL 研究 [J]. 计算机科学 ,2008,35(3):25-27.

[26] 吴刚 . 基于主题的中文事件抽取技术研究及应用 [D]. 苏州 : 苏州大学 ,2009.

[27] 吴刚 , 许荣华 , 朱巧明 , 等 . 一种基于角色匹配的事件抽取方法 [J]. 微计算机信息 , 2010, 26(9):187-189.

[28] 张奇 . 信息抽取中实体关系识别研究 [D]. 合肥 : 中国科学技术大学 ,2010.

[29] 江超男 . 面向社会网络应用的关系抽取研究 [D]. 南京 : 南京理工大学 ,2010.

[30] 孟令恩 , 李颖 , 何彦青 , 等 . 基于 SRL 的专利主题提取研究 [J]. 图书情报工作 ,2014(19):19-24.

[31] 张帆 , 乐小虬 . 领域科技文献创新点句中主题属性实例识别方法研究 [J]. 现代图书情报技术 ,2015(5):15-21.

[32] 薛伟 . 文本分类相关问题研究 [D]. 济南 : 山东大学 ,2012.

[33] 王可 . 走近自然语言理解：言语意图分类研究 [D]. 大连 : 大连理工大学 ,2014.

[34] 张祎挺 .SRL 及其在句子相似度计算上的应用 [D]. 北京 : 北京邮电大学 ,2008.

[35]Dekai Wu, Pascale Fung.Semantic roles for SMT: a hybrid two-pass model[C]// NAACLHCT,2009:13-16.

[36]Ding Liu,Daniel Gildea. Semantic Role Features for Machine Translation[C]//the 23rd International Conference on Computational Linguistics, 2010:716-724.