语义角色标注研究综述

祝 娜[1]，王效岳[2]

（1. 山东理工大学科技信息研究所，山东淄博 255049；
2. 山东理工大学机械工程学院，山东淄博 255049）

摘要：通过查阅国内外语义研究领域的相关文献，梳理并分析了目前流行的语义角色标注（semantic role labeling，SRL）方法，展望了 SRL 技术的发展趋势。最后，重点综述了 SRL 技术在情报分析中的应用进展。

关键词：情报学；语义角色标注；综述；语义角色
中图分类号：G250.1 **文献标识码**：A **文章编号**：1674-2850(2015)08-0752-06

Review of semantic role labeling research

ZHU Na[1], WANG Xiaoyue[2]

(1. *Institute of Scientific & Technical Information*, *Shandong University of Technology*, *Zibo*, *Shandong* 255049, *China*;
2. *School of Mechanical Engineering*, *Shandong University of Technology*, *Zibo*, *Shandong* 255049, *China*)

Abstract: This paper searched and analyzed the relevant literatures of the semantic field from domestic and abroad, combed the current popular method of semantic role labeling (SRL), outlooked SRL technology's trends. Finally, we focused on the applications on SRL technology in the intelligence analysis.

Key words: information science; semantic role labeling; review; semantic role

0 引言

情报分析中的语义分析技术是将自然语言句子转化为反映这个句子意义（即句义）的某种形式化表示。用户检索信息时输入的是自然语言，而自然语言的分析和解释实际上都联系着不同的语法理论。

技术按现代语法知识将词语序列分组，并以语义角色对它们进行分类，它体现句子基本意思，该方法不对整个句子做详细的语义分析，只是标注句子中给定谓词（动词、名词等）的语义角色（参数），从而使计算机对语句有一个“浅层”的理解[1]，就是对给定句子中的每个谓词（动词、名词等）分析其相应的语义成分，并作相应的语义标记，如施事、受事、工具或附加语等。SRL 是浅层语义分析的一种实现方式，它不对整个句子进行详细的语义分析，实质是在句子级别进行浅层的语义分析，具有分析任务定义明确、便于评价等优点。目前，研究比较热门的技术有基于依存关系的 SRL，基于特征向量的 SRL 等。

在情报分析领域，海量信息的出现，无疑增加了分析处理信息的难度。情报学的研究领域已经向知识管理领域转变，广大科研人员研究中必须准确快速地把握科技文献创新内容，构建知识元，减少人力资源浪费，在此背景下 SRL 技术应用而生。

SRL 技术的出现帮助不同用户理解并获取知识，围绕知识搜集、描述、组织、检索和使用构建知识

作者简介：祝娜（1988—），女，硕士研究生，主要研究方向：文本数据挖掘技术
通信联系人：王效岳，教授，主要研究方向：数据挖掘与信息处理技术. E-mail: 346700849@qq.com

库与用户模型。知识组织 SRL 模型及其可行性方案的实现将会解决知识获取、求精和结构化等问题。目前，SRL 技术已经被成功地应用于问答系统、信息抽取、自动文摘、文本蕴涵、词义消歧、信息检索、指代消解、机器翻译、生物信息学等领域。其中信息抽取、自动文摘、信息检索等都是 SRL 在情报分析领域的典型应用。

1 数据分析

研究数据来源于 EV2（Compendex）以及中国知网。中国知网中以 SRL 为篇名检索字段，共检索 105 篇相关文献（人工去除 24 篇语言类等非相关技术文献）；在 EV2 数据库检索到 170 篇外文文献，人工整理发现 62 篇文献来源于中国，由于汉语语义的天然属性，对其进行 SRL 的研究比较多，发表文献也较多。运用数据库本身的分析功能，梳理了 SRL 技术的相关文献，如图 1 所示，对 2004 年至今国内外 SRL 和语义标注文献数量年分布再现。SRL 是近十年来研究的一种新的标注技术（如图 1 所示），国外这个领域的研究发展速度快，但是由于中文语义的特殊性，发表相关文献较多。

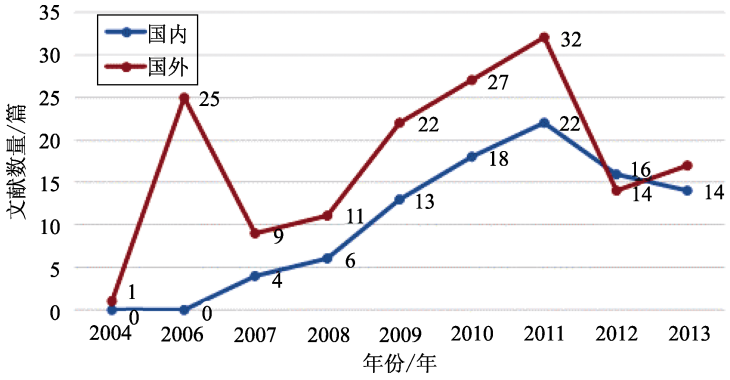

图 1 国内外 SRL 技术相关文献年分布图

Fig. 1 Relevant literature distribution of SRL technology from domestic and abroad

通过篇名的关键词为语义标注（模糊匹配）检索，国内张庆旭[2]在 1993 年发表了第一篇语义标注方面的专业文献，论述清华大学计算机系汉语真实文本自动语义标注取得突破性进展，自动标注正确率高达 93%. 以“SRL”命名的文献在 2006 年才出现（如图 1 所示），此文献也表明国内实际研究早于理论成果。LOWE 等[1]在 1997 年首次提出基于语义标注技术的框架语义标注方法，然而国外关于语义标注的研究早于国内很久。国内外 SRL 技术早在 1993 年之前就进入实用化阶段，而理论研究从 1993 之后才开始，实际应用早于理论研究，但理论研究发展速度快，发表的文献量迅速增多。但是篇名为 SRL 的文献（精确匹配）国内从 2006 年开始出现第一篇科技文献，国外从 2004 年已出现并进入快速增长期，国内比国外晚了两年，但是检索英文文献发现来自中国研究者有 59 篇。基于中文的语义的特殊性，相对于其他语言，SRL 技术对于中文更有实用价值。

2 SRL

2.1 语义标注技术出现

SRL 技术开始进入研究应用阶段，对计算机、自动化技术以及情报分析领域的研究都有重要影响。自从 FILLMORE 等[3]关于“深”语义类型的动词参数开创性论文的出现，关于语义角色是否存在、存在的多少以及它们的类别，科学界一直存在争论。1968 年，FILLMORE[4]提出的格语法理论是第一次系统化地将表面句法成分与潜在语义角色论点进行映射，奠定了 SRL 研究的理论基础。对句法成分的分析刚开始时，有些 SRL 研究进展缓慢，CHOMSKY[5]定义的 theta 标准指出，名词短语与 theta 角色之间一一对应，为句法成分分析服务，以此引入语义论点，取得了很好的效果。JACKENDOFF[6~7]试图通过使位于原始语义角色作为参数传递给深层概念的谓词，建立了词汇概念的结构理论。DOWTY[8]的研究放弃了

众多的语义角色标签，通过定义原型代理与其本身，以便涵盖一系列不同的语义角色类型。LEVIN[9]在寻找具有共同点的语法表达的基础上，将动词组合到某一类，这是为了捕捉同一共性的语义角色。同时，FILLMORE[10]继续探索词汇语义学领域，实现了动词类型的语义框架集合，该集合包括超过 2 000 多种的语义角色，被称为框架元素。而 SRL 技术的研究真正迅速积累不到十年时间，目前仍然是一个新兴领域，下面将进行更详细地解释。

2.2 SRL 方法

2.2.1 基于依存句法关系的 SRL

依存关系可以是句中词与词之间的句法关系，也可以是语义关系。依存句法使用了不同的信息组成方式，信息重组对于本地化特定谓词的语义角色非常有用。通过依存关系，依存树可以直接对与谓词节点相连的语义角色结构进行编码。英文方面，HACIOGLU[11]首次采用基于依存分析的方法来实现 SRL；LAFFERTY 等[12]在研究自然语言处理曾应用过此机器学习方法。相比基于句法成分的英文 SRL，中文基于依存关系的标注对词汇的依赖性较弱，鲁棒性较高。

2.2.2 基于特征向量的 SRL

比较文本信息中不同特征，以基本的特征和语料数据资源为基础，筛选便于识别和分类的特征进行 SRL. DING 等[13]提出了一种层次化特征选择策略，XUE 等[14]在 GILDEA 标准特征集合的基础上尝试了组合特征，BOXWELL 等[15]提出一种基于丰富特征的 SRL 方法。目前的研究一般将基本的特征加以组合来降低特征空间的维度，提高 SRL 的性能；人工智能知识表示中的“框架-槽”、模式识别中的“特征”、多媒体数据库中的“特征向量”等已经开始应用此技术。

2.2.3 基于最大熵分类器的 SRL

基本思想是为所有已知的因素建立模型，把所有未知因素排除在外。预测了全部能够在句法分析树中找到匹配成分的角色后，采用简单的后处理规则去识别那些找不到匹配成分的角色。使用最大熵模型进行 SRL 的有 XUE 等[14]和刘挺等[16]。XUE 等[14]使用最大熵模型进行实验，在基于单一句法树的基础上，详细验证了 GILDEA 等[17]研究中的各个基本特征在 SRL 各阶段的贡献，并提出了新的特征。刘挺等[16]选取了较多的特征采用最大熵分类器对句子中谓词的语义角色同时进行识别和分类，取得很好效果。

2.2.4 基于核函数的 SRL

一般使用核函数的目的是将低维线性不可分问题映射到高维空间，使之成为线性可分问题。可以通过计算核函数隐式达到，从而降低时间和空间复杂性很好地融入到 SRL 技术中。MOSCHITTI[18]最早使用核函数的方法来实现 SRL，CHE 等[19]在其研究基础上，将 PAK 核分为路径核和成分结构核，通过线性组合集成 2 个核。ZHANG 等[20]指出传统树核函数都是“硬”匹配，不利于计算相似成分或近义的语法标记，支持向量机（support vector machine，SVM）、感知器等学习算法。

2.2.5 基于条件随机场的 SRL

条件随机场（conditional random fields，CRFs）模型擅长处理序列标记问题，是一种无向图模型，考察给定输入序列对应的标注序列的条件概率，训练目标是使条件概率最大化。它以浅层句法分析为基础，把短语或命名实体作为标注的基本单元，将 CRFs 用于句子中谓词的 SRL. COHN 等[21]在 PropBank 句子的完全句法分析树上建立树 CRFs 标注模型，标注结果好于最大熵模型；董静等[22]考虑句法树中水平层次上的角色标签之间的马尔科夫依赖关系，在 PropBank 语料基础上进行了 SRL 实验；YU 等[23]使用 CRFs 模型研究了英文 PropBank 的 SRL 问题，使用浅层短语块和命名实体块作为标注单位。以上使

用 CRFs 模型的 SRL 实验都取得了不错的效果。

3 SRL 在情报研究中的应用

随着互联网的普及和推广，信息量剧增，对情报分析的研究越来越注重利用计算机技术自动获取情报，过滤无用信息，从而使计算机具备一定程度的语言理解和处理能力。现阶段使用机器学习的方法进行语义句法分析是自然语言处理领域的一个热门方向，并且取得了不错效果。SRL 给研究者提供了一个很好的测试机器学习技术的平台。

3.1 数字图书馆建设

数字图书馆实质上是一种多媒体制作的分布式信息系统，把各种不同载体、不同地理位置的信息资源用数字技术存贮，以跨越区域面向对象的网络查询和传播的一个大型信息系统。目前，数字化图书馆的建设将知识作为工作对象，围绕知识搜集、描述、组织、检索和使用展开研究。运用 SRL 技术可以创新和丰富数字图书馆知识组织系统的语义关系层面，探索出数字图书馆知识组织 SRL 模型及可行性方案；并且如果将 SRL 模型应用于数字图书馆信息抽取、语义标注、知识库构建、用户模型构建等知识组织环节中，最后会构建出面向各学科领域知识库，帮助各领域学者方便快速查找所需要的科技知识。

3.2 信息检索

情报分析中必然要运用信息检索技术收集信息。信息爆炸时代，绝大部分知识是用文本描述的，对文本进行分析处理必然涉及自然语言处理技术。现代化信息检索技术是应用信息技术和自然语言处理技术让计算机模拟人类的思维活动，提高计算机的智能水平，目的旨在解决知识获取、求精和结构化等问题。在信息检索过程中运用 SRL 技术来帮助并引导用户进行检索；可以根据不同用户的行为习惯，提供不同的检索方式，易于不同用户的理解与帮助。

信息检索过程中的自动标注技术能对数据进行自动化处理——浅层语义分析步骤能够识别出一个动词的施事、受事以及该动作发生的时间、地点等信息，这对回答一些针对该动作的问题是非常有帮助的。SRL 技术在信息检索领域的其他典型应用有语义网，社会标注系统，导航系统及数字参考咨询系统等。

3.3 信息抽取

信息抽取是把文本里包含的信息进行结构化处理，变成表格一样的组织形式，即输入原始文本，经过处理后，输出如命名实体、实体关系、事件等固定格式的信息点。

利用 SRL 可以很好地完成这种事件的抽取，需要找出每个事件对应的触发动词，然后标注出它的各种语义角色作为事件的参数，并根据参数的特点，进一步确认是否为需要事件，以提高实体抽取的准确率。在实际应用中，参数往往是较长的人名、地名、机构名、时间和数字等，因此对这些实体的识别变得异常重要。同时，浅层语义分析本身也可以看作是一种通用的信息抽取，只是人们事先定义的结构化信息并不局限于特定的领域，因此浅层语义分析技术的发展，必将会促进信息抽取技术的发展。SURDEANU 等[24]已经利用浅层语义分析技术进行了信息抽取实验，得到了较理想的效果。

3.4 科技文献知识抽取

科技文献中蕴含着大量知识，有很高的应用价值，所以对科技文献信息定位要求越来越精确。对科技文献的情报分析研究已经向知识获取、组织和利用的方向发展，其中必然涉及自然语言处理技术。研究人员可以利用 SRL 技术对科技文献进行分析标注，抽取研究者需要的研究背景、研究方法、关键技术、应用前景、发展趋势等知识信息，然后让知识信息以统一的形式集成在一起以方便检查和比较。这些信

息可以导入关系数据库等结构化数据管理工具，方便进行查询、分析和研究等操作。

科技文献大部分以文本格式存在，SRL 又是浅层语义分析领域研究的关键技术，其对文本进行处理并标注（此外还包括相似度的计算），让反馈的信息按照用户预定的模式输出，可以利用此技术突破科技文献文本信息语义研究瓶颈，比如词义消歧技术。可以在特定的上下文中确定多义词的意思，以及其在学术抄袭检测系统、参考文献自动标注、文献自动综述系统等方面的应用都取得很好的效果。

4 结论与展望

捕捉准确的信息是情报分析研究的重点工作，然而目前各种语义模型，虽然研究得很热，但极少能真正大规模应用；另外信息检索系统对性能的苛刻要求，使语义检索、个性化检索等知识密集型的技术很难应用。情报分析的根本任务是通过查询、分析、定位，最后反馈给用户特定格式的信息。目前大部分信息存在于文本中，对信息进行语义层面的分析，有利于准确高效地定位并提取信息，而深层语义分析技术研究不是很理想，很多学者都想在浅层语义分析方面寻求突破，希望找到一种新的更有效的方法，SRL 技术是一个突破口。

目前 SRL 技术可利用的语料资源有限，浅层语义分析又遭遇数据稀疏等问题，所以要构造更具有泛化能力的系统对词块、短语和从句等进行识别标注。另外，SRL 也需要有效的机器学习算法和统计分析模型，使目标语块不存在重复的语义角色，使整个模型对文本内容进行覆盖实现对句子成分的分类。

SRL 技术有几个有前途的途径，用分而治之的方法降低完全句法分析的难度，用组块的分析来简化句子的结构。通过改进语法/语义交互，打破句法分析性能的瓶颈，在更好概括并有效利用 PropBank[25]，VerbNet[26]和 FrameNet[27]之间语义链接的基础上达到动词类的有效利用，将 SRL 映射到平行语料库，对新的语言产生训练数据，把独立的谓词结构参数与复杂的事件陈述链接，把推理规则与 VerbNet 和 FrameNet 语义谓词相互关联，创造出更丰富的语义特征，以及为这些任务发展监督和无监督的方法等。

这里从文献计量分析的角度，梳理了关于 SRL 技术的科学文献；概述了目前此技术主要的研究热点及其未来展望，后续的研究将进行 SRL 技术在具体应用在情报分析中的作用。基于 SRL 的情报分析无疑为图书情报领域构造了一个完美的语义空间，找到较为完善的数学模型，完善了已有的人工构造的语料集，将会进一步提高情报分析过程中对自然语言处理的效果。

[参考文献]（References）

[1] LOWE J B, BAKER C F, FILLMORE C J. A frame-semantic approach to semantic annotation[C]//Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How. 1997: 4-5.

[2] 张庆旭. 汉语真实文本自动语义标注取得突破性进展[J]. 语文建设，1993（9）：37-44.
ZHANG Q X. The breakthrough of Chinese real text automatic semantic annotation[J]. Languages Construction, 1993(9): 37-44. (in Chinese)

[3] FILLMORE, CHARLES J. The case for case[C]//Universals in Linguistic Theory. New York: Holt, Rinehart & Winston, 1968: 1-88.

[4] FILLMORE C J. Toward a modern theory of case[M]. USA: Department of Health, Education and Welfare, 1968.

[5] CHOMSKY N. Lectures on government and binding: the pisa lectures[M]. Berlin: Walter de Gruyter, 1993.

[6] JACKENDOFF R S. Semantic interpretation in generative grammar[M]. Cambridge, MA: MIT Press, 1974.

[7] JACKENDOFF R S. Semantics and cognition[M]. Cambridge, MA: MIT Press, 1983.

[8] DOWTY D. Thematic proto-roles and argument selection[J]. Language, 1991, 67(3): 547-619.

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

[10] FILLMORE C J. Frames and the semantics of understanding[J]. Quaderni di Semantica, 1985, 6(2): 222-254.

[11] HACIOGLU K. Semantic role labeling using dependency trees[C]//Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg: ACM Press, 2004: 1273.

[12] LAFFERTY J. McCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning 2001(ICML 2001). ACM Press, 2001: 282-289.

[13] DING J T, WANG H L, ZHOU G D, et al. On optimized combination of features in semantic role labeling[J]. Computer Applications and Software, 2009, 26(5): 17-21.

[14] XUE N W, PALMER M. Calibrating features for semantic role labeling[C]//Proceedings of the EMNLP. Stroudsburg: ACL, 2004: 88-94.

[15] BOXWELL S A, MEHAY D, BREW C. Brutus: a semantic role labeling system incorporating CCG, CFG, and dependency features[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language. ACM Press, 2009: 37-45.

[16] 刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007，18（3）：565-573.
LIU T, CHE W X, LI S. Semantic role labeling with maximum entropy classifier[J]. Journal of Software, 2007, 18(3): 565-573. (in Chinese)

[17] GILDEA D, PALMER M. The necessity of parsing for predicate argument recognition[C]//Proceedings of the 40th Meeting of the Association for Computational Linguistics. Philadelphia, PA: ACM Press, 2002: 239-246.

[18] MOSCHITTI A. A study on convolution kernels for shallow semantic parsing[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: ACM Press, 2004: 335-342.

[19] CHE W X, ZHANG M, LIU T, et al. A hybrid convolution tree kernel for semantic role labeling[C]//Proceedings of the COLING/ACL on Main Conference Poster Sessions. Stroudsburg, PA: ACM Press, 2006: 73-80.

[20] ZHANG M, CHE W X, ZHOU G D, et al. Semantic role labeling using a grammar-driven convolution tree kernel[J]. IEEE Transaction on Audio, Speech and Language Processing, 2008, 16(7): 1315-1329.

[21] COHN T, BLUNSOM P. Semantic role labelling with tree conditional random fields[C]//Proceedings of the Ninth Conference on Computational Natural Language Learning. Stroudsburg, PA: ACM Press, 2005: 169-172.

[22] 董静，孙乐，吕元华，等. 基于线性链条件随机场模型的语义角色标注[C]//中国中文信息学会二十五周年学术会议，北京，2006：32-37.
DONG J, SUN L, LÜ Y H, et al. Semantic role labling based on conditional random[C]//Chinese Information Processing Society Conference Twenty-fifth Anniversary of China. Beijing, 2006: 32-37. (in Chinese)

[23] YU J D, FAN X Z, PANG W B, et al. Semantic role labeling based on conditional random fields[J]. Journal of Southeast University (English Edition), 2007, 23(3): 361-364.

[24] SURDEANU M, HARABAGIU S, WILLAMS J, et al. Using predicate-argument structures for information extraction[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: ACM Press, 2003: 8-15.

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

[26] KIPPER K, DANG H T, PALMER M. Class-based construction of a verb lexicon[C]//Proceedings of the Seventeenth National Conference on Artificial Intelligence. Austin, TX: AAAI, 2000: 1-6.

[27] BAKER C F, FILLMORE C J, LOWE J B. The berkeley framenet project[C]//Proceedings of the 17th International Conference on Computational Linguistics. Stroudsburg, PA: ACM Press, 1998: 86-90.