朝鲜语句子语义角色标注研究*

毕玉德 陈洁 吕春燕

解放军外国语学院亚非语系 洛阳 471003

E-mail: biyude@gmail.com; chenjie003@yaohoo.com.cn; caiqianhuily@sina.com

摘要：语义角色标注是目前自然语言处理的一项研究热点。本文采用理性主义和经验主义相结合，以实用主义为原则，从语义信息处理角度，提出了一种朝鲜语语义角色标注的研究思路，即以朝鲜语动词句法语义层次框架为理论基础，辅之以基于特征向量的方法，并结合指称类概念分类标注库，以标注语料库为试验对象，并进行测试的方法，进行语义角色标注研究。

关键词：朝鲜语；语义角色标注；句法语义；特征向量

A Study on Semantic Role Labeling of Korean Sentence

BI Yu-de, Chen Jie & Lv Chun-yan

PLA University of Foreign Languages, Luoyang 471003

E-mail: biyude@gmail.com; chenjie003@yaohoo.com.cn; caiqianhuily@sina.com

Abstract: Semantic Role Labeling is a hot research in natural language processing. Together with rationalism and empiricism, and with the principle of practicality, this paper poses a study method to label the semantic role of Korean from the side of language processing. This method, based on the theory of syntax and semantic framework on Korean verbs, and with the assistance of feature vector, is used to test signed corpus for the study on semantic role labeling.

Keywords: Korean, Semantic Role Labeling, Syntax and Semantic, Feature Vector

1 引言

语义分析一直是自然语言理解的主要目标之一。所谓语义分析，指的是根据句子的句法结构和句中每个实词的词义推导出能够反映这个句子意义的某种形式化表示。通过语义分析，可以理解自然语言语句，并进行深入的知识获取和推理，从而计算机能够与人类无障碍的沟通。语义角色标注，又被称作浅层语义分析，是对深层语义分析的一种简化，它只标注与句子中谓词有关的成份的语义角色，如施事，受事，时间和地点等等。它在问答系统、信息抽取、机器翻译、词义消岐等自然语言理解的诸多领域，都有着广泛的应用，能够产生巨大帮助。如在信息抽取领域，语义角色的使用已经能够建立更健壮的统计系统；在统计模型中增加语义角色的标注信息，可以提高句法分析器和语音识别的准确率。

近年来，国内外在语义角色标注方面已经进行了一些卓有成效的研究。目前大多采用统计学习的方法解决语义标注问题。如最大上模型、支持向量机、决策树模型、基于树结构的条件随机场模型、基于记忆的学习方法和基于转换的错误驱动学习方法等等。

浅层语义分析，和其它基于统计的自然语言处理技术一样，同样需要好的语料资源。

* 该研究得到国家自然科学基金项目（编号：60673036）支持。

目前，英语较为知名的浅层语义分析资源为 FrameNet 和 PropBank。其中、U.C.Berkeley 开发的 FrameNet 以框架语义为标注的理论基础对英国国家语料库进行标注。它试图描述一个词汇单元(动词和部分名词以及形容词)的框架，同时也试图描述这些框架之间的关系。除英语外，许多语言都建立了各自的浅层语义标注库，例如，SALSA 是德语版的 FrameNet；我国山西大学以框架语义学为理论基础，参照 FrameNet，构建了一个以有限词语集合为描述对象的汉语框架语义网（CFN，Chinese FrameNet），其中队汉语 1760 个词元构建了 130 个框架，标注了 8200 条句子。

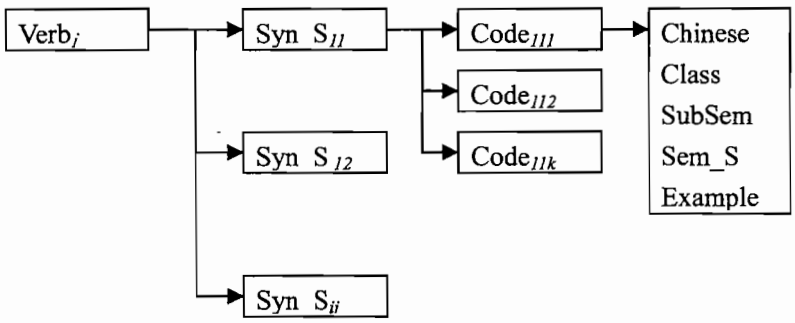
朝鲜语研究方面，Kim Byoung-Soo 等采用基于非指导学习的方法对朝鲜语副词格进行了语义标注研究，还采用 bootstrapping 算法对朝鲜语格助词进行了语义标注研究；Shin Myung-Chul 等利用韩国世宗计划电子词典和利用功能动词句法与概念的相似度分别对朝鲜语副词格进行了语义角色标注研究。

2 基本理念

2.1 研究思路

长期以来，我们一直从事朝鲜语/韩国语动词句法语义知识库方面的研究，几经修改完善形成了符合朝鲜语/韩国语言事实的动词句法语义信息的框架体系。在设计思路上，我们汲取对自然语言进行整合描写的理念，将词典和语法融为一体，其基本思路就是词汇语法化和语法词汇化，它充分反映了通过句法描写语义的技术路线和句法语义一体化描写的思想。该框架库以动词为基本单元，它包括四个层次，第一层是朝鲜语形态动词；第二层是基本句法结构信息，第三层是句法语义分类代码信息，第四层包含了汉语释义、词性信息、语义结构、语义特征（因子）以及例句等词库信息。

该框架中主要把句法知识和语义知识整合在了一起，框架结构格局如下：



我们将主要以该层次框架为理论基础，辅之以基于特征向量的方法，并结合指称类概念分类标注库，以韩国世宗计划完成的 1000 万短语标注语料库为试验对象，进行语义角色标注研究。将来我们计划选取其中 5000 条语句进行自动标注，然后进行人工校对，并对试验结果进行比对分析，根据校对的结果，修改特征向量；最后对 10000 条语句进行标注。

2.2 标注集的确定

经过长期的研究，我们建立现代朝鲜语的语义角色系统。该系统以陈望道在《修辞学发凡》中提出的“六何”思想为源，在充分观察朝鲜语言事实的基础上，以“假设－演绎”模型为研究模式，从语言反映的对象世界中的各种认知联系出发，本着语义结构研究最终要落实到语形结构（表层结构、句法结构）的精神，以语义结构及其构成因子——语义参

与者为基础，系统地全面地观察朝鲜语可能存在的所有语义结构，最终构拟出朝鲜语事件语义结构的层级推演系统。这种系统化的语义角色清单不仅对以动词为中心的语法描写有利，而且对机器翻译等应用领域也大有裨益。这个语义角色系统分为三个层级。在本课题的研究中我们将以此为标记集，对语料库中选取的语句进行语义角色标记。

2.3 语言资源的选择

韩国世宗计划 2002 年完成了 1000 语节的语料库并进行了标注，包括书面语（80%）和口语（20%）两大部分，其中书面语部分涉及报刊、杂志、书籍等，内容包括政治、经济、社会、外交、媒体、生活、综合等，语料库附有详细的说明和目录。标注采用自动标注加人工检查的方式。我们将选取 1000 条语句进行实验，并选取 1000 条语句作测试语料。例如：

무엇보다도 쌀 수매 정책의 후퇴를 중단해야 한다.

→ 무엇/NP+보다/JKB+도/JX 쌀/NNG 수매/NNG 정책/NNG+의/JKG 후퇴/NNG+를/JKO 중단/NNG+하/XSV+아야/EC 하/VX+ㄴ다/EF+./SF

3 标注步骤

3.1 标记单元的确定

语义标注的基本单元可以使句法成分、短语、词或者依存关系等等，现在大多数语义角色标注系统通常都以句法成分为基本标注单元。

语义角色标注系统一般通过三个阶段实现：首先把多数不可能是语义角色的成分过滤掉；其次进行语义角色标注；最后使用多类分类器把识别的语义角色分到对应的类别。

鉴于朝鲜语的句法特点，同时利用我们设计的包含句法和语义信息的层次框架结构体系，我们初步设计了语义角色标注步骤，下面以例句说明如下：

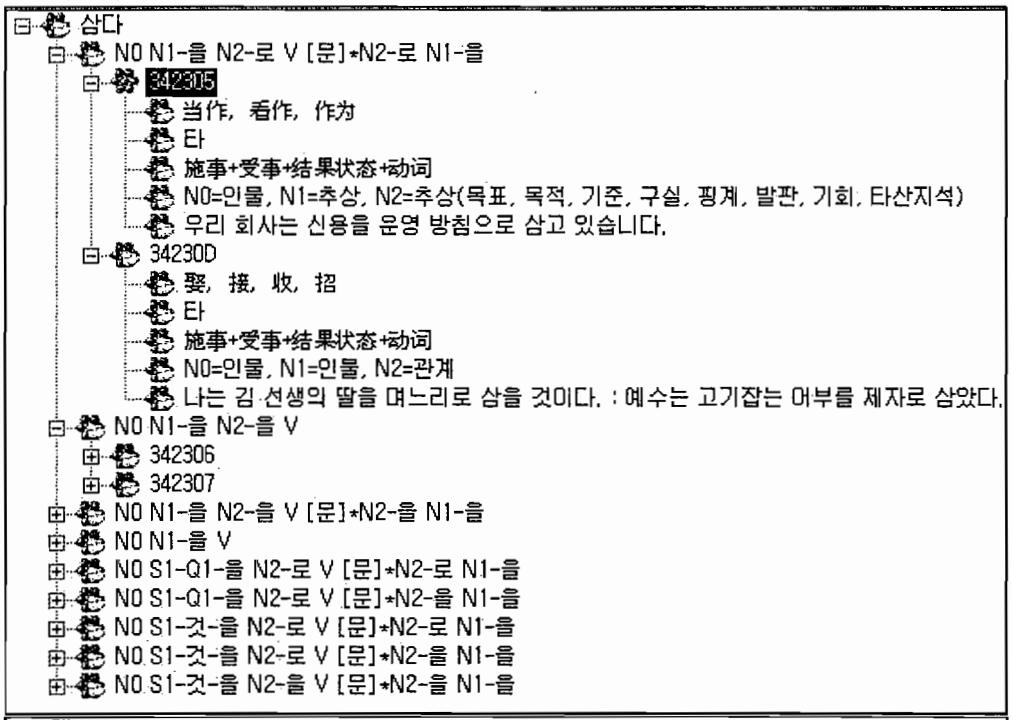
a. 原文：나는 그 기쁨과 슬픔을 모두 그리움의 양식으로 삼아 놓기 위해서 이 소설을 엮었다.
b. 译文：我为了把喜悦和痛苦都当作思念的食粮编写了这个小说。
c. 标记：나/NP+는/JX 그/MM 기쁨/NNG+과/JC 슬픔/NNG+을/JKO 모두/MAG 그리움/NNG+의/JKG 양식/NNG+으로/JKB 삼/VV+아/EC 놓/VX+기/ETN 위하/VV+아서/EC 이/MM 소설/NNG+을/JKO 엮/VV+었/EP+다/EF+./SF

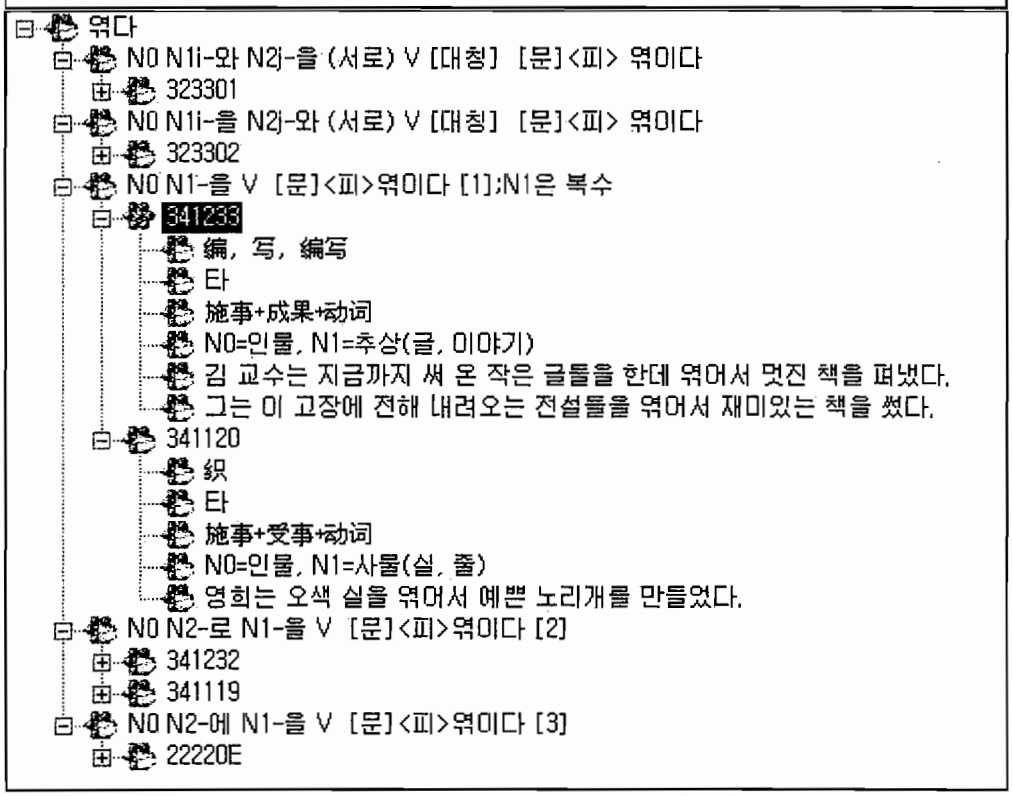
①首先进行单复句区分，根据朝鲜语连接词尾把复句拆分单句；例句中“-기 위해서”是表示目的的词尾，由此该句分为连个单句；由于主语“나는”的形态标识是“는”，所以是主句的成分（做主语），同时也是从句的主语，所以目的状语从句是“그 기쁨과 슬픔을 모두 그리움의 양식으로 삼아 놓-/把喜悦和痛苦当作思念的食粮”，而主句为“나는 S'-기 위해서 이 소설을 엮었다./我为了S'编写了这部小说。”

②选中谓词：朝鲜语属于 SOV 型语言，动词在句子结尾（或分句结尾），是句尾动词语言(Verb-final language)，根据这一特点以及标注的结果[1]，可以比较容易地确定谓词，同时去除谓词词干后的情态成分即非语义角色成分。以上例句中的谓词是“삼/VV+아/EC 놓/VX→삼아

[1] 世宗计划语料库句法标注采用的标记集较我们构建的朝鲜语动词句法语义知识库采用的句法标记集更为详细，为此我们可以建立一个对应关系表。

놓-(从句)”和“엮/VV→엮-（主句)”，其中从句谓词的主动词是“삼다”、辅助动词是“놓다”，主句谓语动词是“엮다”。





③确定与该谓词相关的句法成分即论元：因为一个谓词在组句时潜在多种句法结构和多个句

法语义项（如上图所示），对此我们首先可以根据句中标注的各个论元的句法特征，选定该句中谓词所属的句法结构类型。例句中，从句主动词“삼다”的句法结构类型是“NO N1-를 N2-로 V”，其中 NO=나/我（实际上与主句主语一致），N1=그 기쁨과 슬픔/那喜悦和痛苦，N2=그리움의 양식/思念的食粮；主句动词“엮다”的句法结构类型是“NO N1-를 V”，可知 NO=나/我，N1=이 소설/这部小说。

④根据各论元中心词的语义特征，确定其所属语义结构类型，从而确定该动词在该句中所属句法语义项。此时句中主要论元的语义角色基本可以确定下来。上例从句中，根据 N1 和 N2 的语义特征，可判断出动词“삼다”所取句法语义项是 342305，对应的语义结构是“施事+受事+结果状态+动词”，由此可判定 N1 的语义角色是“受事”，N2 的语义角色是“结果状态”；主句动词相应的句法结构下有两个句法语义项，根据论元 N1 的语义特征（抽象：作品）可判定为属第一句法语义项 341233，对应的语义结构为“施事+成果+动词”，因此，NO 的语义角色是“施事”，N1 的语义角色是“成果”。事实上，第二个句法语义项的语义结构与第一个相同，但由于 N1 要求的语义特征不同（抽象或具体），故在层次框架结构中分别列出。

⑤根据句子出现的论元的语义特征和句法特征确定次要语义角色。如“오늘-에는”中，“오늘/今天”是时间名词，而“-에는”表示时间或存在场所，据此我们可以判定论元“오늘/今天”的语义角色为“时间”。

3.2 需要特殊处理的内容：

①未列入基本句法结构的论元的语义角色，需要根据其语义特征及句法标识来确定。在语义角色（主要角色和次要角色）的句法标识已经确定，因此需要根据指称类概念的分类体系建立指称类概念标注库，目前我们已经对 22 万条概念进行了初步标注。下一步工作是根据分类体系进行更细致的标注。

②特殊结构的处理：该类情况可以枚举，可以建立基本特征列表，根据其句法特征，进行识别。如：对于结构“N1 와 N2”，可根据 N2 后接助词来确定，当后接宾格助词“를/을”时，该结构为合成宾语；当后接主格助词“가/이， 께서(는)， 는/은”时，该结构为合成主语；当后接宾格助词“에，로”等时，其谓词为趋向动词，若 N1 为人称代词或人名，则 N1 为伴随主语，N2 为达到场所或方向。

③未知谓词的处理：根据相关句子成分，计算该谓词与已知谓词数据库中句法语义项的相似度，判断其归属。

④复句的处理：首先枚举所有连接词尾并标记其意义（并列、转折、让步等），并据此区分主从句，然后分别处理。

⑤定语从句的处理：由于朝鲜语是句尾动词语言，定语从句的成分全部在从句谓词之前，这些成分的语义角色标注可按一般单句处理，关键是从句前面边界的确定，究竟是主句成分还是定语从句的成分，对此建立特征集的办法来处理。

3.3 需要解决的问题

我们的研究方案是以朝鲜语动词句法语义层次框架为理论基础，辅之以基于特征向量的方法，并结合指称类概念分类标注库，进行语义角色标注研究。目前我们基本构建了朝鲜语动词句法语义层次框架库，还需要解决以下几个关键问题。

①特征向量集的确定：特征在句法成分的识别和语义角色的判断上，起着非常重要的作用，朝鲜语是形式化非常发达的语言，我们将根据这一特点，结合指称类概念分类标注库，进行综合判断。

②指称类概念分类标注库：目前我们已经对 22 万条概念进行了初步标注。下一步工作

是根据分类体系进行更细致的标注。

4 结束语

本文试图采用理性主义和经验主义相结合，以实用主义为原则，从语义信息处理角度，提出了一种朝鲜语语义角色标注的研究思路，它首先以类似于 FrameNet 的层次框架为基础，然后辅之以句法语义特征向量，这对于形态特征非常突出的朝鲜语句子的语义角色标注，非常有效；同时，采取语料库检验的方法，可以避免出现容易忽略语言中那些经验性的、小粒度的知识以及难以覆盖各种复杂纷繁的语言现象的问题，从而大大提高标注的准确率。

参 考 文 献

[1] Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles [J]. Computational Linguistics, 2005, 31(1).

[2] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet project [C]. Proceedings of ACL&Coling-1998. 86-90.

[3] Gildea D, Jurafsky D. Automatic labeling of semantic roles [J]. Computational Linguistics, 2002, 28(3): 245-288.

[4] Thompson CA, Levy R, Manning CD. A generative model for semantic role labeling [C]. Proceedings of ECML-2003, Springer Berlin Heidelberg, 2003. 397-408.

[5] Liu T, Che W X, Li S, et al. Semantic role labeling system using maximum entropy classifier [C]. Proceedings of CoNLL-2005. Ann Arbor, Michigan, 2005.189-192.

[6] S. Pradhan, K. Hacioglu, V. Krugler, et al. Support vector learning for semantic argument classification [J]. Machine Learning Journal, 2005.

[7] N. Xue, M. Palmer. Automatic semantic role labeling for Chinese verbs [C]. Proceedings of IJCAI2005, 2005.

[8] P. Koomen, V. Punyakanok, D. Roth, et al. Generalized Inference with Multiple Semantic Role Labeling Systems [C]. Proceedings of CoNLL-2005, 2005. 181-184.

[9] C.F. Baker,C.J. Fillmore and J.B. Lowe. The Berkeley FrameNet project. InC.Boitet Proceedings of pages

[10] Byoung-Soo Kim, Yong-Hun Lee and Jong-Hyeok Lee. Unsupervised Semantic Role Labeling for Korean Adverbial Case [J]. Software and Application [34-2], 2007. 112-122.

[11] Byoung-Soo Kim, et al. Bootstrapping for Semantic Role Assignment of Korean Case Marker[C]. Proceedings of Korea Computer Conference(B), 2006. 4-6.

[12] Shin, Myung-Chul, et al. Semantic Role Assignment for Korean Adverbial Case Using Sejong Electronic Dictionary[C]. KISTI Press, 2005. 122-130.

[13] Yude Bi, Binhong Wu. A Study on the Structure of Korean Knowledge Database[C]. Proceedings of PACIC20, 2006.

[14] 刘开瑛等. 汉语框架元素自动标注实验[C]. 第四届信息检索与内容安全学术会议(NCIRCS2008), 2008.

[15] 于江德，樊孝忠，庞文博. 事件信息抽取中语义角色标注研究[J]. 计算机科学，2008 (3).

[16] 毕玉德. 现代韩语动词语义组合关系研究[A], 北京民族出版社, 2005.

[17] 韩国国立国语研究院. 21 世纪世宗计划语料库. 2004.