韩国语名词短语结构特征分析及自动提取

安帅飞 毕玉德

解放军外国语学院语言工程系 河南洛阳 471003

E-mail: anshuaifei2013@sina.cn biyude@gmail.com

摘 要： 名词短语作为语言中一种普遍的语法现象，在自然语言处理领域日益受到了研究人员的关注。目前，对其研究范围主要集中在边界识别、语法分析、语义分析及其分类等方面。本文通过研究分析韩国语书面语名词短语的左右边界规则，从大规模标注语料库中自动提取出名词短语。实验结果表明：语料中的高频名词短语相对集中于 8 个类型之中。根据提取结果分别建立不同类型的名词短语库，为进一步建立双语平行短语语料库打下基础，以便于以后的机器翻译、信息检索等自然语言信息处理工作。

关键词：韩国语； 名词短语； 标注语料库； 边界界定； 自动提取

Structure Characteristic Analysis and Automatic Extraction for Korean Noun Phrase

AN Shuaifei, BI Yude

PLA University of Foreign Languages, Luoyang 471003

E-mail: anshuaifei2013@sina.cn, biyude@gmail.com

Abstract: These years， noun phrase， as a common grammatical phenomenon， has attracted eyes of many scholars in the field of natural language processing. At present， most researches on noun phrase lie in boundary identification， grammatical analysis， semantic analysis， categorization and some other aspects. This thesis abstracts noun phrases from a large-scale tagged corpus through studying and analyzing rules of left and right boundaries of noun phrases in written Korean. From the experimental result， we can see that high-frequency noun phrases mainly lie in 8 categories. Different kinds of corpus for noun phrases can be built according to the result of the abstract， which lays the foundation of building paralell corpus. It will also be convenient for machine translation， information retrieval and other work in language information processing in the future.

Key words: Korean; noun phrase; tagged corpus; boundary identification；automatic extraction

1.前言

名词短语在自然语言中出现频繁，是自然语言处理领域中重要的资源，而且相比单个词汇，名词短语的确定性更高，可以解决绝大部分的局部歧义结构问题，从而为进行更深入的语块分析和完全句法分析打下基础。同时，名词短语的分析结果可以在一定程度上简化句子的结构，降低句法分析的复杂度，在自然语言处理领域有着重要的意义。另外，在机器翻译中的歧义消解、快速匹配等方面，名词短语库都是重要的知识资源。然而手工构建名词短语库存在诸多缺点：工作量极大、覆盖度不全、不易更新。因此，本文着眼于名词短语库的自动构建，通过研究分析名词短语的左右边界规则，自动提取出名词短语并建立分类短语库。

2.韩国语名词短语及语料库相关说明

2.1 韩国语名词短语

BaseNP（基本名词短语）这一概念是 Church 在英语中首次提出的，他将英语 baseNP 定义为“简单的非嵌套的名词短语”，也就是说，一个 baseNP 内部不能再包含有更小的名词短语（Church K，1988）。在汉语的基本名词短语定义方面，赵军（1992）给出了形式化的定义：

BaseNP→BaseNP+BaseNP

BaseNP→BaseNP+名词|名动词

BaseNP→限定性定语+BaseNP

BaseNP→限定性定语+名词|名动词

限定性定语→形容词|区别词|动词|名词|处

所词|西文字串|（数词+量词）

在韩国语中，名词短语被定义为句子中起名词作用的短语，而名词被定义为表示事物的名称的词性范畴。（李基文，1997）本文将参照Church和赵军对英、汉基本名词短语的形式化定义，根据韩国对名词短语的定义，通过语料库提取句中的名词性成分，然后对韩国语名词短语进行归纳分析。需要特别说明的是，韩国语中谓词作定语时，是通过谓词词干加上冠形词性转成词尾实现的，因为谓词是一个句子的中心，所以包含谓词的限定性成分一般被视为句子降格做句子成分，相当于英语中的定语从句，此类情况不在本文的讨论范围之内。

2.2 韩国语名词短语识别研究现状

自然语言处理的研究方法从大类可分为规则方法和统计方法。目前很多韩国学者对韩国语名词短语的研究大多是综合考虑语序、词性、分写法等上下文相关信息，并赋予其不同的权值进行统计学上的分析处理。其中，양재형（2000）的名词短语B（begin）I（inside）O（out）标记法， 황영숙（2002）的结构树方法等研究都取得了不错的进展。尤其是서충원（2003）等人通过构建三元隐马尔科夫模型研究基本名词短语，并添入了短句的中心词等相关信息，通过计算形态素与标记的对应概率值，利用统计出的概率高的对应组帮助机器识别出基本名词短语，取得了很好的效果。本文将通过规则方法识别名词短语，在名词短语的分类上将高频的名词短语集中分类。

2.3 韩国语标注语料库及相关说明

本文实验所用的语料库是韩国国立国语研究院“21 世纪世宗计划”开发的标注语料库。“21 世纪世宗计划”是韩国政府为推动韩文信息化发展，自 1998 年开始实施的计划，并于 2007 年建成了大规模的语料库。该语料库包含新闻、杂志、小说等各类体裁。实验时选取了体裁为新闻和小说的 207634 个句子作为训练语料。

3.韩国语名词短语的边界识别

강인호等人（2000）的研究结果表明，在用统计方法处理韩国语名词短语时，优先考虑右边界的识别效果要好于左边界优先。这表明了不同权重的左右边界识别方法会影响到名词短语的识别效果，对用规则的方法处理韩国语具有一定的参考意义。本文分析了韩国语名词性成分（即单个名词及名词短语）[1]在句子中出现的特点，结合其语言学特征，可以发现名词性成分左右边界相邻词的出现有一定的规律。以紧邻词作标记，界定出名词短语和单个名词，分析归纳其左右边界的规律如下。

3.1 左边界的界定

通过语言学规律分析和迭代分析，我们可以看到名词短语的左边界分两种情况，一种是直接以名词或名词短语开头，无左边界。另一种是左边界紧邻词为助词、冠词形转成词尾、连接词尾、副词以及省略号等符号。具体如下：

（1）句子直接以名词或名词短语开头，其左边界紧邻词不存在。

例：신라/NNP 상인/NNG 들/XSN 의/JKG 무역/NNG 루트/NNG 를/JKO 따르/VV 아/EC…（跟随新罗商人们的贸易之路……）

（2）句子以非名词或非名词短语开头，存在左边界紧邻词。左边界紧邻词分为以下 5 种情况。

第一、左边界紧邻词为助词的，按助词类别又分为以下 4 种情况。

①左边界紧邻词为主格助词（JKS）

例：최인호/NNP 가/**JKS** 온/MM 힘/NNG 을/JKO 쏟/VV 아/EC … （崔仁浩倾尽全力……）

②左边界紧邻词为目的格助词（JKO）

例：한국/NNP 영화/NNG 를/**JKO** 영어/NNP 자막/NNG 으로/JKB 관람/NNG 하/XSV ㄹ/ETM 수/NNB 있/VV 는/ETM 외국인/NNG（能看懂韩国电影英文字幕的外国人）

③左边界紧邻词为副词格助词（JKB）

例： 역사학자/NNG 가/JKC 아니/VCN 기/ETN 때문/NNB 에/**JKB**역사/NNG 적/XSN 정설/NNG 을/JKO 내/VV ㄹ/ETM 수/NNB 는/JX 없/VA 습니다/EF ./SF（他不是历史学家，所以无法给出历史的定论。）

④左边界紧邻词为补助词（JX）

[1] 因为单个名词和名词短语在句中的作用和位置一样，所以本文在分析和提取名词短语过程中，把单个名词一并考虑在内，但是对名词短语归类时将其排除在外。

例：음절/NNG 수/NNG 로/JKB 는/**JX** 2/SN 음절/NNG 어휘/NNG 가/JKS 14/SN 만/NR 1765/SN 개/NNB (/SS 32/SN ./SP 2/SN %/SW)/SS 로/JKB 가장/MAG 많/VA 았/EP 고/EC… (14 万 1765 个双音节词汇占总音节数的 32.2%，数量最多……)

第二、左边界紧邻词为冠词形转成词尾（ETM）。

例：사라지/VV ㄴ/**ETM** 백제왕국/NNP 의/JKG 영광/NNG 을/JKO… （消逝的百济王朝的辉煌……）

第三、左边界紧邻词为连接词尾（EC）。

例：그/NP 에게/JKB 역사/NNG 는/JX 생생/XR 하/XSA 게/EC 살/VV 아/**EC** 오늘/NNG 의/JKG 우리/NP 를/JKO 비추/VV 어/EC… （历史栩栩如生，和如今的我们相比……）

第四、左边界紧邻词为副词（MAG/MAJ）。

例：다만/**MAG** 문제/NNG 제기/NNG 를/JKO 하/VV ㄹ/ETM 뿐/NNB 이/VCP 지요/EF ./SF （但是不局限于问题的提出。）

第五、左边界紧邻词为省略号标记（SE）等符号。

例：… /**SE** 한/MM 장/NNB 의/JKG 벽돌/NNG 을/JKO 쌓/VV 아/EC 올리/VV 었/EP 다/EF ./SF （一方砖石堆积而成。）

3.2 右边界的界定

名词短语的右边界分两种情况，一种是直接以名词或名词短语结尾，无右边界或为符号；另一种是紧邻词为助词和指示词。具体情况如下：

（1）句子以名词或名词短语结尾，其右边界紧邻词不存在或符号。

例：신라/NNP 상인/NNG 들/XSN 의/JKG 무역/NNG 루트/NNG 를/JKO 따르/VV 아/EC 장보고/NNP 의/JKG 흔적/NNG 을/JKO 추적/NNG 하/XSV ㄴ/ETM 소설가/NNG 최인호/NNP ./SF（沿循新罗商人贸易之路探寻张寶高足迹的小说家崔仁浩。）

（2）句子以非名词或非名词短语结尾，存在右边界紧邻词。右边界紧邻词分为 2 种情况：

第一、右边界紧邻词为助词，按助词类别分为以下 6 种情况。

①右边界紧邻词为主格助词（JKS）

例：대명사/NNG 중/NNB 에/JKB 는/JX 인칭/NNG 대명사/NNG 가/**JKS**…（代词中人称代词……）

②右边界紧邻词为目的格助词（JKO）

例：장보고/NNP 의/JKG 뱃길/NNG 을/**JKO** 따르/VV 아/EC…（沿着张寶高的水上航线……）

③右边界紧邻词为副词格助词（JKB）

例：사투리/NNG 중/NNB 에/**JKB** 는/JX 경상도/NNP 사투리/NNG 가/JKS… （方言中庆尚道的方言……）

④右边界紧邻词为呼格助词（JKV）

例：나/NP 의/JKG 신부/NNG 여/**JKV** !/SF （我的新媳妇呀！）

⑤右边界紧邻词为补格助词（JKC）

例：병원/NNG 에서/JKB 만나/VV ㄴ/ETM 변호사/NNG 가/**JKC** 되/VV 기/ETN 도/JX 하/VX ㄴ다/EF .（医院里见到的那位律师也行。）

⑥右边界紧邻词为补助词（JX）

例：역사/NNG 는/**JX** 참으로/MAG 흥미/NNG 롭/XSA ㄴ/ETM '/SS 존재/NNG '/SS 이/VCP 거든요/EF ./SF （因为历史真的很有意思。）

第二、右边界紧邻词为指示词（VC），按指示词类别又分为 2 种情况。

①肯定指示词(VCP)

例：…보/VX 는/ETM 거울/NNG 이/**VCP** 다/EF ./SF （……看到的镜子。）

②否定指示词（VCN）

例：이/NP 를/JKO 전혀/MAG 모르/VV 는/ETM 척/NNB 하/VV ㄹ/ETM 수/NNB 없/VA 는/ETM 것/NNB 이/JKS 현실/NNG 아니/**VCN** ㄴ가/EF ./SF （事实是不能装作完全不知道此事。）

3.3 特殊名词短语的边界界定

（1）在韩国语中，“名词+名词派生词尾+肯定指示词+冠形词转成词尾+名词”这类特殊的名词短语的边界判定不完全适用于上面论述的左右边界的判定，需要对其单独判定。

肯定指示词在语料中被标记为 VCP，而 VCP 又是普通名词短语的右边界，如果使用前文的右边界界定规则匹配，此类特殊名词短语

将被漏掉。因此，实验时单独使用正则表达式来直接匹配“名词+名词派生词尾+肯定指示词+冠形词转成词尾+名词”，以避免与提取普通名词短语发生冲突。使用正则表达式对语料匹配时，又出现了两种情况：

其一，直接以名词开头的，无左边界，该特殊短语右边界为 JKB、JKS、JKC、JKO、JKV、JX 、VCP、VCN。

例：보편/NNG 적/XSN 이/VCP ㄴ/ETM 의의/NNG 를/**JKO** 상실/NNG 하/XSV 고/EC… （丧失了普遍的意义……）

其二，非名词开头，左边界为JKB、JKS、JKO、JX、EC、ETM、MAJ、MAG、SE，该特殊短语右边界为JKB、JKS、JKC、JKO、JKV、JX、VCP、VCN。

例 ： 개인/NNG 의/JKG 자기/NNG 권리/NNG 인식/NNG 과/JC 공공/NNG 의/JKG 업무/NNG 에/JKB 대하/VV ㄴ/**ETM** 능동/NNG 적/XSN 이/VCP ㄴ/ETM 참여/NNG 를/**JKO** 강화/NNG 하/XSV 고/EC… （强化个人的自我权利意识和对公共事务的主动参与……）

（2）在韩国语中，还有一些单个名词和名词短语在语料中往往省略了形态标记，无法用上述的形态标记界定左右边界。经过统计分析，这些名词性成分一般均分为 4 类：“（数字|数词|代词）+表示时间的依存名词或名词（如，21 세기/21 世纪, 2013 년/2013 年, 8 월/8 月, 5 일/5 日, 10 시/10 时, 30 분/30 分，10 초/10 秒）；表示时间的名词 （오늘/今天，어제/昨天，내일/明天）；时段|冠词（이, 그）+동안（如，3 년 동안/3 年期间，그 동안/那段时间）；名词|冠词（이, 그）+때（如，고등학교 때/高中时，그 때/那时）。因此，我们可以在依照上述边界规则提取名词性成分的基础上，从含有此类短语的名词性成分中将它们分离出来。

例 ： ① 그/NP동안/NNG 비/XPN 공식/NNG 사무/NNG 총장/NNG 접촉/NNG（那期间非正式事务长官会晤）

② 5/SN 일/NNB 일본/NNP 무역/NNG 진흥회/NNG （5 日日本贸易振兴会）

4.名词短语的自动提取实验及结果分析

4.1 实验流程设计

按照前文归纳的名词性成分左右边界的界定规则，从“21 世纪世宗计划”标注语料库中自动提取出名词短语和单个名词。

具体流程如图 1 所示：

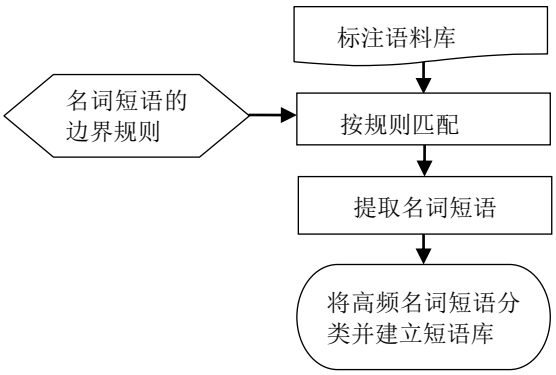

图 1 名词短语分析提取流程图

4.2 自动提取的实现

实验时选取了“21 世纪世宗计划”标注语料库中的 207634 句语料作为训练语料，利用前文给定的左右边界界定规则进行名词短语的自动提取。编程实验的核心思想是将词和词性的标记形式分开，对每个词的标记形式与名词短语的左右邻接词的形式进行匹配，将左右边界界定的中间部分提取出来，放入数组之中，进行后期处理，并将处理后的名词短语放入哈希函数之中，计算各个名词短语的出现次数并按序输出。在 207 634 句训练语料共提取出 410 680 个名词短语，从前 500 位的高频短语中按序选取了 10 个具有代表性的短语，如表 3 所示：

表 3 名词短语自动提取结果表

排序	名词短语	出现次数
1	모든/MM 것/NNB	382
2	그런/MM 것/NNB	321
11	시민/NNG 단체/NNG	220
15	방/NNG 안/NNG	204
22	우리/NP 나라/NNG	177
45	그/NP 의/JKG 얼굴/NNG	117
160	21/SN 세기/NNG	59
380	희수/NNP 와/JC 미니/NNP	39
406	일/NR 년/NNB	37
499	사회/NNG 적/XSN 약자/NNG	28
…	…	…

4.3 提取结果及其分析

对信息提取结果的评价，最为常用的两个指标是召回率和准确率（俞士汶，2003），在信息提取系统中，这两个指标的定义为：

$$召回率=\frac{正确提取的结果数量}{所有正确结果的数量}\times 100\%$$

$$准确率=\frac{正确提取的结果数量}{所有结果数量}\times 100\%$$

为了验证实验结果，选取了训练语料之外的 500 句语料作为测试语料来计算召回率和准确率，具体结果如下：

自动提取结果中共有 801 个短语，其中正确的名词短语为 736 个，人工在测试语料中找到 792 个名词短语。因此，实验结果为：

$$召回率=\frac{736}{792}\times 100\%\approx 92.9293\%;$$

$$准确率=\frac{736}{801}\times 100\%\approx 91.8851\%;$$

实验中出现了未被召回和被错误召回的名词短语，经分析发现了以下两种典型的错误：

其一，名词短语之后紧邻的是名词+되다/하다结尾，而机器无法判断具体匹配到哪个名词，不能满足第二章中的右边界识别而被漏掉。

例：……는/JX 3/SN 년/NNB 여/XSN 간/NNG계속/NNG 되/XSV 었/EP 다/EF ./SF（……持续了三年多）

上例中本该被提出的“三年多（3/SN 년/NNB 여/XSN 간/NNG）”被漏掉。

其二，语料库中标注出现错误。

例：在语料中우리나라(我们的国家)被标注为两种形式：우리나라/NNG 和 우리/NP 나라/NNG。而根据本文的名词短语界定规则，只有后一种标注形式才是正确的。

5.韩国语名词短语分类及库的构建

本节将根据前文提取的名词性成分（单个名词和名词短语），在去除单个名词（包括单个名词、代词以及派生词和合成词）后，对名词短语进行语言学分类，建立不同类型的名词短语库。

5.1 名词短语的分类

根据语言学规律和实验提取结果，我们将训练语料中出现频率较高的名词短语分为 8 个类型，具体如下：

（1）名词|代词+의+名词|名词叠加，该类名词短语是“~的~”型，例如：

그/NP 의/JKG 얼굴/NNG（他的脸）[117] [2]

공공/NNG 의/JKG 이익/NNG（公共利益）[44]

나/NP 의/JKG 마음/NNG 속/NNG（我的内心里）[18]

（2）两个或两个以上名词（代词）混合叠加。例如：

시민/NNG 단체/NNG（市民团体）[220]

우리/NP 사회/NNG（我们的社会）[102]

자원/NNG 봉사/NNG 활동/NNG（志愿活动）[27]

조선/NNP ㅡ/SS 미국/NNP 조약/NNG（朝美协定）[23]

（3）名词|代词+接续助词|特殊的副词+名词|代词。其中，接续助词在该语料中体现为：와（과）、(이)나，特殊的副词体现为：그리고、및。该类名词短语可以归结为“~与~”型。例如：

희수/NNP 와/JC 미니/NNP（熙淑和米尼）[39]

그녀/NP 와/JC 나/NP （她和我）[14]

정부/NNG 나/JC 기업/NNG（政府和企业）[8]

도구/NNG 와/JC 장비/NNG 및/MAG 기술/NNG（道具、装备及技术）[3]

때/NNG 그리고/MAJ 그때/NNG（此时与当时）[2]

（4）冠形词+名词|代词

该类名词短语为“什么样的~”型，冠形词起到修饰中心名词的作用。例如：

모든/MM 것/NNB（所有的东西）[383]

그런/MM 것/NNB（那样的东西）[324]

그/MM 여자/NNG（那个女人）[282]

그/MM 누구/NP（那谁）[50]

（5）数字|数词+名词

该类名词短语多表示日期和时间，例如：

21/SN 세기/NNG（21 世纪）[59]

일/NR 년/NNB（一年）[37]

[2]例子中名词短语后面的数字均为从 207 634 句原始语料中提取的该名词短语的出现次数。

（6）名词|名词叠加+적+名词

表示属性的“적（的）”将两个名词性成分组合在一起，构成“某个性质的~”型名词短语。例如：

사회/NNG 적/XSN 약자/NNG（社会的弱者）[28]

환상/NNG 적/XSN 기준/NNG（虚幻的基准）[19]

중앙/NNG 집중/NNG 적/XSN 편의/NNG（中央集权的益处）[3]

（7）名词+名词派生接尾词+肯定指示词+冠形转成词尾+名词

该类名词短语大部分是“名词+적인+名词”结构，也有部分是“名词+가（或네等其他词）+이란+名词”等其他结构，表示“怎么样的~”。例如：

보편/NNG 적/XSN 이/VCP ㄴ/ETM 의의/NNG（普遍的意义）[7]

길/NNP 가/XSN 이/VCP 란/ETM 놈/NNB（那个叫吉佳的家伙）[5]

금순/NNP 네/XSN 이/VCP 란/ETM 여자/NNG（名字为金顺的女人们） 2

（8）名词|代词+数词+（依存名词）

在表示“几个~”时，韩国语的表达与汉语不同，汉语是“数词+量词+名词”，而韩国语是“名词+数词+（依存名词）”，其中依存名词起到汉语中量词的作用，可省略。例如：

방/NNG 하나/NR（一间房子）[9]

나/NP 하나/NR（我一个人）[9]

벼/NNG 오백/NR 섬/NNB（五百包稻子）[4]

根据上述 8 类的名词短语提取结果，计算各类名词短语的规模，统计结果如图 2 所示。

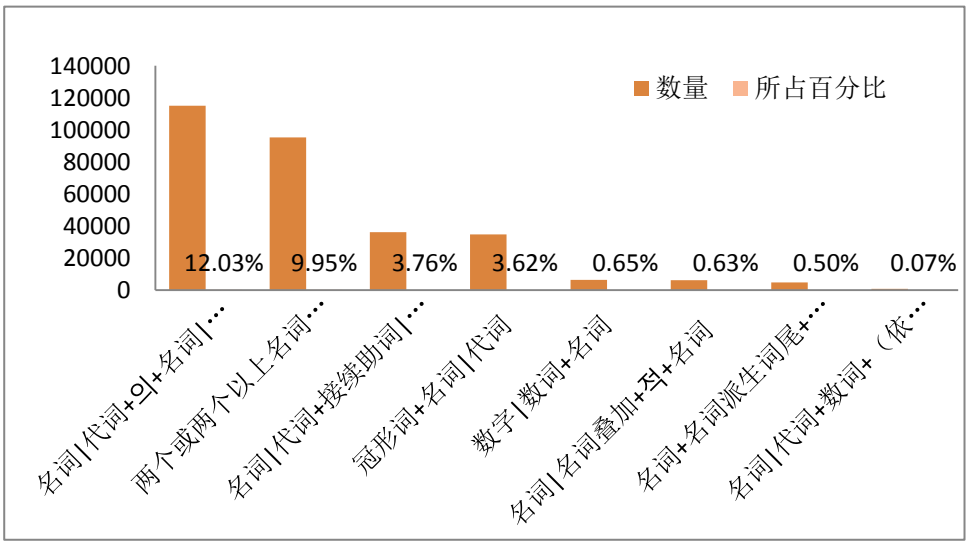

图 2 各类名词短语分布图

5.2 分类构建名词短语库

根据 5.1 中的名词短语分类，根据不同的分类分析出具体的匹配规则，从测试语料库中分别提取出归属于上述 8 类的名词短语，建立 8 个名词短语库，为以后建立韩汉双语短语库打下基础，可以有效地消除一词多义和一词多译现象。

6.结语

本文研究了基于大规模标注语料库的名词短语的结构及分类，通过界定名词短语的左右边界，对名词短语做了统计学上的分析和研究。实验结果的正确率约为 91.89%，召回率约为 92.93%，表明文中的界定规则是可行的。根据实验结果将高频的名词短语分成 8 个类型，并建立不同类型的名词短语库，以便于机器翻译领域里的双语快速匹配、歧义消解、短语对齐等，从而提高机器翻译的效率和精确度，使其更能满足人们的实用需求。

参考文献

［1］ 李基文［韩］. 东亚国语大辞典［D］. 斗山东亚，1997:754.

［2］ 俞士汶. 计算语言学概论［M］. 商务印书馆. 2003:322-333.

［3］ 赵军. 基于转换的汉语基本名词短语识别模型［J］. 中文信息学报. 1998:1-2.

［4］ K. W. Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Test ［A］. Proceedings of the Second Conference on Applied Natural Language Processing. 1988:136-143.

［5］ 강인호,전수영,김길창. 최대 엔트로피 모델을 이용한 한국어 명사구 추출［J］.제 12 회 한글 및 한국어 정보처리 학술대회. 2000.

［6］ 서충원,오종훈,최기선. 어절의 중심어 정보를 이용한 한국어 기반 명사구 인식［J］. 한국정보과학회 언어공학연구회. 2003.

［7］ 양재형. 규칙 기반 학습에 의한 한국어의 기반 명사구 인식［J］. 정보과학회 논문지. 2000: 제 27 권 제 10 호.

［8］ 황영숙,정후중,박소영,곽용재,임해창. 자질집합선택 기반의 기계학습을 통한 한국어 기반구 인식의 성능향상［J］. 정보과학회 논문지. 2002: Vol 29. Number 9.