# 격틀 자동구축과 격틀평가 방법에 관한 연구

최용석, 이주호, 최기선
한국과학기술원 전산학과/전문용어언어공학연구센터
{angelove, mywork, kschoi}@world.kaist.ac.kr

# Study on Automatic Construction and Evaluation method of Caseframe

Yong-Seok Choi, Ju-Ho Lee, Key-Sun Choi
Department of Computer Science KAIST/KORTERM

**요 약**

격틀이란 동사에 대해 필요한 격들과 그 격에 알맞은 단어집합으로 이루어져 있는 것으로 명사와 동사의 의미적 호응을 표현한다. 격틀은 자연언어처리분야에서 주요한 정보로 사용할 수 있다. 의미구분이라든지 번역에서 한국어 생성, 정보검색에서 중요정보 추출 등 잘 구성한 질 높은 격틀은 여러 연구의 질을 높여줄 수 있다. 따라서, 질 좋은 격틀을 구성하기 위한 여러 노력들이 현재 이루어지고 있다.
본 논문에서는 기계 가독형 사전과 말모듬을 이용해서 자동으로 격틀을 구성한다. 자동구성 방법으로 먼저 기계가독형 사전을 이용해서 상위개념 정보를 가지는 분류정보를 구성한다. 말모듬과 사전의 예문들을 형태소 분석한 후에 각각의 예문들을 분류정보를 이용하여 최상위 개념으로 바꾼다. 그리고, 말모듬과 사전의 예문에서 나온 정보들을 통합하므로 해서 자동으로 격틀을 구성한다. 자동으로 격틀을 구성한 후에 수동으로 구성한 격틀과 비교해 본다. 비교하기 위한 평가방법에 대해서 논의한다.

## 1. 머릿글

격틀(case frame)은 명사와 동사의 의미적 호응을 표현한다. 격틀에는 동사가 주어지고, 그 동사에 필요한 격들을 나열하고, 그 격에 쓰일 수 있는 명사류의 단어들을 다시 나열해 준다. 격틀은 이렇게 명사와 동사간의 관계를 설정하므로 질 좋은 격틀을 가지고 있다면, 의미구분으로 구문분석 애매성 해소, 기계번역에서 한국어 생성, 정보추출에서 사용할 수 있는 패턴정보 등으로 활용할 수 있다. 질 높은 격틀은 자연언어처리에 큰 역할을 하므로 현재 좋은 격틀을 만들기 위한 연구들이 행해지고 있다.

이렇게 격틀의 유용성에도 불구하고, 격틀을 구축하려면 동사와 명사의 관계를 파악할 수 있는 고급인력과 구축에 필요한 시간 등 많은 비용이 필요한 일로 구축이 쉽지 않았다. 또한, 어느 동사부터 격틀을 구축해야 하는 것인지도 쉽게 결정할 수 없었으며 많은 비용을 들여서 격틀을 구축하고도 격틀을 제대로 구성했는지 평가하는 방법도 찾기 어려웠다.

이에 본 논문[1]에서는 비교적 적은 비용을 들여서 기계가독형 사전과 말모듬(corpus)을 이용해서 자동으로 격틀을 구축한다. 자동으로 격틀을 구축하면서 격틀에 필요한 최소 요소를 제시하고, 격틀의 규모에 대한 평가를 내리려고 한다.

사전에서 추출한 격틀 정보는 기본적으로 명확한 예제와 보편적인 사용자들이 사용하는 말에서 추출했다는 가정을 할 수 있다. 말모듬에서 추출한 격틀 정보는 현실에서 쓰이고 있는 말을 반영할 뿐만 아니라, 그 빈도 정보를 이용해서 기본적으로 필요한 정보들은 빈도가 높은 정보라는 것을 말할 수 있다.

이렇게 2가지 방법으로 구축한 격틀을 통합해야 한다. 명확한 정보인 사전에서 뽑은 격틀과 현실 반영 정보인 말모듬에서 뽑은 격틀은 서로 약간 다른 성질을 갖는다. 2가지를 통합하기 위해서 하나의 형식으로 바꿔주는 작업이 필요하다. 분류정보(taxonomy)를 이용해서 예로 올라와 있는 명사들을 상위정보로 바꿔 준 후 격틀을 통합한다. 이런 방법으로 격틀을 자

1) 이 연구는 과학기술부와 한국과학기술평가원의 "핵심소프트웨어기술개발" 중 "대용량 국어정보 심층 처리 및 품질 관리 기술 개발" 과제로서 지원받았음.

동으로 구축할 수 있으며, 이 격틀은 격틀의 출발선을 제시할 수 있다.

본 논문에서는 격틀에 필요한 최소 동사수, 최소 예제수의 기준을 제시할 격틀을 자동으로 구축한 후에, 이것을 근거로 격틀에 대해서 동사수와 예제수를 기본 축으로 하는 벡터를 만들어 그 길이로 평가방법을 제시한다. 자동구축 격틀의 평가를 위해서 다른 수동으로 만든 격틀과 비교를 할 것이다.

이 논문의 구성은 다음과 같다. 2장에서는 격틀에 관해서 살펴보고 논문에서 구축할 격틀의 형태를 제시한다. 3장에서는 말모듬으로부터 격틀 추출하는 방법에 대해서 다루고, 4장에서는 기계가독형 사전으로부터 격틀을 추출하는 방법에 대해서 다룬다. 5장에서는 2가지 격틀을 통합하는 방법에 대해서 다루고, 6장에서 평가 방법과 기준을 제시하며, 다른 격틀과 본 논문에서 구축한 격틀을 비교한다. 마지막으로 7장에서 결론을 맺는다.

## 2. 격틀

격틀이 포함해야 할 요소는 동사와 명사의 호응관계에 관한 정보이다. 이 정보를 구성하는 방법으로 크게 자동으로 구성하는 방법과 수동으로 구성하는 방법이 있다. 또한 정보 표현 방법으로 사람을 위한 표현이냐, 기계처리를 위한 표현이냐로 나눌 수도 있다. 이런 구성 방법들에 따라서 격틀은 다양한 형태로 나타나게 된다. 다양한 형태의 격틀로 인해서 격틀간의 정보교환이 쉽지 않고, 격틀 비교도 쉽지 않다. 이 장에서는 2개의 격틀 형태를 살펴보고, 본 논문에서 사용할 격틀 형태를 정의한다.

### 2.1. 과기원 격틀[송영빈 1999]

과기원 전문용어언어공학연구센터에서 현재 만들고 있는 격틀이다. 수동으로 직접 구축하는 격틀로 기본적으로 기계사용을 목적으로 한다고 볼 수 있다. 그 형태는 그림 1과 같다.

형식을 살펴보면 '매다6'은 '매다' 동사의 6번째 의미라는 뜻으로 그 뜻을 먼저 기록한다. 이 부분은 사람을 위한 부분이라고 볼 수 있다. 다음 N0으로 주체를 표시하고 N1이 가지는 조사와 거기에 올 개념을 정리하며 그 개념의 하위 개념으로 실제 예를 가져다가 써 놓는다. 또한 다른 조사에 대해서 'N2에'라고 표시하고 거기에 따르는 개념과 실제 예들이 들어가 있다.

이러한 구성이 기계를 위해서 만들어 졌다고 생각하는 이유는 사람이 이해하기 좋은 형태가 아닌 단순한 형태로 만들어졌다는 것이다. 의미를 풀어놓은 부분도 사람이 수동으로 격틀을 구축할 때 도움을 받기 위해서 기록해 놓았다고 볼 수 있을 정도이다. 수동으로 구축했기 때문에 다양한 예와 비교적 명확한 개념들로 표현하고 있다. 그러나, 수동으로 구축했기 때문에 기계에서 사용하려면, 각 시스템에 맞게 다시 한번 가공의 필요가 있다.

### 2.2. 신중호 격틀[신중호 1999]

신중호 격틀은 클러스터링 기법을 사용해서 동사를 분류하기 위해서 만들어졌다. 자동으로 구축했으며 동사를 분류하는 기계적 작업에 적합하도록 구성했다. 격틀을 사용해서 동사를 분류하고 그 분류를 이용해서 정확한 언어분석을 할 수 있도록 하려는 목적으로 만든 격틀이다. 아래는 그 일부분이다.

가결시키 (로 jca)
가결시키 (로 jca) (를 jc)
가결시키 (로 jca) (으로써 jca)
가결시키 (를 jc)
가결시키 (를 jc) (로 jca)
가결시키 (를 jc) (으로 jca)

여기에 격틀로서 필요한 명사정보들은 말모듬을 사용해서 가중치를 주고 뽑았다. 격에 관한 정보는 조사의 형태정보로만 할 경우에 말모듬에서 자료가 부족하게 되므로 조사와 형태정보를 묶어서 사용하였다. 조사의 문법적 성격이 같을 경우 같이 취급하도록 정규화한 것이다.

신중호 격틀은 기계에 사용하기 적합하도록 가공되었으며, 특히 말모듬의 자료부족문제(data sparseness problem)를 회피하기 위해서 조사의 형태정보 부분을 정규화한 것이 특이한 점이다. 또한 명사와의 호응관계는 말모듬에서 가중치를 뽑아서 따로 관리하는 것도 실제 시스템에 적합하도록 격틀을

| 매다6 | 의류부분을착용함 | N0 | N0 | N1을/를 | N1 | N2에 | N2 |
|---|---|---|---|---|---|---|---|
| | | 사람 | | 의류 | 허리띠 | 의류 | 도포 |
| | | | | | 넥타이 | 동물(부분) | 목 |
| | | | | | 옷고름 | | 허리 |
| | | | | | 대님 | | 머리 |
| | | | | | 댕기 | | |
| | | | | | 두건 | | |
| | | | | | 스카프 | | |
| | | | | | 망건 | | |

그림 1 격틀의 예[송영빈 1999]

구성한 것이라 볼 수 있다.

### 2.3. 격틀 형태 정의

앞에서 본 바와 같이 사용목적에 따라 다양한 형태의 격틀이 있음을 알 수 있다. [홍재성 1997]과 같이 사람이 보고 이해하기 쉬운 형태로 가능한 격틀 정보들을 나열해 놓은 책도 존재한다. 이런 형태는 사람이 그 동사와 명사간의 호응 관계를 파악하기는 쉬우나 기계에 적합하도록 하기 위해서는 구조화라는 단계를 거쳐야 한다. 본 논문에서는 이런 다양한 형태의 격틀을 통합하고, 평가하기 위해서 아래와 같은 형태로 격틀을 정의한다.

| | | |
|---|---|---|
| 가르 | jco jco[2] | 빗장 식물 |
| 가르치 | jcm jco jcc | 너 행동 곳 |
| 가르치 | jco | 국어 |
| 가르치 | jco | 절차탁마 |
| 가르치 | jco | 하나 |
| 가르치 | jco jcm jco jcc | 곳 너 행동 곳 |
| 가름하 | jco | 승패 |
| 가리 | jca jco | 수건 장식물 |

동사, 격을 이루는 조사의 형태정보[이공주 1996], 그리고 그 격에 나타난 명사들을 나열하는 식으로 격틀을 표현한다. 의미구분(sense disambiguation)에 필요한 다의어적인 동사에 대한 구분은 하지 않는다. 이런 정보들은 말모듬과 기계가독형 사전으로부터 각각 구축할 것이고, 그 후에 상위 정보로 바꾸어서 통합할 것이다. 통합하기 위해서 추출 가능한 몇 가지 정보를 포기하고, 공통 정보를 뽑아놓은 형태로 격틀형태를 정의했다.

## 3. 말모듬으로부터 격틀 추출

문화체육부와 과학기술처의 연구과제 국어정보처리기반구축과 STEP2000에서 구축된 과기원 말모듬 중에서 97년에 만든 품사부착 말모듬을 격틀을 추출하는 데 사용했다. 이 품사부착 말모듬은 1160만 어절 수준이다.

추출 대상은 동사로 지시동사와 일반동사로 품사부착한 것을 뽑아냈다. 뽑아낸 동사 앞에서 격조사를 찾아 뽑아내고 격조사와 함께 나온 명사 역시 뽑아낸다.

총 1만 8609개의 서로 다른 동사를 추출했다. 추출한 동사에는 '시시콜콜하다'같은 형용사에 동사 품사가 부착되어 있는 것과, '희번득이다'와 같이 사전에는 등록되지 않은 빈도수가 낮은 동사도 실제 말모듬에서 나왔다. 또한 맞춤법 오류 등의 문제도 사전에 등록되어 있지 않은 동사를 등장시킨다. 그러나, 실제로 글쓴이가 의도하고 사용했을지도 모르는 동사를 사전에 없다고 제외하지는 않았다.

총 195만 4238개의 격틀 후보들을 뽑았다. 동사 개수만 대상으로 했을 때는 '하다'(7.11%), '있다'(4.65%), '없다'(3.51%), '되다'(1.96%) 순으로 발생빈도가 높았다. 격까지 대상으로 하면 '목적격+하다'(1.53%), '주격+있다'(1.13%) 순이었으며, 사용예로 명사까지 고려하면 '수+주격+없다'(0.25%)가 빈도가 가장 높았다.

빈도정보는 여러 자연언어처리 분야에서 유용하게 쓰일 수 있는 정보지만 본 논문에서는 통합을 위해서 제거했다. 또한 동사가 문장 처음에 수식어로 나와서, 격틀을 구성하는 격정보와 명사정보를 추출할 수 없는 경우도 있는데, 추출할 수 없다는 정보 역시 그 동사의 특성정보로 사용할 수 있지만 통합을 위해서 그런 격틀 후보는 제거했다.

## 4. 기계가독형 사전으로부터 격틀 추출

한글학회에서 나온 '우리말 큰 사전'을 사용해서 격틀을 추출했다. '우리말 큰 사전'은 북한말, 옛말, 이두 등을 포함해서 40만 2305개의 항목을 가지고 있다. 그 중에서 자동사, 타동사, 자타동사를 대상으로 했다. 대상으로 한 동사수는 4만 5703개이다.

### 4.1. 사전 구조화

원본 사전에는 출판을 위한 기호를 쓰고 있으며 나열식으로 되어 있으나 기계에서 본격적으로 다루기 위해서는 그 형태를 변화시켜야 원하는 정보를 자동으로 뽑을 수 있다. 형태를 구조적으로 바꿔 주고 그 구조를 읽어들여서 원하는 일을 한다. 구조화하기 위해서 먼저 사전의 구조를 자세히 파악하고 다음의 세 단계의 일을 해야 한다.

첫 번째 단계에서는 원래 사전에서의 올림말에서 특수문자(':', '-' 따위)를 제거하여 새로운 올림말을 만든다. 이렇게 만든 새로운 올림말과 단어번호가 사전 항목의 키가 된다. 또한 구조화 작업을 하기 전에 물결표시('~')로 나온 부분은 올림말로 치환하여 원래의 정보를 살린다.

두 번째 단계에서는 사전에서 쓰이는 특수기호를 가지고서 새로운 태그를 붙인다. 특수기호만으로 어떤 태그를 붙일지 불명확할 때는 경험규칙을 사용하여 태그를 붙이게 된다. 위의 예에서 보면 올림말부분에서 '['가 나오면 이것은 올림말이 끝나고 한자나 어원이 시작됨을 뜻하는 것 등이 경험 규칙이 될 수 있다.

세 번째 단계에서는 구조적인 부분까지 고려하여 전체적으로 표준범용표기언어[3](SGML)[ISO 8879]형태로 구조화시킨

2) 조사기호[이공주 1996] - jcs: 주격, jco: 목적격, jcc: 보격, jcm: 관형격, jcv:호격, jca: 부사격, jcj: 접속격, jct: 공동격, jcr: 인용격

3. 문서작성언어표준 이라고도 한다.

다.

아래는 기계가독형 사전에 나오는 형태이다.

\가:관-스럽다[可觀--]. M ㅂㅋ 꽤 볼 만하다. ¶동물원의 돌고래가 배구하는 재주는 참으로 가관스러운 일이었다.

위 형태를 아래와 같이 구조화 시켰다.

<단어>
<올림말>가관스럽다</올림말>
<단어번호>1</단어번호>
<사전표기>가:관-스럽다</사전표기>
<한자,어원>可觀--</한자,어원>
<불규칙활용>ㅂ<불규칙활용>
<의미><의미번호>1</의미번호>
<의미풀이>꽤 볼만하다.</의미풀이>
<예문>동물원의 돌고래가 배구하는 재주는 참으로 가관스러운 일이었다.</예문>
</의미></단어>

### 4.2. 격틀 구축

격틀 구축은 그림 2 같은 순서로 이루어진다. 사전에서 대상 동사가 예문을 가지고 있다면, 그 예문을 형태소 분석한 후 품사 부착을 한다. 품사부착 후에는 말모듬에서와 마찬가지로 격조사를 보고 격조사와 명사를 뽑아내서 격틀을 만든다.

예문을 가지고 있는 동사수는 8275개였다. 그 중에서 격을 추출할 수 있는 동사수는 3899개였다. 어떤 특정동사의 예문에는 여러 동사가 같이 나타난다. 올림말 동사의 예문에는 올림말이 아닌 동사도 등장하지만, 명확한 추출을 위하여 올림말 동사만의 격정보를 추출했다. 격정보를 추출하기 위해서 형태소 분석기를 사용했고, 간단히 만든 품사부착기[신중호 1994]를 사용해서 예문에 품사를 부여했다. 사전의 특성상 일반적으로 예문이 단순 명료하므로 구문해석기 없이도 좋은 격정보를 추출할 수 있었다.

한 동사가 여러 가지 뜻을 가질 수 있는데, 이런 동사는 다의어에 속한다[조정미 1998]. 사전에는 이런 다의어 정보를 상세히 기록해 놓고 있으며, 이 정보는 의미구분에 유용하게 사용할 수 있다. 본 논문에서는 통합을 위해서 다의어의 의미정보를 고려하지 않고, 모두 같은 동사로 보았다.

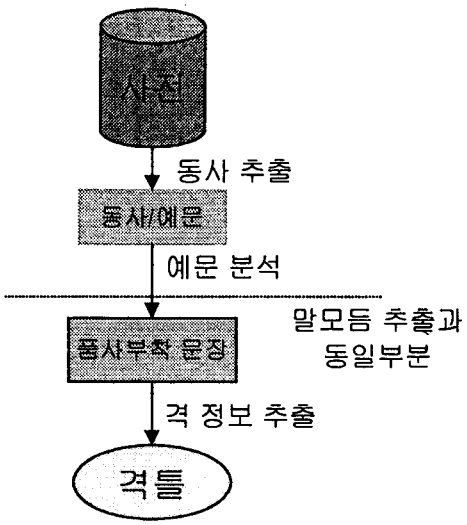


그림 2 기계가독형 사전으로부터 격틀 추출

## 5. 격틀 통합

본 장에서는 앞에서 만든 2개의 격틀을 통합하는 방법에 대해서 다룬다. '우리말 큰 사전'으로부터 분류정보(taxonomy)를 구축한 후 분류정보에서 상위개념(genus term)을 찾아내어 명사를 대치한 후에 통합했다. 명사를 예 상태로 그대로 놓아 둘 경우, 너무 다양한 예들이 나타나고 각 격틀의 성격에 따라서 예의 개념 수준이 달라지게 된다. 예로 나타난 명사들을 최상위 개념으로 바꿔줘서 예들의 성격을 통일한 후에 통합한다.

### 5.1. 분류정보 구축

기계작업을 위해서 앞에서 구조화시킨 사전을 이용한다. 사전의 단어 정의부를 이용해서 단어의 상위개념을 찾고 분류정보를 구축하는 것이다[조평옥 1997]. 한국어의 특성으로 대개 단어 정의부의 뒷부분에 상위 개념이 나온다. 이를 이용하여

| | |
|---|---|
| 경험법칙들<br>(다음과 같은 말이 단어 정의의 뒷부분에 나오면 A가 상위 개념이 된다.) | A들, A의 (한)가지, A의 (한)갈래, A의 궁중말, A의 낮은말 A의 낮춤말, A의 높임말, A의 뜻, A의 무당말, A의 변말, A의 변한말, A의 (한)부문, A의 (한)부분, A의 (한)분야, A의 비유, A의 소경말, A의 속된말, A의 심마니말, A의 어린이말, A의 원말, A의 (한)이름, A의 (한)종류, A의 하나, A의 한가지, A의 힘줌말 |
| 상위개념에서 제외한 단어들 | 감, 것, 냄, 말, 븜, 일, 전, 줌, 한, 함 |

표 1 상위개념을 위한 경험법칙

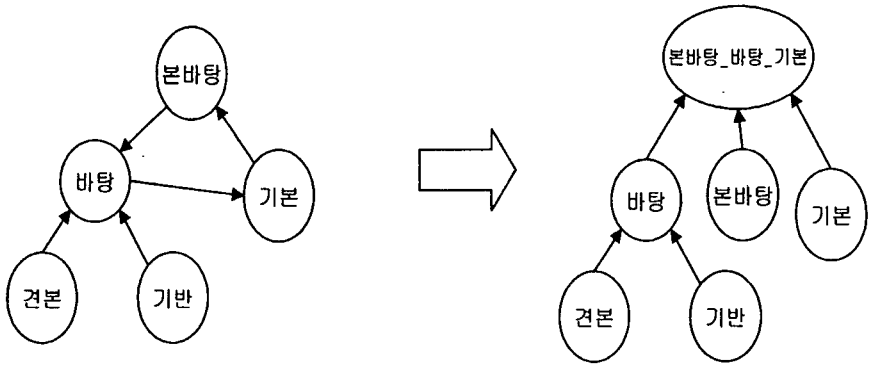


그림 3 순환현상이 나타나는 노드의 처리

사전의 내용을 이해하거나 형태소 분석 또는 구문분석을 하는 노력을 하지 않고도 상위 개념을 찾을 수 있다고 가정한다. 여기에서는 명사에 대해서만 단어 분류정보를 구축하도록 한다. 명사의 경우에는 명사가 상위개념이 된다.

같은 올림말에 대해서 여러 개의 단어가 사전에 포함 될 수 있는데 의미가 많은 단어가 상위 개념으로 잘 쓰인다고 보고, 가장 의미가 많은 단어를 선택한다. 또, 한 단어 내에서도 정의부가 여러 개인 것이 있다. 이런 경우에는 '우리말 큰사전'의 경우에 일반적으로 자주 쓰이는 의미가 앞에 나오는 경우가 대부분이므로 먼저 나오는 의미를 우선적으로 선택하여 상위 개념을 찾는다. 결과의 정확도를 높이기 위해서 여기에 몇 가지 경험 법칙을 사용한다. 또한, 이 때 이 경험법칙을 사용할 때 자주 나오는 옳지 않은 상위 개념은 제외시킨다. '봄', '감', '냄' 같은 동사가 명사형 전성어미로 끝나는 단어들이 제외되는 상위 개념들의 예이다. 모든 명사에 대하여 그것의 상위 개념을 뽑고 나서 상위 개념으로 많이 나타나는 단어들에 대해서 제대로 나왔는지, 제대로 나오지 않았다면 제대로 나오기 위한 새로운 경험 법칙을 추가하여 다시 실험한다.

경험법칙과 상위개념에서 제외한 단어들은 표 1과 같다. 경험법칙은 [조평옥 1997]을 따랐으며, 본 논문의 통합 목적에 맞게 약간 수정을 가했다.

이런 규칙으로 각 명사들을 상위개념과 쌍으로 연결할 수 있다. 이것을 가지고 트리를 구성하게 된다. 이 때, 어쩔 수 없이 상위노드가 하위노드를 참조하는 순환현상(cycle)이 발생한다. 순환현상이 나타나는 곳을 찾아서 대표개념을 설정하는 일이 필요하다. 그런 경우에 먼저 순환현상이 나타나는 노드들을 찾아내고 그것을 전부 합친 새로운 노드를 만들고, 기존의 노드의 상위 개념으로 새로 만든 노드를 가리키게 한다. 이것을 그림으로 나타내면 그림 3과 같다.

이런 식으로 트리를 구성할 때, 상위개념으로 갈수록 추상적인 개념이 많이 나오고 그러다가 전혀 의미가 다른 쪽으로 노드가 연결될 수 있으므로 이럴 경우 어쩔 수 없이 사람이 개입해서 직접 검증하고 수정해야 한다.

사전에 나오는 전체 단어의 수는 40만 2305개이고, 그 중에

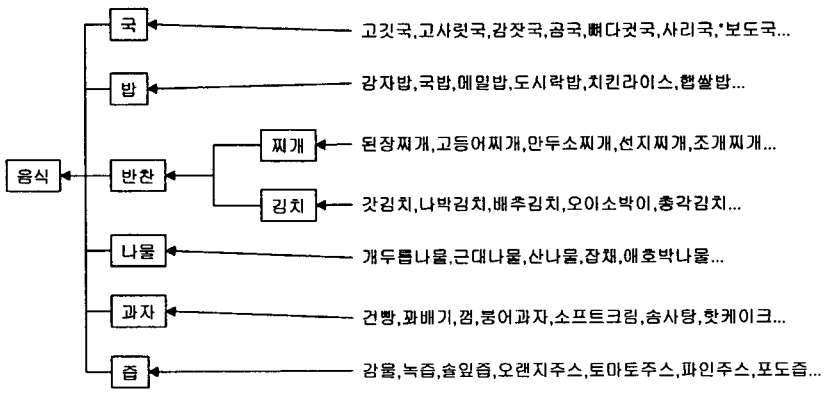


그림 4 '음식'에 관한 분류정보

서 명사의 수는 30만 4788개이다. 여기에서 중복된 올림말을 제외시키고, 상위개념을 찾을 수 있는 13만 3385개의 명사를 대상으로 분류정보(taxonomy)를 구축했다. 구축한 트리의 수는 3795개이고, 순환현상이 나타나 묶은 노드 수는 33개이다. 가장 깊은 트리는 깊이가 16이다. 그림 4는 만들어낸 분류정보 중에서 '음식'에 관한 부분의 일부이다.

### 5.2. 명사를 상위 개념으로 치환

격틀의 형태는 다양하며, 말모듬과 사전으로부터 만든 격틀의 예들도 다양한 형태로 나타난다. 사전의 예문에 나오는 명사는 단순한 몇 가지로 이루어져 있지만, 말모듬에서 추출한 격틀에는 다양한 명사들이 출현한다. 이들은 비교하기 위해서 명사들의 개념수준을 맞추어 주는 정규화가 필요하다. 본 논문에서는 정규화할 정보로써 앞에서 구축한 분류정보를 이용한다. 격틀에 나타난 예들을 분류정보를 이용해서 최상위 개념으로 치환한다.

"학교에 가다" 같은 문장에서 다음과 같은 격틀 정보를 추출할 수 있다.

가 jca 학교

이 격틀에서 '학교'를 최상위 개념으로 치환해 준다. 그러면 아래와 같이 변한다.

가 jca 데

'학교'의 최상위 개념은 분류정보에서 '데'로 되어 있다. 이런 식으로 치환 가능한 예들을 모두 최상위 개념으로 치환한다. 그러나, '서울', '부산' 이나 많은 명사들이 분류정보에 포함되어 있지 않았다. 포함되어 있지 않을 경우에는 그냥 그 예를 써서 격틀을 표현한다.

앞에서 구축한 2개의 격틀의 예들을 최상위 개념으로 치환한 후에 격틀을 통합한다.

## 6. 격틀 평가 방법과 기준

앞에서 2개의 격틀을 통합했다. 본장에서는 2개의 통합결과와 격틀 평가 방법에 대해서 다룬다.

### 6.1. 격틀 통합 결과

각각 구축한 격틀의 특성은 표 2와 같다. 사전에서 나오는 총 동사수는 4만 5703개이다. 이 중에서 예문을 가지고 있는 동사는 8275개인데, 그 예문을 분석해서 격틀을 추출할 수 있는 동사는 3899개이다. 격틀이 나왔을 때 예로 나오는 명사부분을 무시했을 때 격틀 수는 7160개였고, 예에 나오는 명사정보가 다르면 다른 격틀로 보고 그 수를 따지면 1만 493개이다. 예를 분류정보를 이용해서 치환한 후에 격틀수는 9713개가 나온다.

마찬가지 방법으로 품사부착 말모듬에서 추출한 격틀에 대해서도 수치정보를 뽑아낸다. 양쪽에서 뽑은 격틀이 수치정보의 대소는 같은 형태로 나타나고 있음을 볼 수 있다.

2개의 격틀을 통합한 후에 구할 수 있는 정보는 격을 가진 동사수와 예를 치환한 격틀수이다. 통합하기 위해서 예를 최상위 정보로 치환하기 때문에 다른 정보는 의미가 없다. 통합한 후에 격을 가진 동사수는 1만 5661개이고, 예를 치환한 격틀수는 59만 796개이다. 서로의 격틀 정보가 크게 중복되지 않고 상호 보완적인 관계로 나타남을 알 수 있다.

| 추출 대상 | 기계가독형사전 | 품사부착 말모듬 |
|---|---|---|
| 동사수 | 4만 5703 | 1만 8609 |
| 격을 가진 동사수 | 3899 | 1만 3552 |
| 격틀수(격) | 7160 | 22만 9962 |
| 격틀수(예) | 1만 493 | 75만 6500 |
| 격틀수(예 치환) | 9713 | 58만 3210 |

표 2 격틀 추출 결과

### 6.2. 격틀 평가 방법

격틀은 개발의 목적에 따라서 다양한 형태로 나타나고 있다. 따라서, 격틀의 비교와 평가를 쉽게 할 수 없다. 격틀을 구축할 때, 목적에 맞는 정보를 중점적으로 구축하기 때문에 정보의 관점과 질이 다르기 때문에 평가가 쉽지 않다. 본 논문에서는 격틀에 공통적으로 나타나는 요소를 추출해서 그 요소들을 비교함으로써 격틀을 평가하려 한다.

평가하기 위한 가장 기본적인 요소는 다루는 동사수이다. 격틀이 어느 정도의 동사를 다루고 있는지를 살펴봐야 한다. 또한, 각 동사에 대해서 어느 정도의 예제를 가지고 있는가도 격틀의 질을 결정하는 중요요소이다. 물론 예제를 가지고 있는 정도가 격틀의 목적에 따라서 상당히 다르기 때문에 앞에서 이용한 분류정보로 예제를 최상위 개념으로 바꾸어서 정규화했다. 평가에 필요한 요소들을 이용해서 벡터를 만들어 준다. 본 논문에서는 두 요소를 사용해서 격틀 평가 벡터를 만들어 준다. 만약, 평가요소가 더 있다면 벡터공간의 차원을 늘려서 평가벡터를 만든다.

$\overrightarrow{\text{격틀}}$ = (동사수, 동사당 예제수)

이렇게 벡터로 표현한 후 격틀의 평가 척도는 격틀 벡터의 길이가 된다. 이런 경우 동사수는 만 단위이고, 동사당 예제수는 십 단위이기 때문에 단순히 길이를 구하면 동사수의 영향이 너무 크다. 따라서, 두 요소를 정규화 시킨 후에 길이를 비교해야 한다. 정규화 시킬 값을 정하는 방법으로 동사수는 5만

개를 최대로 보고, 동사당 예제수는 50으로 본다. 물론 이를 넘을 수 있는 값이 나올 수 있지만, 기본적으로 우리 세계에서 나올 수 있는 동사수는 사전에 나오는 동사수로 할 수 있고, 동사당 예제수는 말모듬에서 가능한 예제수 정도일 것으로 예측하기 때문이다. 사전에 나오는 동사수는 4만 5703개이고, 말모듬에서 추출한 동사당 예제수는 43.03개이다.

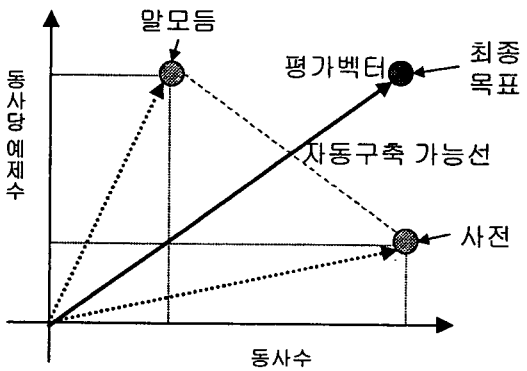


그림 5 평가 좌표 평면

그림 5는 평가할 좌표평면의 예를 보여준다. 사전에서 뽑은 격틀벡터는 동사수가 많이 나올 것이고, 말모듬에서 뽑은 격틀벡터는 예제수가 강조된 위치에 나타날 것이다. 두 격틀을 통합한 격틀벡터는 그 연장선상 부근에 나타날 것으로 예상한다. 우리가 최종 목표로 하는 격틀 벡터는 사전에서 나오는 동사수만큼의 동사를 가지고, 말모듬에서 나오는 동사당 예제수만큼의 예제수를 가질 것이다. 벡터의 길이가 길어지면 길어질수록 좋은 격틀이라 평가할 수 있다.

### 6.3. 격틀 평가 결과

앞 절에서 기술한 평가방법으로 우리가 구축한 격틀을 평가한다.

$\overrightarrow{말모듬}$=(13552/50000, 583210/13552/50)=(0.27, 0.86)

$\overrightarrow{사전}$=(45703/50000, 9713/3899/50)=(0.91, 0.05)

$\overrightarrow{통합}$=(15661/50000, 590796/15661/50)=(0.31, 0.75)

위와 같이 각각의 벡터 값을 구할 수 있다. 사전에 나오는 동사의 항목수는 4만 5703개이지만 실제로 그 중에서 예문이 있는 동사는 8275개이고, 격을 추출할 수 있는 동사수는 3899개에 불구하다. 4만여개의 동사는 격정보를 구축하는 데 별 도움을 주지 못했다. 예문은 있는데 격정보를 추출하지 못한 4천여개 정도의 동사는 문장의 앞부분에서 수식하는 역할로 많이 사용한다는 정보는 얻을 수 있지만, 예문도 없는 동사는 격에 관한 어떤 정보도 주지 못 한다고 볼 수 있다. 하지만, 가능한 동사수를 알려준다는 의미와 초기 벡터공간을 정의하기 위해서 사전에서 추출한 동사수를 4만 5703개로 했다. 말모듬에서도 마찬가지 현상이 생기지만 격정보에 도움을 주지 못한 동사는 제외하고 동사수를 계산했다. 일반적인 격틀은 격을 가지지 않은 동사는 수록하지 않는다.

통합한 격틀에서는 격을 가지지 않은 동사까지 통합할 수는 없었다. 사전 벡터는 특수하게 격을 가지지 않은 동사수도 계산했지만, 통합격틀에는 격이 없는 동사는 존재하지 않는다. 만약 사전과 같은 비율로 격이 없는 동사를 통합격틀에 포함시키면, 자동구축 가능선상에 통합벡터가 있을 것이다. 계산한 벡터의 길이는 아래와 같다.

$|\overrightarrow{말모듬}|$= 0.90

$|\overrightarrow{사전}|$=0.91

$|\overrightarrow{통합}|$=0.82

여기에 격틀의 목표로 삼는 벡터는 말모듬의 예제수와 사전의 동사수를 가지는 것이다. 그 벡터의 길이는 아래와 같다.

$|\overrightarrow{목표}|$ = | (0.91, 0.86) | = 1.26

네 가지 벡터를 좌표평면에 나타내면 그림 6과 같다.

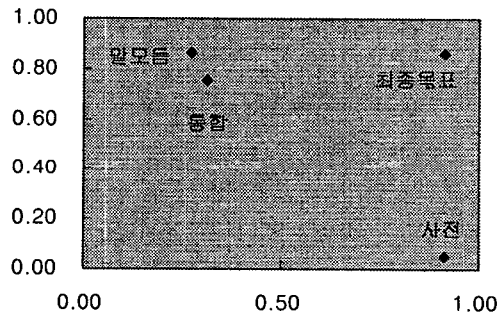


그림 6 좌표평면상의 실제 격틀벡터

| 추출 대상 | 의미구분 | 의미구분 무시 |
|---|---|---|
| 동사수 | 2706 | 460 |
| 격틀수(예 치환) | 1만 2952 | 1만 1306 |
| 평가벡터 길이 | 0.11 | 0.49 |

표 3 [송영빈 1999] 분석 결과

[송영빈 1999]의 격틀을 평가해 보았다. 이 격틀은 의미구분을 할 수 있도록 한 가지 동사에 대해서 번호를 붙여서 다양한 의미로 취급하고 그 사용예를 달아 놓았다. 의미구분을 한 동사를 다른 동사로 취급한 경우와 의미구분을 무시하고 한 가지 동사로 취급한 경우 나오는 결과는 표 3과 같다.

본 평가방법에서는 의미구분을 인정하지 않는다. 인정하고 결과를 구해보니 평가벡터의 길이가 작은 값이 나왔다. 하지만, 앞에서 제시한 평가방법에 준해서 의미 구분을 무시하고 같은 동사로 다루었을 경우 동사당 예제수가 많아져서 더 좋은 평가벡터의 길이를 보이고 있다. 평가한 격틀이 의미구분을 중시해서 수동으로 다양한 예들을 입력했기 때문에 이 같은 결과가 나온 것이다.

[송영빈 1999] 격틀이 아직 완성한 격틀이 아니기 때문에 단순히 비교할 수는 없으나, 평가벡터로 평가하며 자동으로 구축한 통합 격틀이 더 나은 결과를 보여주고 있다. 따라서, 격틀을 구축할 때 통합 격틀을 기준으로 시작한다면 좋은 격틀을 구축할 수 있을 것이다.

## 7. 맺음글

본 논문에서는 격틀을 자동구축하고, 격틀을 평가하는 방법을 제시했다. 자동 구축한 격틀을 격틀의 기준선으로 삼고 격틀이 목표하는 바를 수치화해서 격틀 평가의 기준을 만들었다. 자동으로 구축한 격틀의 평가결과는 0.82였으며, 목표로 하는 격틀의 평가치는 1.26이었다. 자동구축으로도 괜찮은 격틀을 만들 수 있다는 것을 보였고, 평가방법을 제시함으로 해서 서로 다른 격틀을 비교할 수 있도록 했다. 평가에는 동사수와 동사당 예제수를 사용한 2차원 벡터를 사용했다. 평가에 더 많은 요소가 필요하다면 벡터의 차원을 늘리면 된다.

말모듬에서 추출한 격틀은 실제 생활에서 사용하는 다양한 예들을 보여줬으며, 사전에서 추출한 격틀은 간단 명료한 정보를 줌으로써 격틀의 신뢰성을 높였다. 두 개의 격틀을 통합함으로써 서로의 장점을 함께 가지는 격틀을 만들었다.

평가를 위해 두 격틀을 통합하면서 각각의 특성이 되는 정보를 제거했다. 말모듬에서 추출한 격틀은 빈도 정보를 제거했으나, 이 정보를 이용하면 격틀을 구축할 때 빈도수가 높은 격부터 구축할 수 있다. 사전으로부터 나오는 다의어적인 특성도 무시했으나, 이를 모아서 격틀로 구성하면 의미구분에 유용하게 쓸 수 있다.

향후 과제로는 말모듬에서 격틀 추출할 때 구문해석기를 써서 더 확실한 격틀을 만들 수 있도록 하는 것이 있다. 현재 구축한 격틀에서 무시한 다의어적인 동사에 대한 구분을 어떤 식으로 격틀에 표현할 것인가도 중요 연구과제이다. 다의어 정보는 격틀 응용에 있어서 중요한 정보이다. 또한, 예들을 정규화하기 위해서 사용한 분류정보는 시소러스, 온톨로지 등으로 확장해 좀 더 명확한 평가 기준이 되게 할 수 있는 연구도 필요하다. 그리고, 사전에서 조합형 문자 처리를 하면 조금 더 많은 동사들을 다룰 수 있을 것이다.

## 감사의 글

형태소 분석기와 품사 부착기를 함께 구현해준 이운재, 김선배, 김길연, 서충원에게 감사의 마음을 표한다.

## 참고문헌

[송영빈 1999] 송영빈, 채영숙, 최기선, "동사의 애매성 해소를 위한 구문의미사전 구축", 한글 및 한국어 정보처리 학술대회, 전주, 1999년 10월 8-9일

[신중호 1994] 신중호, 한영석, 박영찬, 최기선, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사태깅", 한글 및 한국어 정보처리 학술대회, 389-394쪽, 대전, 1994년 11월 18-19일

[신중호 1999] 신중호, 박혁로, 한선화, "클러스터링 기법을 이용한 동사분류", 한국인지과학회 춘계 학술대회, 232-238쪽, 서울, 1999년 5월 29일

[이공주 1996] 이공주, 김재훈, 최기선, 김길창, 구문 트리 부착 코퍼스 구축을 위한 한국어 구문 태그, 인지과학, Vol. 7, No. 4, 7-24쪽, 1996년

[조정미 1998] 조정미, "코퍼스와 사전을 이용한 동사 의미 분별", 박사학위 논문, 한국과학기술원 전산학과, 1998년

[조평옥 1997] 조평옥, 옥철영, "한국어 명사 의미 계층 구조 구축", 한글 및 한국어 정보처리 학술대회, 129-165쪽, 부산, 1997년 10월 10-11일

[한글학회 1991] 한글학회, "우리말 큰 사전", 어문각, 1991년

[홍재성 1997] 홍재성, 김원근, 김현권, 류시종, 박만규, 박진호, 심봉섭, 안근종, 우순조, 임준서, "현대 한국어 동사 구문 사전", 두산동아, 1997년 1월

[Cho 1999] Pyeong-Ok Choi, Cheol-Yeong Ock, Soo-Dong Lee, Jae-Duk Park and Dong-In Park, "A Korean Noun Semantic Hierarch Based on Semantic Feature", International Conference on Computer Processing of Oriental Languages, pp. 211-216, Tokushima Japan, March 24-26, 1999

[ISO 8879] ISO 8879. Information Processing - Text and Office Systems - Standard Generalized Markup Laguage(SGML), International Organization for Standardization, 1986.

[Li 1998] Hang Li and Naoki Abe, "Generalizing Case Frames Using a Thesaurus and the MDL Principle", Computational Linguistics, Volume 24, Number 2, pp. 217-244, June 1998