세종전자사전을 활용한 한국어 구문분석

성열원
서울대학교 인지과학협동과정
superbia@naver.com

Korean Parsing using Sejong Dictionary

Seong, Yeolwon
Dept. of Cognitive Science

요 약

본 논문에서는 세종전자사전의 정보를 활용하여 논항 결합의 정확도를 향상시키는 한국어 구문분석 모델을 제안한다. 구문분석 과정에서 노드간의 결합 가능성을 계산할 때, 세종전자사전 동사사전의 격틀 정보, 논항 제약 정보와 명사사전의 의미부류 정보를 활용하여 가산점을 부여하여 사전의 내용과 일치하는 결합이 선호되도록 하였다. 이 과정에서 구조적 오류를 해결할 수 있었고, 결합에 참여하는 동사와 명사의 의미 중의성도 해소할 수 있었다. 평균 13어절 길이의 실험용 문장 50개를 대상으로 실험한 결과, 35% 정도의 오류 감소 효과를 볼 수 있었다. 또한 구문분석 결과 정보를, 전자 사전에 기술된 정보의 완결성을 시험하고 보완하는 데에도 활용하였다.

1. 서론

구문분석기의 연구에서, 사전 정보의 활용이 필요하다는 것은 오래 전부터 인식되어 왔다. 그러나 구문 분석기에서 활용할 수 있을 정도로 정밀하게 정보가 기술되어 있고, 또한 많은 엔트리를 대상으로 하는 사전은 지금껏 없었다. 이 때문에 현재까지 대부분의 구문 분석 연구에서는 어휘 단위의 상세한 결합 정보를 활용하지 못했었다.

본 논문은 올해 10년간의 연구가 완료되는 세종전자사전의 정보를 구문분석기에 적용해보고, 어느 정도 성능 향상에 기여할 수 있는지 실험해 본다. 세종전자사전은 한국어 개별 어휘의 통사적, 의미적 특성을 상세히 기술한 대규모 상세 전자 사전으로, 구문분석을 비롯한 언어처리 전반에 활용될 수 있는 많은 정보들을 담고 있다.

덧붙여 구문분석기를 활용하여 사전 자체의 완결성을 높이는 방법에 대해서도 설명한다. 언어학자의 직관에 의존해서 만든 사전이 그 자체로 완결성을 가지기는 어렵다. 구문분석기를 이용하면 사전에 기술된 정보에 문제가 없는지 확인할 수 있어, 사전의 수정 보완에 큰 도움이 된다.

2. 기존 연구

구문분석기의 초창기 연구에서는 주로 사람이 기술한 문법을 주요 정보로 삼았다. 그러나 사람이 문법을 완벽하게 기술하는 것 자체가 어려운 일이고, 또 언어학자가 기술한 문법을 전산 처리에 알맞은 형태로 바꾸는 작업도 매우 어려운 일이어서, 대부분의 문법 기반 구문분석 시스템은 구축하기도 어렵고, 강건성의 측면에서도 문제가 있었다.

1990년대 후반부터는, 통계적 기법이 많이 적용되었다. 문법 기반 시스템의 문제점인 구축의 어려움과 강건성의 문제를 한꺼번에 해결할 수 있어서, 많은 연구자들이 이 방식을 택하였다. 대용량 트리-태그드 코퍼스가 구축되고, 컴퓨터의 성능이 좋아지면서 통계적 기법은 더욱 각광을 받게 되었다. 통계적 기법이 사용되면서 구문분석 연구는 한 단계 도약했다고 할 수 있다. 그렇지만 여전히 실용화(자동번역 등에서 사용될 수 있는) 수준에는 미치지 못하였는데, 이는 Sparseness 문제 때문이다.

최근 들어 상당히 큰 트리코퍼스가 구축되긴 했어도, 그 정도로 언어의 다양한 사용을 다 포착할 수는 없다. 품사나 구절기호 수준을 통계의 기본 단위로 삼는 경우에는 현재 구축된 코퍼스로도 충분한 정보를 뽑을 수

있겠지만, 이 방식에서는 개별 어휘의 결합 특성을 고려할 수가 없어 성능 향상에 한계가 있다.

이러한 문제 때문에 어휘화된 통계 기법, 즉 통계의 수준을 구절기호에서 어휘 수준까지 상세화하는 기법이 도입되었다. 이 경우, 어휘별로 상세한 결합 특성을 통계적으로 뽑아낼 수 있지만, 문제는 정보의 양이 매우 부족하다는 것이다. 하나의 품사에 포함되는 어휘의 수는 만개가 넘는 경우도 많다. 또한 결합 정보의 관점으로 본다면 그 개수는 엄청나게 늘어난다. 예를 들어 N과 V의 결합을 구절기호 수준에서 본다면 한가지가 되겠지만 5만개의 명사와 1만개의 동사를 가정한다면 5억 가지가 된다. 이 정도로 자세한 정보를 무리 없이 뽑을 수 있는 트리코퍼스를 구축하는 것은 당분간 매우 어려울 것으로 생각된다.

이를 돌파하기 위해, 다양한 연구들이 진행되었다. 가장 많이 적용된 것이 공기 정보의 활용이다. 공기 정보는 코퍼스 상에서 연속되어 함께 나타난 어휘쌍의 집합으로, 이를 활용하면 어휘적 결합 특성을 어느 정도 처리할 수 있다. (김나리 1997, 윤준태 1997, 이공주 1998, 2002)

또한 트리 코퍼스의 다른 정보를 활용하는 연구도 있었다. 결합 대상이 되는 요소 뿐만 아니라, 그 앞뒤나 부모를 문맥의 개념으로 추가적으로 고려하는 방식이다. (이공주 1998, 박소영 2002, 2004)

이러한 흐름과 별도로, 개별 어휘의 상세한 결합 정보를 활용하려는 연구도 지속되어 왔다. 초기의 연구는 사람이 기술한 정보를 활용하는 것이었는데, 이러한 연구들은 대부분 시험적인 적용에 그치고 있다. (조성원 1991, 송영빈 1999, 이수선 1999)

이후에는 코퍼스로부터 어휘적 결합 정보를 자동으로 뽑아내는 연구들이 진행되었다. (류법모 1996, 최용석 1999) 그러나 아직 이러한 연구들도 충분한 수준에는 미치지 못하고 있다.

3. 세종전자사전의 구조

세종전자사전은 XML 형식으로 기술되어 있으며, 품사마다 그 구조가 다르다. 본 연구에서는 세종전자사전 중, 체언사전과 용언사전의 정보를 주로 활용하였다.

먼저 체언사전의 구성은 다음과 같다. “가격”이라는 어휘의 예로 설명한다. (항목별로 주요 내용만을 추려서 보였다.)

```
<superEntry>
  <orth>가격</orth> // 철자
  <entry n="1" pos="nng_s"> // 엔트리
    <see></see> // see also
    <morph_grp> // 형태소 정보
      <org lg="si">價格</org> // origin
      <comp pos="n" type="suffix">~수준</comp> // 복합명사
    </morph_grp>
    <sense n="1"> // 첫번째 의미 부분
      <sem_grp>
        <eg>공산품 ~이 크게 오를 것으로 우려되고 있다.</eg> // 예문
        <trans>price</trans> // 대역어
        <sem>값</sem> // 의미부류
      </sem_grp>
      <syn_grp>
        <n_aj type="appr">~이 높다</n_aj>
        <n_v type="narg_vpred">
          <form vcompound="no">~을 올리다</form>
        </n_v>
      </syn_grp>
    </sense>
  </entry>
  <entry n="2" pos="nng_s">
    <morph_grp>
      <comp pos="v" type="suffix">~하다</comp>
      <comp pos="v" type="suffix">~당하다</comp>
    </morph_grp>
    <sense n="1">
      <sem_grp>
        <eg>챔피언은 도전자의 복부에 연속으로 ~을 하였다.</eg>
        <trans>strike</trans>
        <sem>방향성행위</sem>
      </sem_grp>
      <syn_grp>
        <n_aj type="appr">~이 주효하다</n_aj>
        <n_v type="npred_vsup">
          <form vcompound="yes">~을 하다</form>
          <frame>X-이 Y-을|에게 ~을 하다</frame>
          <frame>X-이 Z-을|에 ~을 하다</frame>
        </n_v>
      </syn_grp>
    </sense>
  </entry>
</superEntry>
```

이 내용 중에서 많은 것들이 구문 분석에 필요하지만, 그 중에서도 가장 중요한 것은 <sem_grp> 하위의 <sem> 정보이다. 이는 그 어휘 (정확히 말하면 어휘 하위의 의미단위)가 어떤 의미 부류에 속하는지를 표시하는 것으로, 격틀 사전에서 논항 매칭을 할 때 사용된다.

다른 정보들 중 <syn_grp>에 속하는 것들은 명사를 핵으로 하는 구조를 분석할 때 필요한 것들이다. (본 논문의 시스템에서는 아직까지 이 정보를 적용시키지 못하고 있다.)

세종 사전의 명사 의미 부류는 다른 일반적인 연구와 마찬가지로 트리 형태로 구성되어 있다. 모두 590개의 의미 항목을 가지고 있다. 실제 언어 생활에서 중요하게 다루어지는 ‘인간’과 같은 의미 항목은 대단히 세밀하게 구분되어 있다.

동사의 경우, 큰 구조는 명사와 비슷하지만 더 많은

정보가 기술되어 있다.

```
<superEntry>
  <orth>가결하다</orth>
  <entry n="1" pos="vv">
    <mnt_grp>
    </mnt_grp>
    <morph_grp>
      <var type="spr">가결을 하다</var>
      <cntr opt="opt" type="VDelCCon"/>
      <org lg="si">可決_</org>
      <infl type="yeo"/>
    </morph_grp>
    <sense n="01">
      <sem_grp>
        <class>결과행위</class>
        <trans>pass</trans>
        <trans>adopt</trans>
      </sem_grp>
      <frame_grp type="FTR">
        <frame>X=N0-이 Y=N1-을 V</frame>
        <subsense>
          <sel_rst arg="X" tht="AGT">인간집단</sel_rst>
          <sel_rst arg="Y" tht="THM">(법률|안건)</sel_rst>
          <n_appr arg="X"></n_appr>
          <eg>국회가 이번 안건을 만장일치로 가결했다는 것은 놀라운 일이다.</eg>
          <eg>이 협회에서 만장일치로 법안을 가결한 것은 이번이 처음이자 마지막 일이 될 것이다.</eg>
        </subsense>
        <subsense>
          <sel_rst arg="X" tht="AGT">인간집단</sel_rst>
          <sel_rst arg="Y" tht="THM">S것</sel_rst>
          <n_appr arg="X"></n_appr>
          <eg>국회는 이번 회기에 동티모르에 군대를 파병할 것을 가결했다.</eg>
        </subsense>
      </frame_grp>
      <com>
        <col_grp>
          <col type="magn">만장일치로</col>
        </col_grp>
      </com>
    </sense>
  </entry>
</superEntry>
```

위의 내용 중에서 가장 중요한 정보는 <frame_grp>의 내용이다. <frame> 안에는 격틀의 정보가 들어 있고, <subsense>의 <sel_rst> 안에는 각 논항의 선택제약 정보가 들어 있다. 선택제약 안에 기술된 내용은 의미부류 (여기서는 '인간집단') 이거나 혹은 실제 어휘들 (이 경우에는 '(..)'안에 들어있다. 예에서는 법률|안건 이 기술되어 있다.)이다.

이와 같은 동사의 격틀 정보, 논항별 선택제약 정보와, 명사의 의미부류 정보, 그리고 명사 의미부류 트리를 이용하면 구문 분석에서 논항 결합이 적절히 분석되었는지를 판단할 수 있다.

본 연구에서는 세종 전사사전의 상세 전자사전 전체의 명사/동사/형용사 정보를 사용하였다. 그 엔트리 숫자는 다음과 같다.

< 세종 사전의 엔트리 수 >

	명사	동사	형용사
개수	25554	15180	4400

4. 구문분석기에 적용

세종전자사전의 정보를 활용하도록 구문분석기를 개발하였다. 본 구문분석기는 상향식 차트 파서로 바이너리 결합만을 인정한다. 정보로서는 기본적으로 문법을 사용하며 세종 사전의 정보 중에서는 프레임 정보와 명사 의미 정보를 활용하여 논항 결합 가산점과 명사 병렬 결합 가산점을 계산한다. 이 외에도 어미 위계 정보, 허사 단위 결합 정보, 문법적 숙어 정보, 동적 구조 정보 등 다양한 정보를 사용한다. 각 정보를 이용한 결합 강도 조절 부분은 모듈화 되어 있어, 문법을 제외한 다른 정보는 사용하지 않고도 동작이 가능하다. 두 개의 노드가 결합하여 새로운 노드를 생성할 때, 각 노드는 점수를 부여받는데, 이 점수는 형태소 분석기에서 출력한 어휘의 점수와 결합을 허용한 문법의 점수, 기타 각 정보 모듈에서 얻어진 가산점을 합산하여 계산된다. 부분 점수들이 통계에 의해서 뽑혀진 것이 아니기 때문에 1을 총점으로 하는 확률 형태의 점수는 사용하지 못했다.

본 논문에서는 세종 사전의 정보 활용 중에서도 논항 결합 부분을 설명한다.

세종 사전을 이용한 프레임 매칭을 위해 먼저 사전의 정보 중, 필요한 것만을 뽑아내어 별도의 데이터로 구축하였다.

< 용언 사전 >

```
#가격하다
@n : 1 pos : vv
blow=+AGT[인간]이  +THM[인간|신체부위]을  +INS[구체물]로 V
attack=+AGT[국가]이 +THM[국가]을 +INS[구체인공물]로 V
blow=+EFF[구체인공물]이 +THM[인간|신체부위]을 V
attack=+EFF[구체인공물]이 +THM[국가]을 V
```

< 체언 사전 >

```
#가격
-1 :nng_s @값
-2 :nng_s @방향성행위
```

용언 사전의 경우, 위와 같은 형태에서 성능의 향상을 위해 엔트리의 정보와 각 결합관계별 논항 정보로

다시 나누어서 관리한다.

위와 같은 정보를 활용하여, 동사를 헤드로 하는 노드 결합시, 사전에 기술되어 있는 결합인지를 판별하여 결합 점수에 가산점을 부여한다. 각 단계는 다음과 같다.

1. 헤드 동사의 어휘를 프레임 사전에서 검색
2. 문법에서 판단한 결합 관계(AGT 등)에 해당하는 논항 제약 정보만을 뽑아옴
3. 사전의 논항 제약 정보의 허사 제약 정보와 실제 결합한 논항의 허사 정보를 매칭 (이 경우 격조사가 아이라 보조사가 사용된 경우와 관계문 결합의 경우에는 점수를 조금 낮추어서 허용)
4. 사전의 논항 정보에 기술된 어휘목록과 실제 결합한 논항의 어휘 정보를 매칭, 매칭된 경우 만점을 반환, 단 복합명사의 처리를 위해 뒷부분만 일치하는 경우 일정 점수를 부여
5. 사전의 논항 정보에 기술된 의미목록과 실제 결합한 논항의 의미 정보를 매칭, 이 때 의미목록의 트리를 활용함. 목록상 어떤 위치에서 매칭이 되는가에 따라 다른 점수를 부여

프레임 매칭에 소요되는 시간을 줄이기 위해, 한번 매칭된 정보는 저장하여 다음 번에 검색하여 사용하도록 하였다. 프레임 매칭에서는 각 노드의 헤드 정보만을 사용하기 때문에, 파싱이 진행되었다고 해서 다시 매칭을 해야 하는 것은 아니다. (이에 반해 노드의 구조 상태를 보는 문법 등에서는 매번 매칭을 해야 한다. 파싱의 매칭을 동적 구조를 사용하는 것과 그렇지 않은 것으로 나누어서 고려해야 한다.) 이러한 기법은 윤준태 1997의 전역적 연관표(GAT)와 유사점이 있다.

그러나 분류사나 단위사의 같은 경우에는 그 의미적 특성을 자식에서 가져와야 하는 경우가 있다.

밥 한 공기를 먹었다.

위와 같은 문장에서는 '공기를 먹다'가 아니라 '밥을 먹다'로 의미적 분석을 해야 한다. 현재 분석기에서는 가까운 것끼리 먼저 결합하게 되므로, 먼저 '공기를 먹다'가 결합되게 되고, 이 때는 '기체를 먹다'의 의미로 결합된다. (연탄가스를 먹었다의 의미) 그 후에 밥과 공기가 결합한 후 마시다와 결합할 때는 공기가 의미태그를 밥에서 가져오도록 하여 '음식을 먹다'의 의미로 결합되도록 한다.

본 구문분석기는 바이너리 결합만을 허용하므로, 단일 결합 판단시에는 프레임의 전체 모습을 고려할 수가 없다. 이 경우, 동일 용언의 서로 다른 프레임에서 사용된 논항들이 한 번에 결합할 가능성이 있다. 이를 방지하기 위해, 전체 문장 분석 후, 용언별로 결합된 논항들을 점검하여 같은 프레임에서 사용된 것들만 남기는 후처리를 추가하였다. 또한 동일한 결합 관계의 논항이 2번 이상 나오는 경우도 방지하였다. (그러나 아직 토픽에 해당하는 보조사들의 처리에는 문제가 있다.)

5. 가산점의 계산

사전의 프레임 정보를 잘못 사용하면, 오히려 오분석이 늘어날 가능성도 있다.

나는 아침에 학교에 갔다.

위의 문장을 처음 시험하였을 때는 '날아다니는 아침'으로 분석되었다. '아침'에는 '하루중단계'와 '음식'의 의미부류가 들어있고, '날다'에는 '구체물이 날다'의 프레임이 들어있으며, 의미트리에는 '음식'이 '구체물'의 하위에 들어있기 때문이다.

이 문제를 해결하기 위해서는 '아침'의 의미가 문장 내에서 하나로 유지되도록 분석할 필요가 있다. 그렇지만 다음과 같은 경우는 또다시 문제가 된다.

나는 아침을 먹고 학교에 갔다.

위의 경우는 '아침'이 '하루중단계'가 아닌 '음식'의 의미이기 때문에, '날아다니는 아침'으로 해석되어도 걸러낼 수가 없다.

이 문제를 해결하기 위해서는 '날아다니는 음식'이 말이 되지 않음을 판단해야 하는데, 그러기 위해서는 다음과 같은 확률을 판단해야 한다.

1. '아침'이 '음식'으로 쓰일 확률
2. '날다'의 전체 프레임에서 '구체물이 날다' 형태로 쓰일 확률
3. '음식'이 '구체물'로 쓰일 확률

위의 확률은 모두 태그드 코퍼스(어휘들이 의미별로 구분이 되어 있어야 한다)를 분석하여 실제로 얼마나 사용되었는지 카운트하면 알 수 있을 것이다. 그러나 이러한 정보를 뽑기에는 대용량의 정밀한 태그드 코퍼스가 필요하므로, 앞으로 좀 더 연구가 필요할 것이다.

위의 정보들 중에서 세번째의 경우는 의미트리를 잘 분석하는 것으로 대략적인 점수를 뽑아 낼 수 있다.

점수에 관여하는 요소는 다음과 같다.

1. 상위어와 하위어의 깊이 차이

상위어와 하위어가 다 같이 '인간' 이라면 점수가 높을 것이다.

상위어가 '온' 이고, 하위어가 '소속인가' 이라면 점수가 낮을 것이다.

2. 상위어의 깊이

상위어가 '소속인간' 이라면 점수가 높을 것이다.

상위어가 '온' 이라면 점수가 낮을 것이다.

3. 상위어의 자식수
상위어의 자식이 많다면 점수가 낮을 것이다.
같은 깊이라도 '생물' 아래에는 70여 개의 자식 태그가 있지만, '상자' 아래에는 아무것도 없다.

4. 하위어의 자식수
하위어의 자식이 많다면 점수가 낮을 것이다.

위와 같은 기준으로 만든 현재 모듈에서는 다음과 같은 점수를 부여하고 있다.

인간 - 화시적인간 : 60
장소 - 기관건물 : 46
인간집단 - 교육기관 : 57
행위 - 물리적행위 : 35

6. 중의성 해소

논항 사전의 적용으로 기대되는 효과는 중의성 해소이다. 결합시 의미적 가능성을 검토함으로써 여러 후보들 중에 의미상 말이 되는 것을 골라낼 수 있다. 동사와 논항의 중의적 의미 중 적절한 것을 골라낼 수 있을 뿐 아니라 구조적으로 중의적인 것들도 해결할 수 있다. 몇 가지 기본 문장들의 처리 결과를 보이면 다음과 같다.

```
# 나는 가스를 먹었다.

  T_N:나@나이|추상적대상|화시적인간+=은
[10781-12575]-F
  O_N:가스@기체|배설물+=을 [10610-14831]-F
R_V:먹@+=다 [0-0]

Trimmed Frame [먹] :
  (53점) <2번> inhale
    A 나+은 : (17점) <2번> 화시적인간 = 인간
    T 가스+을 : (90점) <0번> 기체 = 기체
  (52점) <13번> consume
    T 나+은 : (5점) <2번> 화시적인간 = 구체물
    T 가스+을 : (100점) <-1번> '가스' = '가스'

# 이 차는 가스를 많이 먹는다.

    M_N:이@신체부위|구체물의부분|추상적대상+=
[9818-12155]
  T_N:차@육상교통기관|구체인공물|음료|나무|식물의부분|
관계수|크기+=은 [9845-12034]-F
  O_N:가스@기체|배설물+=을 [9424-14870]-F
  M_B:많이@+= [8084-7252]
R_V:먹@+=는다 [0-0]

Trimmed Frame [먹] :
  (43점) <13번> consume
    L 차+은 : (30점) <0번> 육상교통기관 = 교통기관
    T 차+은 : (6점) <0번> 육상교통기관 = 구체물
    T 가스+을 : (100점) <-1번> '가스' = '가스'

# 나는 새를 보았다.

  T_N:나@나이|추상적대상|화시적인간+=은
[10005-12140]-F
  O_N:새@새|관계장소|관계시간+=을 [9462-12086]-F
R_V:보@+=다 [0-0]

Trimmed Frame [보] :
  (33점) <0번> see
    A 나+은 : (17점) <2번> 화시적인간 = 인간
    T 나+은 : (5점) <2번> 화시적인간 = 구체물
    T 새+을 : (49점) <1번> 관계장소 = 장소
  (35점) <3번> appraise
    A 나+은 : (17점) <2번> 화시적인간 = 인간
    T 나+은 : (17점) <2번> 화시적인간 = 구체자연물
    T 새+을 : (53점) <0번> 새??? = 구체자연물

# 그는 나는 새를 보았다.

  T_N:그@인간+=은 [9843-11922]-F
    K_V:날@+=는 [9178-11734]-F
  O_N:새@새|관계장소|관계시간+=을 [9736-12146]-F
R_V:보@+=다 [0-0]

Trimmed Frame [날] :
  (90점) <0번> fly
    A 새+을 : (90점) <-1번> '새' = 새
Trimmed Frame [보] :
  (31점) <0번> see
    A 그+은 : (14점) <0번> 인간 = 인간
    T 그+은 : (4점) <0번> 인간 = 구체물
    T 새+을 : (49점) <1번> 관계장소 = 장소
  (33점) <3번> appraise
    A 그+은 : (14점) <0번> 인간 = 인간
    T 그+은 : (14점) <0번> 인간 = 구체자연물
    T 새+을 : (53점) <0번> 새??? = 구체자연물

# 힘이 나는 음식을 먹었다.

      S_N:힘@능력|방향성행위|속성값+=이
[9234-11904]-F
    K_V:나@+=는 [9371-11153]-F
  O_N:음식@음식+=을 [10025-14318]-F
R_V:먹@+=다 [0-0]

Trimmed Frame [나] :
  (74점) <13번> have
    T 힘+이 : (59점) <0번> 능력 = 속성
    L 힘+이 : (1점) <-1번> '힘' = -
    L 음식+을 : (90점) <-1번> '음식' = 음식
Trimmed Frame [먹] :
  (90점) <0번> eat
    T 음식+을 : (90점) <-1번> '음식' = 음식
```

7. 실험

아직 구문분석기의 성능이 코퍼스의 일반적인 문장을 대상으로 할 정도의 수준이 아니고, 또한 세종사전의 영향을 파악하려면 소수의 문장을 대상으로한 정밀한 결과 분석이 필요하다고 판단하여, 중요 문장 50개를 선정하여 시험하였다. (이 문장들은 세종사전팀 내부에서 구문분석기 시험용으로 선정한 것이다. 문장길이는 평균 13.08 어절이다.)

사전의 기여도를 평가하기 위해서, 먼저 문장들에 대해서 기본적 튜닝을 하였다. (단 논항 결합과 관련된 부분은 튜닝하지 않았다.) 그 후, 사전을 결합한 상태와 결합하지 않은 상태에서 각각 구문 분석을 실행하고 그 결과를 수작업으로 분석하였다. 결과는 다음과 같다.

	사전 적용	사전 비적용	차
오류A 노드의 수	11개	19개	8개
오류B 노드의 수	10개	11개	1개
오류A의 확률	2.07%	3.59%	1.52%
오류B의 확률	1.89%	2.07%	0.18%
평균 오류율	3.02%	4.63%	1.61%
상대 오류	5개	8개	3개
전체 처리 시간	17.562초	16.172초	1.39초
문장당 시간	0.3512초	0.3234초	0.0278초
초당 처리 어절 수	37.24개	40.44개	3.2개

* 전체 노드 개수 : 529개
* 전체 어절 개수 : 654개

노드의 오류 여부는 부모를 제대로 잡았는지, 그 관계가 맞는 관계인지의 여부를 판단하였다. 오류A는 명확한 오류이고, 오류B는 기준에 따라서 판단이 달라지는 것들이다. (공유 주어나 복합명사끼리의 병렬 등등) 평균 오류는 오류B의 경우를 0.5개 틀린 것으로 하여 계산한 오류율이다.

상대 오류는 다른 쪽 시도(사전 적용/비적용)에서는 맞게 나왔으나 이 쪽 시도에서는 틀린 경우 중, 사전과 관련 있는 문제들만을 카운트한 것이다. 사전을 적용한 결과 좋아진 경우가 8건, 더 나빠진 경우가 5건으로 분석되었다. 상대 오류의 내용을 보이면 다음과 같다.

< 향상된 것들 >

// 거래가 나타나다 -> 거래가 위축되다
재건축 아파트에 대한 규제를 골자로 한 부동산 대책 이후 서울을 비롯한 전국의 아파트 거래가 급격히 위축된 것으로 나타났다.

// 세력이 생각하다 -> 세력이 무능하다
민주개혁 세력이 국민들이 먹고사는 문제에서는 무능했다고 생각한다.

// 미성년자와 계약 -> 미성년자와 한
민법에 따르면 법정대리인의 동의 없이 만 20세 미만의 미성년자와 한 계약은 취소할 수 있다.

// 남긴 원대 -> 남긴 땅
고인이 된 아버지의 뜻을 기리기 위해 두 아들이 아버지가 남긴 시가 50억 원대 땅을 대학에 발전기금으로 내놓았다.

// 친구가 베풀다 -> 친구가 처하다
진정한 우정이란 친구가 어려움에 처했을 때 무조건 도움을 베푸는 것이 아니라, 친구 스스로 문제를 해결해 나갈 수 있는 힘을 길러 주는 것이다.

// 몇 십 리 살고 있었다 -> 몇 십 리 떨어진
옛날 이 장터에서 몇 십 리 떨어진 마을에 앞을 못 보는 아버지를 봉양하는 여자 아이 하나가 살고 있었다.

// 상대로 하는 일기 -> 상대로 하는 예보
변덕이 죽 끓듯 하는 날씨를 상대로 하는 일기 예보에서 100%의 정확도를 기대하기란 어렵다.

// 자신에게 가진 -> 자신에게 불리한
여러 가지 자신에게 불리한 증거를 가진 사람이 아무 죄가 없을 확률이 높다.

< 악화된 것들 >

// 소식통이 밝히다 -> 소식통이 말하다
미군의 한 고위 소식통이 현지 시간 23일 조지 W 부시 미 대통령이 한국에 전시 작전 통제권을 이양하는 데 공감한다고 말했다고 밝혔다.

// 요즘은 들다 -> 요즘은 투자하다
요즘은 몇 시간을 투자해서 신문을 읽고 나면 머리가 띵하고 내일은 차라리 신문을 보지 말아야지 하는 허탈감마저 든다.

// 스스로 해결하다 -> 스스로 것이다
진정한 우정이란 친구가 어려움에 처했을 때 무조건 도움을 베푸는 것이 아니라, 친구 스스로 문제를 해결해 나갈 수 있는 힘을 길러 주는 것이다.

// 갖고 있었던 생각 -> 갖고 있었던 부모
오랫동안 갖고 있었던 부모에 대한 생각을 바꾸는 것이 쉬운 일은 아니다.

// 목표, 각오를 -> 목표, 것이다
해이해지는 마음을 다스리는 효과적인 방법 중의 하나는 자신의 희망과 목표, 그리고 각오를 글로 쓰는 것이다.

향상된 것들은 대부분 동사와 논항의 결합도가 높아져서 옳은 분석으로 수정된 것들이다. 일반적 논항 이외에도 관형절의 경우 향상된 것들이 있다.

악화된 것들은 의미상으로 양쪽 다 결합이 가능한경우, 일반적 논항이 아닌 토픽에 해당하는 결합 등 다양한 오류들로 분석되었다.

8. 구문분석기를 활용한 사전의 보완

구문분석 결과를 이용하여 사전의 프레임 정보를 수정할 수도 있다. 사전의 오류는 논항 제약 정보가 잘못 기술되어 있거나, 명사에 의미부류 정보가 잘못 기술되어 있는 경우이다. 현재 시스템에서는 의미 정보가 매칭되지 않고 격틀의 허사 정보만 매칭된 경우에도 일부 가산점을 부여하며 그 내용을 결과 트리에 표시하고 있다. 간단한 예를 들면 다음과 같다.

```
# 금리가 추가로 인상될 경우…

  S_N:금리@수량+=이 [7791-8091]-E
  B_N:추가@단독행위+=으로 [7621-8164]
B_V:인상되@+=을경우 [7555-6949]
```

트리의 맨 윗줄 금리.. 부분 뒤에 붙은 '-E'표지가 허사부만 매칭된 경우를 표시한다. (의미까지 매칭된 경우는 '-F'로 표시한다.) 이 경우, 동사 사전을 살펴보면 다음과 같다.

```
#인상되다
go  up=+THM[비율(이자율|환율)|값(가격|전세비)|금전(세금|교통비|월급)]이 V
```

이 문제를 해결하려면, 인상되다의 THM 부분에 '수량'의 정보를 추가하던가, 아니면 '금리'의 의미부류 정보에 '비율'등을 추가하면 된다. (괄호 밖의'비율', '값', '금전'은 의미 부류이고, 괄호 안의 내용은 대표적인 명사 어휘를 표시한다)

현재 사전팀 내부에서 이러한 방식으로 사전의 문제점을 찾아서 개선하는 작업을 진행하고 있다.

9. 향후 과제

본 논문에서는 정밀 사전 자료를 구문 분석기에 적용시켜 보고, 그 과정에서 고려해야 할 여러 요소들을 정리하였다. 한국어 구문 분석과 사전 기반 언어처리라는 두 가지 영역을 하나의 논문에서 다루려다 보니, 세부 주제에 대한 깊은 연구는 부족한 면이 있다. 향후 사전 기반 언어처리 기법이 발전하려면, 이 논문에서 간략히 다루었던 여러 주제들에 대해 더욱 깊은 연구가 뒤따라야 할 것이다.

본 논문에서 사용하고 있는 파서의 처리 방식, 특히 가산점 시스템에 대해서는 다른 논문에서 자세히 설명할 기회가 있었으면 한다. 체언(명사)의 격틀/공기 정보 활용에 대해서도 매우 중요한 주제로 연구가 뒤따라야 할 것이다. 그 외에 복합명사, 의존명사, 분류사, 고유명사, 관용어 등 실용적 수준의 구문 분석을 위해 필요한 세종 사전의 추가적인 정보 활용에 대해서도 연구가 필요하다. 또한 엔진이 어느 정도 안정화 되고 나면 실험의 방식도 대용량 태그드 코퍼스와의 비교를 통한 자동화된 방법을 도입하여야 할 것이다. 그리하여야만 비로서 사전 기반 구문 분석이 다른 방식에 비하여 어느 정도 효과적인지 비교할 수 있을 것이다.

또한 다양한 구문 분석 및 의미 분석, 그리고 형태소 분석 등의 언어처리 시스템에서 세종 사전을 활용한 성능 향상에 대해 연구가 이루어지기를 희망한다.

10. 결론

본 논문에서는 대용량 고정밀 사전의 정보를 활용한 구문분석기의 성능 향상 기법을 설명하였다. 용언 사전의 격틀 정보와 논항 제약 정보, 체언 사전의 의미 부류 정보, 그리고 체언의 의미 트리를 정보로 사용하여, 상향식 차트 파서의 노드 결합 처리시 가산점을 부가하는 방식을 사용하였다.

논항 결합시, 사전에 기술된 결합은 가산점을 부여받게 되어 더 정확한 분석이 가능하다. 또한 각 용언과 논항이 어떤 의미로 사용되었는가도 분석이 가능하다.

소규모의 정밀 테스트 결과, 완성도 높은 구문분석기에 사전 격틀 정보가 적용될 경우, 노드당 결합분석 오류율을 4.63%에서 3.02%로 낮추어 35% 정도의 오류 감소 효과를 볼 수 있었다.

향후 대규모 코퍼스를 대상으로 한 시험 및 튜닝, 그리고 동사의 격틀 및 논항 제약 정보 이외의 다른 정보들의 활용 등의 작업이 뒤따라야 한다.

지금까지 오랫동안 정밀한 사전에 기반한 구문분석기의 가능성이 논의되어 왔지만, 대용량 고정밀 사전이 구축되지 못하여서 실험적인 수준에서만 연구가 이루어져왔다. 이제 10년간의 연구의 결과로 세종 전자 사전이 완결되었으므로, 이를 활용한 다양한 활용 방안을 연구할 수 있게 되었다.

실용적인 구문분석기를 개발하기 위해서는 구절기호 수준의 문법이나 통계정보만으로는 한계가 있다. 이러한 한계를 극복하려면 어휘 수준의 정보를 활용해야 하는데, 어휘 수준에서 통계 정보를 뽑는 방식을 활용할 수도 있고, 정밀한 사전의 정보를 사용할 수도 있다. 어휘 수준의 통계 기법을 사용하려면 매우 큰 용량의 태그드 코퍼스가 필요하지만, 아직까지 그러한 자료는 없다. 세종 사전의 완성을 계기로 어휘 단위의 정밀한 언어처리 연구가 활성화되기를 바란다.

참고 문헌

[1] 김나리, "패턴 정보를 이용한 한국어 구문 분석", 서울대학교 박사학위 논문, 1997.

[2] 김미영, 강신재, 이종혁, “단위 분석과 의존문법에 기반한 한국어 구문분석”, 2000년도 한국정보과학회 봄 학술발표논문집, pp. 327-329, 2000.

[3] 류법모, 이태승, 이종혁, 이근배, “술어 중심 제약 전파를 이용한 2-단계 한국어 의존 파서”, 1996년도 한국정보과학회 봄 학술발표논문집, pp. 923-926, 1996.

[4] 박소영, 곽용재, 정후중, 황영숙, 임해창, "한국어 구문분석의 효율성을 개선하기 위한 구문제약규칙의 학습", 정보과학회논문지:소프트웨어 및 응용 제29권 제10호 (2002. 10), pp. 755-765, 2002.

[5] 박소영, 김수홍, 임해창, "문장성분의 다양한 자질을 이용한 한국어 구문분석 모델", 정보처리학회논문지 B 제11-B권 제6호(2004, 10), pp. 743-748, 2004.

[6] 송영빈, 채영숙, 박용일, 이정민, 설가영, 황혜리, 한나리, 최기선, "동사의 애매성 해소를 위한 구문의미사전의 구축", 한글 및 한국어 정보처리학술대회, pp. 280-287, 1999.

[7] 신서인, “구문분석말뭉치를 이용한 한국어 문형 연구”, 서울대학교 박사학위 논문, 2006.

[8] 윤준태, 김선호, 송만석, "전역적 연관 표를 이용한 한국어 구문 분석", 정보과학회논문지(B) 제24권 제11호, pp. 1297-1306, 1997.

[9] 윤준태, "공기 관계 기반 어휘 연관도를 이용한 한국어 구문 분석", 연세대학교 박사학위 논문, 1997.

[10] 이공주, "언어특성에 기반한 한국어의 확률적 구문분석", 한국과학기술원 박사학위 논문, 1997.

[11] 이공주, 김재훈, 김길창, "제한된 형태의 구구조 문법에 기반한 한국어 구문분석", 정보과학회논문지(B), 제25권 제4호, pp. 722-732, 1998.

[12] 이공주, 김재훈, "중심어 간의 공기정보를 이용한 한국어 확률 구문분석 모델", 정보처리학회논문지(B) 제9-B권 제6호, pp. 809-816, 2002.

[13] 이수선, 박현재, 우요섭, "한국어 분석의 중의성 해소를 위한 하위범주화사전 구축", 한글 및 한국어 정보처리학술대회, pp. 257-264. 1999.

[14] 임홍빈, 이홍식 외, “한국어 구문분석 방법론”, 한국문화사, 2002.

[15] 조성원, 송만석, "사전에 기반한 한국어 문장 해석 시스템 원형의 연구", 한국정보과학회 가을 학술발표논문집 Vol. 18, No. 2, pp. 809-812, 1991.

[16] 최용석, 이주호, 최기선, "격틀 자동구축과 격틀평가 방법에 관한 연구", 한글 및 한국어 정보처리학술대회, pp. 272-279, 1999.

[17] 홍재성 외, “현대 한국어 동사 구문 사전”, 두산동아, 1997.

[18] 홍재성 외, “21세기 세종계획 전자사전 개발 분과 연구 보고서”, 국립국어연구원, 2006.

[19] Manning, Christopher D. and Hinrich Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.