중한 기계 번역 시스템을 위한 형태소 분석기

†강원석, †김지현, †송영미, †송희정, †황금하, †채영숙, †최기선
†한국과학기술원, 첨단정보기술연구센터(AITrc)
†한국과학기술원, 전문용어언어공학연구센터(KORTERM)
E-mail: wskang@world.kaist.ac.kr

A Morph Analyzer For MATES/CK

†Won-Seok Kang, †Ji-Hyoun Kim, †Young-Mi Song
†Hee-Jung Song, †Jin-Xia Huang, †Young-Soog Chae, †Key-Sun Choi
†AITrc, KAIST
†KORTERM, KAIST

요약

MATES/CK는 기계번역 시스템에서 전통적으로 사용하고 있는 세 단계(분석/변환/생성)에 의해서 중한 번역을 수행하는 시스템이다 MATES/CK는 시스템 성능을 높이기 위해 패턴 기반과 통계적 정보를 이용한다. 태거(Tagger)는 중국어 단어 분리를 최장일치법으로 수행하기 때문에 일부 단어에 대해 오류를 범하게 되고 품사(POS : Part Of Speech) 태깅 시 확률적 정보만 이용하여 특정 단어가 다 품사인 경우 그 단어에 대해 특정 품사만 태깅되는 문제점이 발생한다. 또한 중국어 및 외국어 인명 및 지명에 대한 미등록들에 대해서도 올바른 결과를 도출하지 못한다 사전에 있어서 텍스트 기반으로 존재하여 이를 관리하기에 힘이 든다.

본 논문에서는 단어 분리 오류 및 품사 태깅 오류를 해결하기 위해 중국어 태깅 제약 규칙을 적용하는 방법을 제시하고 중국어 및 외국어 인명/지명에 대한 미등록어 처리방법을 제시한다 또한 중국어 사전 관리에 대해 알아본다.

1. 서론

1)자연언어는 각 나라마다의 독특한 언어적 특성으로 인해 기계적으로 처리하기에는 매우 모호하고 무제한적인 언어 현상을 나타낸다. 이러한 자연언어 현상을 기계적으로 처리하는 기계번역 시스템은 전통적으로 사용하고 있는 세 단계(분석/변환/생성)에 의해서 번역을 수행한다. 기계번역 시스템은 특수한 언어적 현상을 해결하기 위해 규칙, 예제, 패턴 기반 및 통계적 기반 등 여러 가지 방법들을 적용해왔다. 중한 기계번역 시스템(MATES/CK)도 위의 세 단계의 방법을 적용하고 있다.

중국어는 공백으로 분리되는 영어나 한글과 달리 단어를 분리할 기준이 명확하지 않고 하나의 단어가 하나 또는 여러 단어로 구성되는 특징을 가진다. 이러한 이유로 중국어 분석에 있어서 단어 분리는 어려운 일이다. 이러한 문제점을 해결하기 위해 많은 연구가 이루어졌다[1]. 그리고 품사 태깅에 있어 다 품사가 존재하는 경우 그 단어의 모호성으로 인해 정확한 태깅이 어렵다[2]. 중국에서 표현되는 중국어 및 외국어 인명 및 지명은 무한 집합이고 절대 부분 사전 미등록어이기에 형태소 분석 및 구문 분석에서 많은 어려움이 있다.

기존 MATES/CK에서 분석 단계에 해당되는 태거(Tagger)는 중국어 단어 분리를 최장일치법으로 수행하기 때문에 일부 단어에 대해 오류를 범하게 되고 품사(POS : Part Of Speech) 태깅 시 확률적 정보만 이용하여 특정 단어가 다 품사인 경우 그 단어에 대해 특정 품사만 태깅되는 문제점이 발생한다. 또한 중국어 및 외국어 인명 및 지명에 대한 미

1) 본 연구는 첨단정보기술연구센터 과제 "다국어 정보검색 연구"와 전문용어언어공학연구센터 과제 "중국어-한국어 경험을 통한 번역지식의 획득"을 통하여 과학재단의 지원을 받았음

등록들에 대해서도 올바른 결과를 도출하지 못한다.

본 논문에서는 MATES/CK에서 나타나는 태거의 문제점을 해결하기 위해 태깅 제약 규칙과 통계적 방법을 접목한 방법을 제시하고 고유명사 인식 및 명사 추출기를 포함한다.

본 논문에서는 2장에서 MATES/CK에서 사용되는 중국어 사전 관리 방법에 대해 설명하고 3장에서 태거에 대해 설명하고 4장에서 실험 결과 및 평가에 대해 설명한다. 마지막 5장에서는 결론 및 향후 계획에 대해 설명한다.

2. 사전 구성

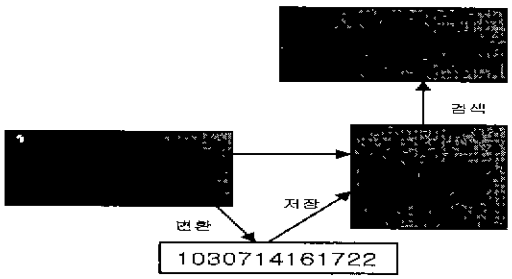
자연언어 처리에 있어서 사전은 시스템의 성능에 중요한 역할을 하는 한 요소이다. 이렇기에 사전 관리가 무엇보다 중요하다. 이러한 사전 관리의 중요성 때문에 사전 관리에 관한 많은 연구가 있었다. 본 시스템에서는 기존에 개발된 사전 관리 구조 TDBM와 SIMTI를 이용한다. TDBM는 Disk 기반으로 많은 데이터를 저장할 수 있는 반면 속도 면에서는 디스크 I/O로 인한 오버헤드가 발생한다. SIMTI는 메모리 기반 저장 구조로 속도 면에서는 빠르나 데이터가 많은 경우 문제가 발생한다. 2.1절에서는 TDBM을 이용한 중국어 사전에 대해 설명하고 2.2절에 SIMTI를 이용한 중국어 사전에 대해 설명한다

2.1 TDBM을 이용한 사전

TDBM은 고유한 Key와 이에 대한 데이터를 구성하며 사전을 관리한다. 이 구조는 Key와 데이터가 모두 스트링 기반으로 구성되어 있다. 본 시스템은 중국어 사전을 관리하기 위해 세 가지 사전으로 구성한다. 첫 번째, 중국어 단어 중 어떤 첫 음절로 시작되는 단어의 최장 길이 정보를 저장하는 사전, 두 번째, 중국어 단어에 대한 품사 정보를 저장하는 사전, 세 번째, 중국어단어와 그에 대한 품사에 기반한 특성 정보 저장 사전이다 두 번째 사전의 Key는 표제어이고 데이터는 이 단어가 가질 수 있는 품사들의 값이다. 세 번째 사전의 Key는 표제어 + POS(Part Of Speech)"로 구성되고(Key = "표제어 + POS") 데이터는 이에 해당되는 특성들을 나타낸다.

2.2 SIMTI를 이용한 사전

SIMTI는 앞에서 언급한 TDBM과 같이 고유한 Key와 이에 대한 데이터로 구성되어 있으나 내부 저장 데이터 형식과 용량이 조금 다르다. SIMTI는 Key는 스트링 기반으로 하는 반면 메모리 기반 특성으로 데이터 저장 공간은 제약을 두고 있다. 그래서 SIMTI는 중한 기계번역에서 사용하지 않고 중국

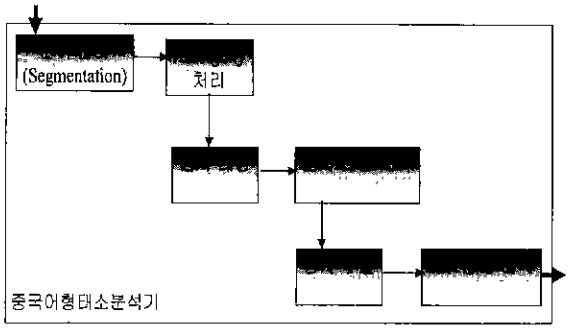


(그림 1) 중국어 사전 저장 구조

어 명사 추출기 시스템(A Noun Retrieval System For Chinese : NRSC)에서 이용하였다. NRSC는 TDBM으로 구성한 사전 중 첫 번째와 두 번째 사전과 같은 형태로 SIMTI로 저장된 사전을 이용한다. 기존 SIMTI는 데이터를 2bytes or 4bytes구성되어 있으나 NRSC에서는 8bytes 데이터를 확장하여 이를 이용하였다. NRSC에서 이용되는 두 번째 사전은 다음과 같이 구성한다. Key는 TDBM과 같이 표제어와 POS로 구성되며 데이터에 해당되는 품사들의 정보는 10진 숫자로 Encoding되어 저장된다. (그림 1)은 저장과 변환 구조를 나타낸다.

3. 중국어 태거

MATES/CK에서 Tagger는 최장일치법에 의한 단어를 분리하고 이들에 대한 품사 태깅은 HMM(Hidden Markov Model)([3])에 의한 방법으로 태깅한다. 태깅되는 품사들이 통계적 정보에 따르기 때문에 중국어 단어가 다 품사인 경우 특정 품사만이 부여되는 경우가 발생한다. 이런 문제점을 보완하기 위해 MATES/CK에서는 품사 태깅 제약 규칙을 적용하여 통계적 정보의 문제점을 보완한다. 그리고 중국어 고유명사 인식 법을 적용한다. (그림 2)는 중국어 형태소 분석 시스템을 나타낸다. 3.1절에서는 품사 태깅 제약 규칙의 구성에 대해 설명하고 3.2절에서는 태깅된 결과를 가지고 명사 추출 형태에 대해 설명하고 3.3절에서는 고유명사 인식(중국어 인명/지명 및 외국어 인명/지명)에 대해서 설명한다. (사용되는 품사 : ***Appendix A***)



(그림 2) 중국어 형태소 분석 시스템

3.1 품사 태깅 제약 규칙 구성

Tagger에서 사용되는 태깅 제약 규칙은 (그림 3)과 같은 기본 형태로 구성된다. (그림 3)과 같이 기본 구성 요소는 7가지 필드로 구성된다. 표제어는 다 품사로 구성된 중국어 단어이고 POS는 다 품사

Word: (표제어)
POS:(품사)
OP: (YES or NO)
Left:
NotLeft:(생략가능)
Right:
NotRight:(생략가능)

(그림 3) 태깅 제약 규칙 정의 형태

중 규칙에 의해 태깅될 품사를 나타낸다 그 밑으로 있는 필드들이 규칙에 해당된다. OP는 단어 A와 단어 B(A+B)로 이루어진 C라는 단어가 있으면 이 단어를 A와 B로 분리 할 것인지를 판단하는데 사용된다. Left와 Right 필드는 현 단어 왼쪽 또는 오른쪽에 존재해야 할 품사 및 단어들을 나타내고 NotLeft와 NotRight는 존재하지 않아야 할 품사 및 단어들을 나타낸다.

Word: 在(재)
POS: d
Left. NULL
Right: v

(그림 4) 在자의 품사 d에 대한 제약 규칙

Word: 在
POS: v
Left:NULL
Right:s,n,f|v

(그림 5) 在자의 품사 v에 대한 제약 규칙

Word: 在
POS: p
Left:NULL
Right:s,n,f,t,NULL+v,s,f

(그림 6) 在자의 품사 p에 대한 제약 규칙

(그림 4)에서 (그림 6)은 在에 대한 품사 태깅 제약 조건을 나타낸 예이다. (그림 4)는 在자 품사가 d로 태깅되기 위해 왼쪽에는 어떤 단어나 품사들이 존재해도 되면 오른쪽에는 반드시 품사 v를 가지는 단어가 존재해야된다는 것을 나타낸다. (그림 5)에서 "|"로 구별된 품사가 있는데 이는 在의 오른쪽에 있는 단어의 오른쪽에 존재하지 않아야 할 품사 및 단어를 나타낸다. 이에 在자가 품사 v로 태깅되기 위해 오른쪽에는 품사 s, n, f를 가지는 단어가 존재해야되고 오른쪽의 오른쪽 단어에는 품사 v가 존재하지 않아야 한다는 것을 나타낸다. (그림 6)에서 "+"로 구별된 품사가 있는데 이는 在의 오른쪽에 있는 단어의 오른쪽에 존재해야하는 단어 및 품사를 나타낸다. 그래서 在자 품사 p로 태깅되기 위해 오른쪽에는 품사 s, n, f, t를 가지는 단어가 존재해야되고 오른쪽의 오른쪽 단어에는 품사 v, s, f가 존재해야 된다는 것을 나타낸다.

Word: 把門(파문)
POS: v
OP:YES-2-2
Left:NULL
Right:NULL
NotRight:u,y

(그림 7) 把門에 대한 단어 분리 규칙

(그림 7)은 앞에서 언급한 것과 같이 두 단어로 구성된 한 단어에 대해 분리할 것인지 그렇지 않을 것인지를 판단할 때 쓰는 제약 조건에 대한 예이다. 여기에 OP필드에 OP:YES로 적으면 표제어가 정의된 품사 조건에 맞으면 A(把)와 B(門)로 분리되고 그렇지 않으면 하나의 단어로 적용된다. 만약 OP필드가 OP:NO로 정의되면 표제어는 정의된 품사 조건이 맞으면 분리되지 않고 그렇지 않으면 A와 B로 분리된다. "OP:YES-2-2"에서 첫 번째 숫자는 분리 될 위치를 나타내고 두 번째 숫자는 분리되어서 앞 뒤 단어들과의 결합 방식에 대해 나타낸다. 단어

결합 방식은 다음과 같다.

숫자 1 : 단어를 분리한 후 앞 단어와 결합을 함
숫자 2 : 단어를 분리한 후 뒤 단어와 결합을 함
숫자 3 : 단어를 분리만 함

만약 A+B-C로 분리된 문장이 있다고 하자. 이를 바탕으로 B가 D+E로 분리되는 제약이 적용된다고 가정하자. 여기서 위에서 설명한 두 번째 숫자가 '1'이 되면 최종 문장은 AD+E+C로 되고 '2'가 되면 A+D+EC가 되고 '3'이 되면 A+D+E-C가 된다.

3.2 명사 추출기

명사 추출기는 앞에서 언급한 것과 같이 태거의 한 모듈로써 태깅된 문장 결과들에 대해서 명사만을 추출한다. 사용되는 사전은 2.2절에서 언급한 SIMTI 버전으로 구성된 것을 이용한다. 아래는 명사로 추출될 후보에 대한 예를 나타낸다.

1) n+u+v : 색인어 NP로 추출
   ex) 世紀(세기)/n 之(지)/u 交(교)/v
2) m+a+n : 색인어 NP로 추출
   ex) 第三(제삼)/m 大(대)/a 産業(산업)/n
3) b+n : 색인어 NP로 추출
   ex) 初級(초급)/b 階段(계단)/n
4) 'm'을 단독으로도 색인어로 인정하지만,
   m+m으로 나타나는 경우는 합친 경우만 인정
   ex) 2700/m 億(억)/m
5) m+q를 색인어로 인정하지만, m+m+q인 경우, m+q는 인정하지 않음
   ex) 3.3/m 萬(만)/m 名(명)/q
6) 'j' 색인어로 인정
   ex) 中外(중외)/j
7) n+n+…+n인 경우 순차적 조합 가능 경우 모두 추출
   ex) 經濟(경제)/n 性質(성질)/n 我國(아국)/n 軟件(연건)/n 企業(기업)/n

3.3 고유명사 인식 - 중국어 및 외국어 인명/지명

중국어 및 외국어 인명/지명은 무한 집합이고 절대 부분 사전 미등록어이기에 형태소 분석에 어려움을 더해주며 구문분석 실패의 원인으로 되기도 한다[4] 중국인 성씨 사전에 포함된 성씨 빈도 정보 및 이런 상용 성씨 문자를 시작으로 하는 문자열의 주어진 텍스트에서의 출현 빈도를 이용하여 중국인 인명 인식을 진행하였으며 중국인 인명 뒤에 따라오는 단어도 참고 정보로 사용하였다[5]. 그러나 그의 방법은 주어진 번역문이 단일 문장일 경우 적용하기 어려운 단점을 가지고 있다.

중국 지명에 대한 인식은 우선 중국 지명 사전 및 태깅된 말뭉치로부터 중국 지명에서 자주 나타나는 문자 집합 및 이들의 출현 빈도를 추출하였으며 다음 지명 인식에서는 이런 지명 상용 문자 출현 빈도를 이용하여 지명 후보를 얻은 후 문장에서의 좌우 문맥정보를 이용하여 휴리스틱 기반의 방법으로 지명 인식을 시도하였다[6].

본 논문에서 제시하는 인명/지명 인식기는 아래와 같은 단계를 거쳐 휴리스틱 결정(Heuristic decision)에 의해 수행되어진다.

Step1. 후보 집합 결정

- 후보에 해당되는 문자들에 대해 점수를 부여한다. 점수 부여 대상은 아래와 같은 경우에 적용된다.

1) 특정 동사 앞에 나오는 모든 단어들로 이들 단어들은 모두 한 문자로 구성됨
   [Ex : 訪問(방문하다), 會見(회견하다) 등]
2) 특정 동사 뒤에 나오는 모든 단어들로 이들 단어들은 모두 한 문자로 구성됨
3) 문장 부호 ":" 앞에 나오는 모든 단어들로 이들 단어들은 모두 한 문자로 구성됨
4) 특정 명사 앞 또는 뒤에 나오는 모든 단어들로 이들 단어들은 모두 한 문자로 구성됨
   [Ex : 主席(주석), 國王(국왕) 등]
5) 연속해서 두 개 이상의 단어들이 모두 한 문자로 이루어진 경우
6) 연속해서 두 개 이상의 단어들 중에 어소(g), 접미사(k), 약어(j) 및 명사(n)를 포함한 경우

Step2. 최종 후보

- 후보 집합 결정 단계에서 두 가지 이상인 경우의 집합에 대해서 최종 후보를 추출한다

1) 중국어 인명/지명 최종 후보 선택

- 중국어 인명

조건) 중국어 인명 문자수는 최대 4 이상을 넘지 않는다.

결정)

· 후보 첫 번째 문자가 중국어 성씨 사전에 포함되면 이를 중국어 인명으로 인식(한 문자로 구성된

중국어 성씨)

· 후보 첫 번째, 두 번째 문자의 결합 문자가 중국어 성씨 사전에 포함되면 이를 중국어 인명으로 인식(두 문자로 구성된 중국어 성씨)

- 중국어 지명

결정) 후보 마지막 문자가 중국어 지명 패턴에 포함되면 이를 지명으로 인식

2) 외국어 인명/지명 최종 후보 선택

조건) 후보 집합들 중에서 문자들이 외국어 인명/지명 문자 셋에 포함되는 경우가 두 문자 이상 이여야 한다

결정)

· 후보 마지막 문자의 품사가 조사(u)이고 이 문자가 외국인 인명/지명 문자 셋에 포함되면 이 문자를 포함하여 인식, 그렇지 않고 외국인 인명/지명 문자 셋에 포함되지 않으면 이 문자는 배제함

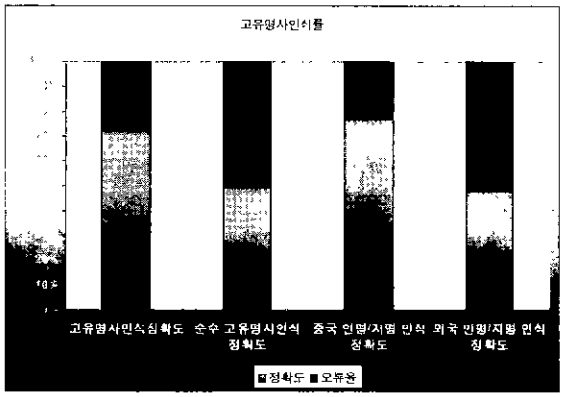
· 후보 마지막 문자가 중국어 상용어휘 문자 셋에 포함되지 않으면 이 문자를 포함하여 인식, 그렇지 않고 외국인 인명/지명 문자 셋에 포함되지 않으면 이 문자는 배제함

4. 실험 결과 및 평가

본 시스템의 평가 환경은 컴파일러로 Visual C++ 6.0 사용하였고 테스트 환경으로 Windows'2000에서 수행하였다. 구축된 태거의 성능을 알아보기 위해 다음과 같은 방법으로 성능을 평가한다.

태거의 성능을 평가하기 위해 테스트 중국어 문장은 태거에서 사용되는 HMM에 대한 확률 값을 얻기 위해 구축된 약 일만 문장의 POS Tagged Tree 코퍼스(약 7만 단어)를 사용한다. 테스트 결과 단어 분리에 관한 정확도는 ***약 92%***의 성능을 보이고 품사 태깅에 관한 정확도는 ***약 82%***의 성능을 보인다

(그림 8)은 본 논문에서 제시한 고유명사 인식에 대한 성능 평가를 나타낸 것이다. 고유명사 인식에서 사용한 평가 셋은 중국 인민 일보 타이틀 100문장을 사용하였다. 여기에서 고유명사 정확도는 100문장 중에 나타난 전체 고유명사 들 중 자동으로 추출한 결과를 나타낸 것이고 순수 고유명사 인식 정확도는 중국어 사전에 등록되지 않은 단어들에 대한 정확도이다. 그리고 중국 인명/지명 인식 정확도는 전체 고유명사들 중 중국 인명/지명에 해당되는 고유명사들에 대한 정확도이고 외국 인명/지명 인식 정확도는 고유명사들 중 외국 인명/지명에 해당되는 고유명사들에 대한 정확도이다.



(그림 8) 인명/지명 명사 인식 정확도

5. 결론 및 향후계획

MATES/CK는 중국어를 한국어로 변환하는 하나의 중한 기계번역 시스템이다. 시스템에서 사용되는 사전를 관리하기 용의하도록 KAIST에서 개발된 TDBM과 SIMTI를 이용한 사진을 구성하였다.

기존 MATES/CK에서 분석 단계에 해당되는 태거(Tagger)는 중국어 단어 분리를 최장일치법으로 수행하기 때문에 일부 단어에 대해 오류를 범하게된고 품사(POS : Part Of Speech) 태깅 시 확률적 정보만 이용하여 특정 단어가 다 품사인 경우 그 단어에 대해 특정 품사만 태깅되는 문제점이 발생한다 이를 해결하기 위해 본 논문에서는 태깅 제약 방법을 적용하는 방법을 제시했다. 그러나 현재 태깅 제약 조건이 그리 많이 구축되지 않아 현재 시스템에서 아주 좋은 결과를 기대하기는 어려울 것이고 앞으로 이에 대한 조건들을 계속 추가하여야 할 것이다. 또한 중국어 및 외국어 인명 및 지명에 대한 미등록돌에 대해서도 올바른 결과를 도출하지 못한다. 이를 해결하기 위해 본 논문에서는 고유명사 인식 및 명사 추출 방법을 제시했다. 그리고 본 논문에서 제시한 방법에서 해결되지 않은 고유명사들에 대해서는 사전 추가 작업과 문의 의미 정보를 이용한 인식 방법을 수행해야 할 것이다.

참고문헌

[1] Yubin Dai and Teck Ee Loh, 1999, "A New Statistical Formula for Chinese Text Segmentation Incorporating Contextual Information", SIGIR'99, pp.82-89.

[2] Lluis Padro and Luis Marquez, 1998, "On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpara.", COLING-ACL'98, pp.997-1002.
[3] 황도삼, 최기선, 김태석, "자연언어처리", 홍릉과학출판사. 1998.
[4] 황금하, 송희정, 김지현, 송영미, 강원석, 서충원, 채영숙, 최기선, 2000, "뉴스타이틀 번역을 위한 중한 기계번역 시스템", 제12회 한글 및 한국어 정보처리 학술대회
[5] Wang Xing, Huang Degen, Yang Yuansheng, 1999, "Identifying Chinese Names based on Combination of Statistics and Rules", in proceeding of JSCL-99, pp.155-161, Chinese.
[6] Tan Honggye, Zheng Jiaheng, Liu Kaiying, 1999, "Research of the Method of Automatic Recognition of Chinese Place", in proceeding of JSCL-99. pp.174-179, Chinese.

Appendix (A)

<중국어 품사 및 구 정보>

품사 표기는 북경대학교 정보 처리용 현대 중국어 분류 체계의 품사 표기집에 의거한다.

[품사]	[Tag]
형용사	a
구별사	b
접속사	c
부사	d
감탄사	e
방위사	f
어소	g
접두사	h
성어	i
약어	j
접미사	k
관용어	l
수사	m
명사	n
의성어	o
개사	p
양사	q
대명사	r
처소사	s
시간사	t
조사	u
동사	v
문장부호	w
비어소자	x
어기사	y
의태어	z

Cf) ap, dp, mp, np, pp, sp, tp, vp, 등은 각각 형용사구, 부사구, 수량사구, 전치사구, 처소구, 시간사구, 동사구를 나타낸다.

Cf) dj, fj, yj, zj 등은 각각 단문, 복문, 인용문, 완전한 문장을 나타낸다.