한국어 문장의 종결유형에 따른 문장형태 단일화에 의한 자연어처리 (lemmatization & stemming of sentence-final ending in Korean language)

1. 소프트웨어 소개

한국어에서는 하나의 용언(동사,형용사에서도 동사 변화가 수십개에 이릅니다. 이러한 용언의 변화를 단일화 시켜 문장을 쉽게 이해할 수 있도록 하는 한국어 자연어처리 시스템입니다.

2. 개요 및 개발 동기

최근 인공지능 비서, 스피커, 그리고 챗봇 관련 기술이 발전하면서 사용자의 말, 의도하는 바를 더 정확하게 이해할 수 있는 자연어처리 기술의 필요와 수요가 늘고 있습니다.
팀 모도리에서는 한국어가 가지고 있는 특성이 자연어처리 기술로서 큰 장점이 될 수 있다고 판단하고, 자연어 처리 시스템 개발을 진행하였으며, 한국어 종결어미 유형별 문장 형태 분석 시스템 개발을 진행하여, 단순화된 체계를 구축할 수 있었습니다.
자연어처리 분야에 있어서, 문장의 종결어미를 단순화함으로써 얻은 효용을 모도리를 통해 소개하고자 합니다.

3. 기능 소개

3-1. 종결어미 단일화

1) 용언(동사, 형용사)의 어미변화 단일기준화.
2) 2개 이상으로 이루어진 문장 종결 용언의 단축 단일화
3) 문장의 형태별 유형 파악:
   가: 문장 형태: 의문문, 평서문, 명령문 (청유문은 종결유형 자체에 포함 시킴)
   나: 시제(tense): 과거형, 현재형, 미래형
   다: 용언종결유형(10가지) 분류
      0:일반, 1:의지(하겠다), 2:명령형, 3: 청유형(하자), 4:의무(해야한다)
      5:추정(일까), 6:경험(해보다), 8:예정(하려한다), 9:희망(싶다), 13:간접인용(있단다)

예시: 평서문: “나는 너를 좋아하고 있다.”
      의문문: “너는 나를 좋아하냐?”

형태(질문,Mod,Tens,	기준문장	대답의 예	시제:tense 부정문 여부
너는 나를 좋아하느냐? senType= Q 0	나ZL좋아하느냐	궁금한가요. 좋아합니다.	현재(5),
나를 좋아했었냐 sentType= Q 0	나ZL좋아하느냐	한때 좋아하기도 했었죠.	과거(3)
너는 나를 좋아하지 않냐 sentType= Q 0	나ZL좋아하느냐	글쎄요. 그런 거 같나요.	현재(5). 부정문(F)
너는 나를 좋아하지 않았냐 sentType= Q 0	나ZL좋아하느냐	글쎄 표시를 안 냈네요.	과거(3), 부정문(F)
너는 나를 좋아할 거냐 sentType= Q 0	나ZL좋아하느냐	글쎄요. 좋은 감정은 가지고 있을 거예요	미래(7)
나는 너를 좋아하지 않아 senType = 0 0	너ZL좋아한다	좋아하지 않는다고요?	현재(5), 부정문(F)
나는 너를 좋아하지 않았어 senType = 0 0	너ZL좋아한다	그럴 리가 없어요.	과거(3), 부정문(F)
나는 너를 좋아하지 않을 거야 senType = 0 0	너ZL좋아한다	글쎄요 그래도 기대할게요	미래(7), 부정문(F)
너를 좋아하고 싶은데. sentType= 0 9(희망)	너ZL좋아한다.	조금씩 가까워질 수 있을 거예요.	현재(5)
너를 좋아하고 싶었는데. sentType= 0 9(희망)	너ZL좋아한다	지나간 일이었나요	과거(3)
너를 좋아하고 있는데 sentType= 0 0	너ZL좋아한다	좋아하는군요.	현재(5)
너를 좋아했었다. sentType= 0 0	너ZL좋아한다	좋아했군요.	과거(3)
….			

참고: '너를'에서 '를'은 '을', '를' 2가지를 단일화 시키기 위해 'ZL'로 표기하고, 공란을 없앴음.

여기서 용언 종결유형을 10가지로 나누었는데 이것은 유동적일 수도 있습니다.
용언 종결유형을 10가지로 나누었을 때,
문장의 종류는 크게 의문문, 평서문 (명령문, 청유문은 별도 처리) 2가지가 되고
문장의 시제도 현재, 과거, 미래형으로 3가지로만 나누어 보았습니다.
부정문을 구분하면 부정문, 일반문으로 2가지가 됩니다.
이론적으로 전체를 계산해보면 문장의 종결유형은 명령문을 제외해도 120가지에 이르게 됩니다.
이것은 용언의 종결어미 변화의 숫자를 제외하고도 그렇습니다.

이것은 다음과 같이 기준 문장 2개로 단일화하여 다음과 같이 처리할 수 있습니다.
의문문: "(너는) 나ZL좋아하느냐"
평서문: "(나는) 너ZL좋아한다"
(참고: 사람이 주어인 경우에 주어는 변화의 종류에서 제외합니다)

이것을 처리하는 프로그램에서 종결어미 특성을 형태별로 태깅할 수 있도록 체계를 구축하였습니다.
문장의 끝에 동사가 2개 이상 연속해서 나올 때에는 종결어미를 단순화 시키고, 동사가 한 번만 나올 때는 동사의 기본형을 사용합니다.
이렇게 기준문장을 단순화시키면, 문장을 유형별로 파악해서 태깅해야 하는 기능이 필요합니다.
문장의 유형 태깅은 다음과 같이 4가지입니다.
의문문:Q, 시제: 3,5,7 문장종결유형:0,1,2,3,4,5,6,8,9,13, 부정문 구분:F,U,P
단형부정문 '안'은 종결어미가 아니므로 별도 처리합니다.

4. 가능성과 활용

한국어는 용언의 활용이 매우 다양하여 하나의 기본형에서도 수십개의 동사변화가 나타날 수 있습니다. 앞서 보신 것과 같이 문장의 종결형태를 단순화 시킬 수 있다면, 무제한적인 용언의 다양한 변화에 대해서도 어렵지 않게 처리할 수 있습니다.
현재 자연어처리 분야에서 단순한 문장을 처리하는 방법에 대해서는 많은 연구와 발전이 있었기 때문에, 단순한 단일 기준의 문장을 처리하는 것은 기존의 많은 연구 결과와 함께 적용하여 활용할 수 있는 것도 큰 장점입니다.
문장의 종결유형을 단순화하는 것을 프로그램으로 처리하고, 용언종결유형 4가지를 태깅하는 것도 프로그램으로 동시에 처리하도록 시스템을 구축하였습니다.

대표적인 것으로 부정문을 문장의 구조와는 별도로 구분하여 알아낼 수도 있도록 하는 것입니다. 구조적으로는 기본이 되는 기준문장으로 단일화 시키고, 문장이 부정문이라는 것을 인식하여 적용하는 것입니다. 이것은 많은 데이터를 하는데 있어서도 큰 도움을 줄 수 있을 것입니다.
인공지능 대화에서도 이러한 종결어미형태를 특성화하는 것은 감성적인 면에서도 다양한 느낌을 표현할 수 있도록 할 것입니다.
아래와 같이 종결어미 60가지의 유형별 대화를 한 페이지에서 정리하는 데이터베이스 화면을 구축할 수도 있습니다.