Ch4 – 텍스트의 전처리

비정형 데이터 : 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 그림이나, 영상, 문서처럼 형태와 구조가 다른 구조화되지 않은 데이터. 변칙과 모호함이 발생해, 데이터베이스의 칸 형식의 폼에 저장되거나 문서에 주석화 된 데이터에 비해 전통적이 프로그램을 사용하여 이해하는 것을 불가능하게 한다.

비정형 데이터의 전처리 : 분석의 위해 raw data를 정형화 하는 과정

텍스트는 비정형 데이터이므로, 아무런 전처리 없이 사용하기에는 힘들다.

정형 데이터	비정형 데이터
행, 열 및 관계형 데이터 형식으로 표시가 가능	행, 열 및 관계형 데이터 형식으로 표시가 불가능
숫자, 날짜, Json 등의 파일형식	시각정보, 음성정보, Raw 텍스트 등
전체 데이터의 약 20%가 정형화 데이터	전체데이터의 약 805가 비정형화 데이터
차지하는 용량이 비교적 적다	차지하는 용량이 비교적 많다
전통적인 방법을 이용한 데이터의 수정 및 사용이 용이하다	전통적인 방법을 이용한 데이터의 수정 및 사용이 용이하지 않다

*정형 데이터와 비정형 데이터의 차이

#2 텍스트 문서의 변환

텍스트 문서를 다룰 때 우선 파일로부터 텍스트를 추출하는 것이 전처리의 첫 번째 단계에 해당된다.

문서들은 docs, hwp, html, pdf 등 사람이 읽기 편한 형식으로 저장되어 있어 컴퓨터에게 불편하다. 각각 파일에 따라 저장된 방법이 다르기 때문

예를 들어 문서 "Welcome to the world of Natural Language Processing"이라는 문장을 크롤링하게 되면, 파일형식에 따라 다르다.

*HTML파일

<h2 style="color: #2e6c80;"> Welcome to the world of Natural Language Processing.</h2>

*PDF파일

Welcome to the world of Natural Language Processing.

그 외에 몇몇 파일 형식은 인코딩을 어떻게 풀어내는가에 따라 읽을 수 없게 될 수도 있다.

이러한 문제 해결을 위해 문서 파일 ➔ 문서 하는 작업을 수행해야한다. 이를 위해 텍스트 내의 입력 문자열은 오로지 목표어휘언어의 문자만 남아있어야 한다. 즉, 여러 특수문자 및 불필요한 타 언어 문자의 제거가 요구. 1차적으론 특수문자를 제거, 2차적으로는 문장과 관련이 없는 특수 커맨드 또는 코딩을 제거하는 방식이다. PDF의 경우에는 텍스트를 오로지 문장 단위로 끊게 해, 마침표만이 문장의 끝으로 인식하는 방식도 있다.

#3 띄어쓰기 교정 방법

띄어쓰기는 단어의 의미 분할 및 전달과 함께 매우 중요하다. 특히 한국어는 띄어쓰기에 따라 의미가 여러 개로 해석될 수 있으므로 주의해야한다. 따라서 텍스트 전처리에서 띄어쓰기 교정은 중의성을 해소하는 매우 중요한 작업이다.

한국어 띄어쓰기 교정은 크게 3가지로 나뉜다. 규칙기반 기법, 통계기반 기법, 그리고 딥러닝 기반 기법이 있다. 이중 딥러닝 기반 기법은 이후 책의 뒷부분에 나옵니다.

3-1. 규칙기반 띄어쓰기 교정 기법

형태소 분석기를 사용하는 규칙기반의 분석적인 방법이 있다. 이때 규칙은 주로 어휘지식, 규칙, 오류유형 등의 휴리스틱 규칙을 이용한다.

장점: 상정내에 있는 상황에서 규칙이 있는 경우에는 100%에 가까운 정확도를 보여준다.

단점: 여러 가지 언어학적 자원을 만들어야 하고 여러 단계의 복잡한 휴리스틱을 적용해야 하기 때문에 비교적 분석과정이 복잡하고 어휘지식 구축관리에 비용이 든다. 또한, 모든 상황에서 사용하기 위해선 모든 경우의 수를 생각해야 하므로 실질적으로 불가능하다.

규칙기반의 휴리스틱을 사용하는 자동 띄어쓰기 방법으로 어절 블록 양방향 알고리즘이 있다.

이 방법에서 어절 블록 인식, 어절 블록 내의 어절 인식, 어절 인식 오류 교정의 세단계를 거쳐 자동 띄어쓰기를 한다.

예시)

규칙

들어가셨다 -> 들어 + space + 가셨다 ➔ 아버지가방에 들어가셨다 -> 아버지가방에 들어 가셨다.

하지만, 문법적인 부분만 고려하도록 한 규칙기반의 경우, 어느 쪽 결과가 더 좋은 지 판별할 수 없기 때문에 의미적으로 중의성을 띄는 구간이나 오타에는 그 성능이 크게 약해진다.

규칙

가방-> 가+space+방 아버지가방에 들어가셨다. -> 아버지가 방에 들어 가셨다.

Non_space+가방-> space+가방 아버지가방에 들어가셨다. -> 아버지 가방에 들어 가셨다.

*규칙기반의 한계

3-2. 통계, 확률기반 띄어쓰기 교정 기법

말뭉치로부터 자동 추출된 음절 n-gram 정보를 기반으로 기계적인 계산 과정을 거쳐 띄어쓰기 오류를 교정한다.

장점: 구현이 더 용이하며, 어휘 지식 구축관리 및 미등록어에 대해서는 견고한 분석이 가능하다

단점: 그만큼 학습 말뭉치의 영향을 크게 받아, 정확도 및 오류율이 높은 경우가 많고, 정확도 개선을 위해선 대량의 학습 데이터를 요구한다.

한국어의 경우에는 통계 확률적인 방법에서 신뢰할만한 n-gram 정보를 얻기 위해 띄어쓰기가 올바른 대용량의 학습 데이터, 즉 학습 말뭉치를 구하기 어렵다. 웹에서는 띄어쓰기가 올바른 문서가 많지 않고, 우리나라의 경우 예외사항도 있기 때문에 더 힘들다.

언어 모델링 방법을 사용하여 학습 말뭉치 내에서 옳은 수정 방향을 확률이 높은 후보들을 차례로 나열하며 이 중 가장 확률이 높은 후보로 교정을 수행한다.

#4 철자 및 맞춤법 교정방법

철자교정

–정확한 의미전송 및 정보교환을 위해 필요 ➔ 의미혼용의 방지 및 정보전달의 실패를 방지하기 위해 필요

-이를 위해 시행하는 것이 '맞춤법 검사'

맞춤법 및 철자 교정기의 두 가지 수행 역할

- 텍스트 내 오류 감지
- 오류의 수정

오타로 인해 발생할 수 있는 오류

- 삽입(Insertion): 추가적으로 문자를 입력하는 오류
- 생략(Deletion): 본래 있어야 하는 문자를 생략하는 오류
- 대체(Substitution): 본래 넣어야 할 문자 대신 타 문자를 대입하는 오류
- 순열(Transposition): 철자 순서를 뒤바꾼 오류

철자 오류 예시

Error	Correction	Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	–	2	deletion
acress	cress	–	a	0	insertion
acress	caress	c, a	a, c	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	–	s	5	insertion
acress	ares	–	c, s	1, 5	insertion

철자 및 맞춤법 교정방법도 크게 규칙기반, 통계기반, 딥러닝 기반으로 나눌 수 있다. 이중 딥러닝 기반은 책 뒷부분에 나온다.

4-1 규칙기반 맞춤법 교정 기법

언어 현상의 규칙성을 추가로 응용하는 방식이다. 어절은 어절보다 작은 단위인 형태소들이 일정한 규칙에 따라 결합하여 이루어짐. 어절을 형태소들로 분절하는 형태소 분석기를 사용한 방식이 존재한다.

장점: 상정내에 있는 상황에서 규칙이 있는 경우에는 100%에 가까운 정확도를 보여준다.

단점: 여러 가지 언어학적 자원을 만들어야 하고 여러 단계의 복잡한 휴리스틱을 적용해야 하기 때문에 비교적 분석과정이 복잡하고 어휘지식 구축관리에 비용이 든다. 또한, 모든 상황에서 사용하기 위해선 모든 경우의 수를 생각해야 하므로 실질적으로 불가능하다.

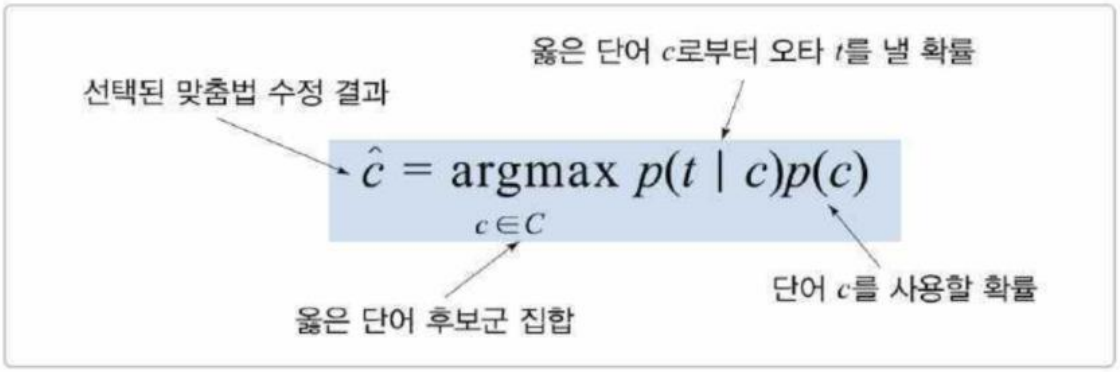
규칙기반 맞춤법 교정 과정

이 방법을 구현하기 위해서는 체언, 용언, 어미, 조사 등으로 구성된 품질이 좋은 사전과 형태소들 간의 접속 가능관계가 표현된 정밀한 접속정보표가 요구된다.

주어진 입력 어절을 성공적으로 분석할 수 없을 경우 해당 어절을 맞춤법이 틀린 어절로 간주된다. 그리고 맞춤법에 맞지 않다고 여겨지는 어절은 조사/어미 등의 문법 형태소의 음절 정보나 대규모의 말뭉치에서 얻어진 음절 간의 결합정보를 이용하여 맞춤법 교정이 이뤄진다.

4-2 통계, 확률기반 맞춤법 교정 기법

간단하게 Bayesian inference model이 있다. 이는 올바른 교정결과를 도출하기 위해 주어진 단어로부터 오타가 일어날 확률을 확률적으로 계산하는 방법이다.



이 수식을 이용해 철자 교정 확률을 관측한다. 이후 가장 확률이 높은 후보군 "actress"를 선택하여 감지한 오타를 대체하도록 하여 철자교정을 시행한다.

Error	Correction	p(t|C)p(c)	%
acress	actress	5.41×10^{-9}	54%
acress	cress	2.02×10^{-14}	0%
acress	caress	1.64×10^{-13}	0%
acress	access	1.21×10^{-11}	1%
acress	across	1.77×10^{-9}	22%
acress	acres	2.22×10^{-9}	23%

*철자 교정 예시