creative commons

COMMONS DEED

저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

단어 추출과 스트링 매치를 이용한 수기 입력 텍스트의 노이즈 처리

Cleansing noisy text using corpus extraction and string match

2013 년 2 월

서울대학교 대학원
산업공학과 데이터마이닝 전공
김 현 중

단어 추출과 스트링 매치를 이용한 수기 입력 텍스트의 노이즈 처리

Cleansing noisy text using
corpus extraction and string match

지도 교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함
2013 년 2 월

서울대학교 대학원
공과대학(산업공학과)
김 현 중

박용원의 공학석사 학위논문을 인준함
2013 년 2 월

위 원 장 _____ 박 종 헌 _____ (인)

부위원장 _____ 조 성 준 _____ (인)

위 원 _____ 윤 명 환 _____ (인)

초 록

텍스트 데이터의 경우 사전에 등록이 되어 있지 않아 인식할 수 없는 단어 역시 노이즈(out-of-vocabulary)로 정의한다. 텍스트 데이터의 노이즈는 크게 두 종류로 나눌 수 있는데, 신조어나 약어, 전문용와와 같이 실제로 사용되지만 사전에 등록이 되어 있지 않는 경우와 오탈자와 같이 의도와 다르게 기술이 되어 인식하지 못하는 경우가 있다. 노이즈를 수작업으로 정제하는 것은 비용을 초래하고, 특정 도메인의 지식(background knowledge)을 필요로 하기도 한다. 텍스트 데이터의 노이즈 처리는 모든 분석의 사전 과정으로써, 이후 분석의 성능에 영향을 준다.

본 연구에서는 효율적으로 텍스트 데이터의 노이즈를 처리하는 방법에 대하여 제안한다. 연구에서 소개하는 방법은 네 단계로 이루어진다. 첫 단계에서 음절 단위의 콜로캐이션(collocation)을 이용하여 연관성이 있는 연속된 음절을 단어로 추출한다. 이를 위하여 전방향 점수(forward score), 후방향 점수(backwoard score), 확산보정 점수(dispersion point normalized score)의 세 가지 콜로캐이션 점수 계산 방법을 제안하였다. 두 번째 단계에서는 첫 단계에서 추출된 부분음절 중에서 단어의 후보를 선택한다. 세 번째 단계에서는 사용자가 추출된 단어의 후보들 중에서 사전에 추가할 단어를 선택한 뒤, 보완된 사전을 이용하여 사전에 등록되지 않은 표현인 노이즈를 판별한다. 네 번째 단계에서는 의도와 다른 형태로 기술된 텍스트 노이즈를 복원하기 위하여 보완된 사전을 이용하여 스트링 매치(string match)를 수행한다. 이를 통하여

의도와 다르게 기술된 텍스트 노이즈를 복원한다.

전자세금계산서에 입력된 업태 데이터의 노이즈를 제거하기 위하여 제안된 방법을 적용함으로써 본 연구가 실제로 유용함을 입증하였다. 사전을 보완하기 전과 비교하여 자동으로 추출된 단어를 이용하여 40.32% 의 노이즈를 처리하였다.

주요어 : 노이즈 처리, 음절 단위의 콜로케이션, 단어 추출, 스트링 매치

학 번 : 2011-21171

목 차

제 1 장 서 론 1
제 2 장 문제 정의 및 관련 연구 7
2.1. 연구의 필요성 및 연구 범위 7
2.2. 관련 연구 8
제 3 장 제안된 연구 방법 12
3.1. 제안된 방법의 절차 12
3.2. 콜로캐이션을 이용한 부분음절 후보 추출 15
3.3. 부분음절로부터 단어 후보 추출 19
3.4. 단어 선택을 통한 사전 보완 및 노이즈 판단 21
3.5. 단어 추정을 위한 스트링 매치 21
제 4 장 실험 28
4.1. 데이터 전처리 및 변수 설정 28
4.2. 실험 결과 31
4.3. 성능 평가 35
제 5 장 결 론 37
5.1. 결론 37
5.2. 향후 연구 방향 38
참고 문헌 40
Abstract 44

그림 목차

[그림 1] 제안된 방법의 절차 13
[그림 2] 제안된 방법의 효과 13
[그림 3] 단어 후보 선정 알고리즘 20
[그림 4] Jaro distance 예시 20
[그림 5] 변형한 Jaro distance 예제 24
[그림 6] 스트링 매치 알고리즘 25
[그림 7] 전자세금계산서에 사용된 빈도수 기준 상위 50개 단어의 등장 비율 30

표 목차

[표 1] 전자세금계산서에 수기로 입력된 데이터 예시.....3
[표 2] 한국표준산업분류표 4
[표 3] 한국표준산업분류표를 이용한 업태 표현 사전.....5
[표 4] 전자세금계산서에 입력된 업태 중 상위 18개의
빈도수 및 비율.....29
[표 5] 음절 단위의 콜로캐이션으로 추출된 194개의 단어 후보32
[표 6] 최종적으로 선정한 149개 업태.....33
[표 7] 보완된 사전의 172개 단어.....34
[표 8] 사전을 보완하기 전과 후의 노이즈 처리 성능 비교
[단어 개수].....36
[표 9] 사전을 보완하기 전과 후의 노이즈 처리 성능 비교
[출현 빈도 비율] 36

제 1 장 서 론

통계학, 기계학습, 데이터마이닝 등에서 자주 사용되는 데이터는 크게 숫자형 데이터와 텍스트(문자형 데이터)로 나눌 수 있다. 숫자형 데이터에 비해 텍스트는 상대적으로 분석 과정이 복잡하여 텍스트 마이닝이라 불리기도 하며, 특히 대상을 구분하기 위하여 명목적인 단어로 사용되는 문자가 아니라 언어로서의 의미를 지닌 문자의 경우 형태소 분석, 동의어 처리 등과 같은 언어학적 지식이 분석 과정에 사용되며, 이러한 분석 과정은 자연어처리(Natural Language Processing) 분야에서 많은 연구가 진행되고 있다.

최근 많은 연구가 이루어지고 있는 트위터(Twitter), 페이스북(Facebook)과 같은 소셜 네트워크 분석에서도 사용자 사이에서 주고 받는 문자형 메시지는 분석의 중요 대상이 되고 있으며, 문서 분류, 오피니언 마이닝 분야에서는 텍스트를 필수적으로 분석해야만 한다. 또한 고객관리, 마케팅, 인구조사에 이용되는 데이터베이스에도 이름, 주소, 불만사항과 같은 텍스트가 존재한다.

분석에 사용되는 모든 데이터는 노이즈(noise)를 가질 수 있다. 숫자형 데이터의 경우 노이즈는 참값과 다른 숫자로 정의할 수 있다. 그러므로 잘못된 숫자 기입이나 측정에 의하여 노이즈가 발생할 수 있다. 텍스트의 경우 노이즈는 참값과 다른 형태로 기록된 경우 (L. Venkata Subramaniam et al. 2009)뿐 아니라

문자에 대한 사전지식(background knowledge)이 없는 경우 (Daniel Jurafsky et al. 2009) 노이즈로 정의할 수 있다. 즉 '맨붕'과 같은 소셜 네트워크 서비스에 등장하는 신조어는 기존의 사전으로 인식할 수 없는 단어다. 이후 본 연구에서는 사전으로 인식할 수 없는 단어를 노이즈로 정의한다.

텍스트의 경우 노이즈가 발생하는 원인을 유형별로 나누면 다음과 같다. (1) 음성인식 과정에서 전기적 신호가 잘못된 글자로 처리되거나, '서빗스'와 같이 '서비스'를 키보드로 입력하는 과정에서 의도와 다르게 문자가 기입된 경우 오탈자가 발생한다. (2) '산업공학'을 줄여서 '산공'으로 사용하는 것과 같이 약어(abbreviations)나 (3) 특정한 영역에서 사용되는 일반적이지 않은 특수 용어 및 구어(non-standard terminology), (4) 조사, 마침표와 같은 기능어가 생략된 경우, (5) 띄어쓰기에 의하여 두 개 이상의 단어가 합쳐질 경우 노이즈가 발생한다 (Craig Knoblock et al. 2007). 이러한 텍스트 노이즈는 발생 원인은 두 가지로 분류하면 첫째는 실제로 사용을 하고 있는 단어지만 사전에 등록되어 있지 않아서 인식하지 못하는 경우와, 둘째로 의도와 다르게 기술이 되어 인식하지 못하는 경우가 있다. (2), (3)은 전자에 해당하며 (1), (4), (5)는 후자에 해당한다.

텍스트가 사용되는 경우에 따라서 주로 발생하는 노이즈의 유형이 다를 수 있다. 트위터나 모바일 문자 메시지(Short Message Service, SMS)의 경우 글자수 제한, 입력기기의 불편함으로 인한 오탈자와 띄어쓰기, 그리고 신조어 등장이 많이

관찰되며, 공정과정에서의 품질관리를 위한 불량 확인 보고서는 전문 용어들이 노이즈의 주원인이 될 수 있다.

텍스트에 노이즈가 발생하는 한 예시로서 전자세금계산서에 수기로 입력되는 업태를 살펴보면 [표 1]과 같다. '제조업'이라는 업태 대신에 '제조외', '제조'로 기술하는 경우도 많으며 '식품음료'를 '식음료'로 약어로 줄여서 사용하거나 '편의점', '담배'와 같이 [표 2]의 통계청의 한국산업표준분류표에서 제안하는 업태 분류가 아닌 단어를 입력하기도 한다. 오탈자, 약어는 사전에 등장하는 단어와 정확하게 일치하지 않기 때문에 알 수 없는 텍스트의 노이즈에 해당하며, '편의점'과 '담배'는 실제로는 사용이 되고 있는 단어지만, 사전에 등록되지 않았기 때문에 인식할 수 없는 텍스트 노이즈에 해당한다.

[표 1] 전자세금계산서에 수기로 입력된 데이터 예시

제조, 도매, 부동산	조립금속제조,	소매업.서비스업.
제조외	기타화학제조	부동산업
제조업	서빗스 도소매	식음료
건설업	편의점, 담배	제조 및 도소, 부동산

실제로 한국산업표준분류표에 등장하는 단어를 이용하여 [표 3]과 같이 업태를 표현하는 단어 사전을 만들 수 있다. 1년 3개월 기간 동안 전자세금계산서에 입력된 업태에 대하여 [표 3]의

단어들이 실제로 사용되는지 살펴보았다. 그 결과 총 48개의 사전에 포함된 단어 중 33개의 단어만이 사용되었으며, 그 비율은 전체 업태 입력의 45.24%에 해당하였다. 이 결과를 통하여 실제로 사용되는 업태의 표현 중에는 사전에 등록되지 않았거나 오탈자로 기입된 단어가 많음을 알 수 있다.

[표 2] 한국표준산업분류표

농업, 임업, 어업	금융 및 보험업
광업	부동산업 및 임대업
제조업	전문, 과학 및 기술 서비스업
전기, 가스, 증기 및 수도사업	사업시설관리 및 사업지원 서비스업
하수 폐기물 처리, 원료재생 및 환경복원업	교육 서비스업
건설업	보건업 및 사회복지 서비스업
도매 및 소매업	예술, 스포츠 및 여가관련 서비스업
운수업	협회 및 단체, 수리 및 기타 개인 서비스업
숙박 및 음식점업	가구내 고용활동 및 달리 분류되지 않는 자가소비 생산활동
출판, 영상, 방송통신 및 정보서비스업	국제 및 외국기관

[표 3] 한국표준산업분류표를 이용한 업태 표현 사전

가구내고용활동	보험	운수
가스	부동산	원료재생
개인	사업시설관리	음식점
건설	사업지원	임대
과학	사회복지	임업
광업	서비스	자가소비생산활동
교육	소매	전기
국제	수도사업	전문
금융	수리	정보
기술	숙박	제조
기타	스포츠	증기
농업	어업	출판
단체	여가	폐기물처리
도매	영상	하수처리
방송통신	예술	협회
보건	외국기관	환경복원

그러나 텍스트의 경우 노이즈가 있는 데이터를 완벽하게 알아보지 못하는 것이 아니다. 사람이 만드는 텍스트의 경우 약어나

오탈자라 하더라도 의미를 유추할 수 있다 (Stephan Gouws et al. 2011). [표 1]에서 사람이 ‘서빗스’를 ‘서비스’로, ‘식음료’를 ‘식품 음료’로 이해하는 것은 어렵지 않다. 그러나 컴퓨터는 단어의 의미를 유추할 수 없고, 추후 본석 또는 올바른 인식을 위해서 노이즈를 제거하여야 한다. 사람은 데이터로부터 의미를 유추할 수 있기 때문에 많은 경우 수작업으로 노이즈를 제거하지만, 수작업은 많은 노동 시간을 필요할 뿐 아니라, 때로는 전문 용어와 같이 특별한 사전지식을 요구하기도 한다. 그렇기 때문에 많은 비용이 드는 텍스트 노이즈를 효율적으로 처리하는 연구가 필요하다.

제 2 장 문제 정의 및 관련 연구

2.1. 연구의 필요성 및 연구 범위

데이터의 전처리 과정은 데이터마이닝을 이용하여 분석을 하기 전의 데이터를 정제하는 필수적인 과정으로서, 분석 성능에 큰 영향을 준다 (Galit Shmueli et al. 2007). 텍스트 데이터의 노이즈를 처리하는 것은 텍스트 데이터를 이용하는 분석이 수행되기 이전의 공통적으로 거쳐야 하는 과정이다 (Lipika Dey, et al. 2009). 한 예시로, 마케팅 전략 수립을 위하여 고객의 구매 내력을 분석하려고 한다. 제품의 이름을 한 사람이 아닌 여러 사람이 입력할 수 있다. 동일한 상품에 대하여 '마담미쉘 알로에 클린징', 마담미쉘/알로에 클렌징'으로 입력하였다면, 두 이름에 대하여 사람은 동일한 상품임을 판단할 수 있지만, 분석의 입장에서 서로 이름이 다르기 때문에 다른 상품으로 오판한다. 고객이 어떠한 상품을 구매하였는지를 명확히 알아야 하지만 상품의 이름이 조금씩 다르다면, 동일한 상품을 서로 다른 상품으로 오판할 수 있다. 또한 트위터와 같은 소셜 네트워크에서 발생하는 이야기의 의미를 분석하기 위해서는 형태소 분석과 같은 언어분석을 수행하지만, 노이즈가 들어간 텍스트의 경우 올바른 형태소 분석을 하지 못한다.

그리고 기하급수적으로 증가하는 상품 이름과 텍스트 데이터의 크기 및 신조어의 종류를 고려하면 수작업으로 노이즈를 처리하는

것은 불가능에 가깝다. 그러므로 반복적으로 이루어지는 텍스트 노이즈 처리 작업을 자동화하는 연구가 필요하다.

본 연구에서는 자연어로 기술되는 문장이 아니라 상품명이나 업태와 같이 명사로 입력되는 텍스트 데이터에 대하여 수작업으로 이루어지는 과정을 자동화하여 효율적으로 노이즈를 처리하는 방법을 제안한다. 또한 전자세금계산서에 수기로 입력되는 업태의 노이즈를 처리하는 과정에 적용하는 실제 사례를 함께 기술한다.

2.2. 관련 연구

기존의 오탈자 교정을 위한 연구에서는 노이즈 체널을 이용(Eric Brill et al. 2000)하거나 주로 발생하는 오탈자의 데이터를 학습한 뒤, 이를 이용하여 사전에 등록되지 않는 단어를 사전에 등록된 단어로 바꾸는 방법, 혹은 사전에 정의된 규칙을 이용하여 오탈자를 교정하는 방법이 제안되었다. 그러나 학습데이터나 사전에 정의된 규칙에 의존하는 경우, 학습데이터에 포함되지 않은 단어에 대하여 오탈자 교정을 할 수 없고, 규칙이 많은 노이즈를 처리할 수 있다고 보장할 수 없다. 특히 최근에는 문자메시지 (SMS)나 소셜네트워크에서 새로운 유형의 텍스트 노이즈들이 발생하기 시작했으며, 실제로 사용되는 언어 (real-world text)라는 단어를 사용하기도 한다. 이러한 노이즈를 해결하기 위해서 기존에 제안되었던 방법처럼 참조할 수 있는 데이터를 만드는 과정이

현실적으로 어렵거나, 참조할 수 있는 데이터가 전혀 없는 경우 사전지식을 활용하지 않으며 텍스트의 노이즈를 제거하고자 하는 비교사 학습의 노이즈 처리 방법(unsupervised cleansing noisy text)이 연구되었다.

단어의 변형이 일어나는 경우 빈번히 발생하는 노이즈들에 대하여 수작업으로 단어 변형의 규칙을 추출한 뒤, 이를 이용하여 다른 노이즈들을 정제하는 방법이 제안되었다 (Stephan Gouws et al. 2011). 그는 사전으로 인식할 수 없는 텍스트를 노이지 텍스트, 사전으로 인식할 수 있는 텍스트를 정제된 텍스트로 정의한 뒤, 노이지 텍스트가 수정되었을 때의 정제된 텍스트 모습을 학습하기 위하여 (노이지 텍스트, 정제된 텍스트)의 쌍을 만들었다. 쌍으로 이뤄진 데이터로부터 어휘의 변형(lexical transformation) 규칙을 학습한 뒤, 이를 이용하여 노이지 텍스트를 정제된 텍스트로 변형하는 방법을 제안하였다. 한 예로 (thinkin, thinking)과 (doing, doin)로부터 단어의 어미의 –ng가 –n으로 축약되는 규칙을 찾을 수 있다. 이 규칙을 이용하여 –n으로 끝나며 사전으로 인식할 수 없는 다른 단어들을 –ng로 복원한다. (노이지 텍스트, 정제된 텍스트)의 쌍을 자동으로 만들기 위하여 2GB의 뉴스기사를 이용하여 각 단어의 분포 유사도 (distributional similarity)와 형태의 유사성을 측정하는 어휘적 유사도 (lexical similarity)를 계산한 뒤, 두 유사도가 모두 높은 단어를 노이지 텍스트와 쌍으로 추출하였다. 그러나 이러한 방법은 트위터에서 사용되는 단어를 포함하기 위하여 많은 외부 데이터 (뉴스)와 하둡 (Hadoop)을

사용할 정도로 많은 계산량을 필요로 한다. 더 중요한 단점은 트위터에서만 사용이 되고, 뉴스에서는 사용이 되지 않는 표현이 존재한다면, 위 방법은 그러한 표현에 대해서 (노이지 텍스트, 정제된 텍스트) 쌍을 만들 수 없다. 즉, 노이즈를 처리하고자 하는 도메인에서만 사용되는 단어가 있다면 그러한 단어의 노이즈는 정제할 수 없다.

노이지 텍스트와 정제된 텍스트를 각각 다른 언어로 가정하고, 노이지 텍스트에서 정제된 텍스트로 기계번역(Statistical Machine Translation)을 하는 방법도 제안되었다 (Danish Contractor et al. 2010). 기계번역의 경우 두 가지 언어(노이지 텍스트, 정제된 텍스트)의 쌍을 학습데이터로 가지고 있어야 한다. 위 연구에서는 형태가 유사한 정제된 텍스트 후보를 생성한 뒤, 언어모델을 이용하여 자주 사용된 표현으로 노이즈를 변환하였다. 그러나 이 방법도 언어 모델을 사용하기 때문에 이름과 같은 단어에 적용하기가 어렵다. 또한 특정 도메인에서만 사용되는 표현을 포함하지 못할 가능성이 있다.

동일인물에 대하여 이름(단어)을 잘못 기입한 경우, 비슷한 이름을 자동으로 선택하기 위하여 형태가 가장 유사한 이름을 찾는 연구가 제안되었다 (William W. Cohen et al. 2003). 저자는 기존에 제안된 스트링 유사도 계산 방법들을 스트링의 각 글자를 비교하는 교정 거리군 유사도(edit-distance like functions)과 부분 글자들의 유사도(token-based distance like function), 그리고 위 두 가지 방법의 혼합 유사도(hybrid distance

function)로 분류한 뒤, 스트링 매치에서 사용할 경우 각 방법의 장단점을 실험으로 비교하였다. 이 외에도 변형이 가능한 노이즈의 후보군을 생성한 뒤, 키보드의 물리적 거리를 이용하여 오탈자의 확률을 계산하여 가장 적절한 단어를 선택하는 방법이 제안되었다 (Nicola Bertoldi et al. 2010). 그러나 위 두 연구는 사전에 사용되는 단어에 대하여 대부분 파악이 가능할 때 좋은 성능을 기대할 수 있다.

위의 관련 연구는 공통적으로 노이즈를 처리하려는 도메인에서 사용되는 단어들을 파악하기 위하여 많은 외부데이터, 혹은 신뢰할 수 있는 사전을 확보하여야 한다. 한 도메인의 텍스트를 정제하기 위하여 사전을 만든다고 하더라도 전문용어 사전은 구축 자체가 매우 어렵다. 데이터 전처리를 위하여 사전을 만드는 일의 비용을 줄이고, 그 과정에서 텍스트 데이터의 노이즈를 줄일 수 있다면 분석을 위한 전체적인 비용을 줄일 수 있다.

제 3 장 제안된 연구 방법

3.1. 제안된 방법의 절차

외부의 데이터를 이용하여 사전을 구축하는 연구는 특정 도메인에서 사용되는 용어들을 포함하지 못할 수 있다. 한 예시로 업태의 경우 많은 사람들은 ‘음식점 및 숙박’ 대신 ‘음숙’을 약어로 사용한다. ‘음숙’은 의도와 다른 형태로 사용되는 노이즈가 아니라 사전지식으로 알지 못하는 단어다. 이 경우에는 ‘음수’와 같은 단어가 ‘음숙’을 의도하지만 제대로 표현하지 못한 노이즈에 해당한다. 또한 사람들이 ‘음식점’이란 단어를 거의 사용하지 않고 ‘음식’을 주로 사용한다면 ‘음식점’은 ‘음식’을 의도하였지만 추가적으로 다른 글자가 합성된 노이즈로 판단할 수 있다.

즉 실제로 언어가 사용되는 모습을 가장 잘 반영하는 외부 데이터가 아니라 분석하고자 하는 데이터다. 사용이 되고 있지만 사전에 등록되지 않은 단어도, 그 단어를 의도하였지만 오탈자가 생긴 단어도 모두 분석하고자 하는 데이터 내부에 존재한다. 그러므로 노이즈를 처리하려는 데이터로부터 사전을 만들어야 실제로 사용되는 단어를 잘 반영할 수 있다.

[그림 1]은 본 연구에서 제안된 방법의 절차이다. 기존에 사전을 가지고 있지 않거나 신뢰할 수 없는 사전을 가지고 있을 경우데이터에서 인식이 가능한 단어는 적다. 1단계에서는 사전을

보완하기 위하여 데이터로부터 음절단위의 콜로캐이션을 이용하여 유사성이 있는 부분음절들을 추출한다. 2단계에서는 추출된 부분음절 후보들 중에서 단어의 후보를 선택하여 추출한다. 3단계에서는 사용자가 추출된 단어 후보들 중에서 사전에 추가할 단어를 선택한 뒤, 보완된 사전을 이용하여 사전에 등록되지 않은 표현인 노이즈를 판별한다. 이 과정을 통하여 실제로 사용은 되지만 사전에 알지 못한 단어들을 인식할 수 있다. 4단계에서는 의도와 다른 형태로 기술된 텍스트 노이즈를 복원하기 위하여 보완된 사전을 이용하여 스트링 매치(string match)를 수행한다. 이를 통하여 의도와 다르게 기술된 텍스트 노이즈를 복원한다. 보완된 사전으로부터 인식이 불가능한 데이터(노이즈)의 양은 이전보다 줄어들 수 있다.

1단계. 콜로캐이션을 이용한 부분음절 후보 추출

2단계. 부분음절로부터 단어 후보 추출

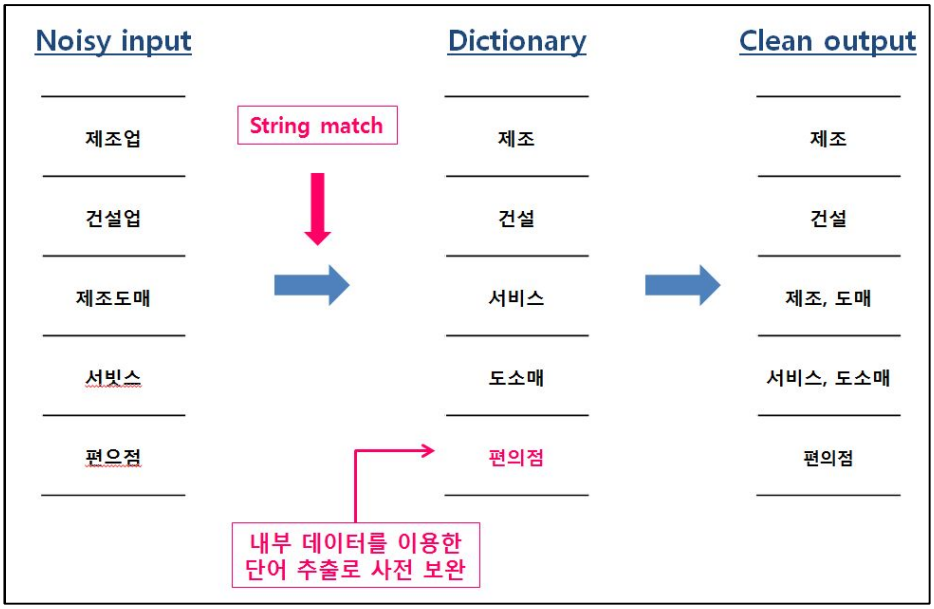
3단계. 단어 선택을 통한 사전 보완 및 노이즈 판단

4단계. 보완된 사전과의 스트링 매치를 통하여 노이즈 중 사전의 단어와 유사한 경우, 사전의 단어로 복원

[그림 1] 제안된 방법의 절차

.

수기 입력으로 이루어진 업태 데이터의 경우 [그림 2]를 통하여 위 과정을 설명하면 한국표준산업분류표로부터 추출한 [표 3]의 사전에는 '편의점'과 '편으점'은 사전에 등록되지 않았기 때문에 노이즈이다. 하지만, 데이터로부터 실제로 사용되고 있는 '편의점'이라는 단어를 추출하여 기존의 사전을 보완하면 '편의점'은 인식할 수 있고, '편으점'은 '편의점'으로 복원 할 수 있다.



[그림 2] 제안된 방법의 효과

사전을 보완하기 위한 가장 간단한 방법은 단어를 빈도 순으로 정렬한 뒤 상위의 단어 중에서 업태로 사용되는 단어를 선택하는 것이다. 그러나 소수의 단어만 자주 사용되고 대부분의 단어들은 자주 사용되지 않는다면, 상위의 몇 개의 단어를 살펴봐야 하는지

결정하기 어렵고, 자주 사용되지 않는 업태는 사전에 보완하지 못할 수 있다.

3.2. 콜로캐이션을 이용한 부분음절 후보 추출

자주 사용도지 않는 단어까지 추출하기 위해서는 빈도수에 영향을 받지 않는 정보를 이용하여 단어를 추출하여야 한다. 이를 위해서 데이터에서 연관성이 높은 연속된 음절을 추출한다. 중국어의 경우 사전에 등록되지 않은 단어를 추출하기 위하여 부분음절의 앞 뒤의 음절의 종류를 고려하는 것처럼 (Haodi Feng et al. 2004), 한국어에서도 부분음절의 앞 뒤 글자는 음절의 연관성을 판단할 수 있는 좋은 정보가 될 수 있다. 본 연구에서는 부분음절의 연관성을 계산하기 위하여 콜로캐이션 방법을 사용한다. 콜로캐이션은 '새빨간 거짓말' 과 같이 서로 함께 나오는 가능성이 많은 두 개 이상의 단어 집합을 추출하기 위한 자연어처리 방법이다 (Christopher D. Manning et al. 1999). 한 단어를 음절로 나눈 다음, 각 음절 단위로 콜로캐이션 방법을 적용한다. 즉, '서비스' 는 [서, 비, 스]가 된다. '서비스' 는 단어로 사용되기 때문에 [서, 비]는 서로 순서대로 함께 등장할 가능성이 높고, [비, 스] 역시 그렇다. 이를 확장하여 [서, 비, 스]가 순서대로 함께 등장할 가능성이 높다면 '서비스'를 단어로 추출한다. 두 음절의 연관성을 계산하기 위해서는 두 음절간의 상호 정보(mutual

information)나 카이 제곱 테스트(chi-square test) 를 수행할 수 있다. 그러나 제안하는 방법은 두 음절의 단어뿐 아니라 임의의 길이의 단어에 대하여 콜로캐이션 분석을 수행해야 하므로, 길이에 관계없이 확장할 수 있는 계산 방법이 필요하다.

한국어의 자동 띄어쓰기 분야에서도 단어의 경계를 자동으로 찾기 위한 방법을 연구한다. 단어 사이에 공백이 있을 확률을 사전에 정의한 개수의 이전 음절의 정보를 이용하여 하는 방법이 한국어의 띄어쓰기 분야에서 제안되었다 (Do-Gil Lee et al. 2007). 한 단어가 n 개의 연속된 음절이라면 단어 W 는 $W = [c_1, \dots c_n]$ 로 표현할 수 있다. 앞의 k개의 음절의 정보를 이용하여 단어의 경계를 판단하기 위해서

$$P(c_1 c_2 \dots c_n) = \prod_{i=1}^{i=n-k-1} P(c_{k+1} \dots c_{i+k+1} | c_k \dots c_{k+i})$$

와 같은 공식을 사용한다. 그 중에서도 적은 계산량을 위하여 이전의 음절만을 고려하는 2음절(bigram) 확률 모델이 제안되었다 (Seung-Shik Kang et al. 2001).

$$P(c_1 c_2 \dots c_n) = \prod_{i=1}^{i=n-1} P(c_i c_{i+1} | c_i c_{i-1})$$

하지만 사전에 정의된 길이의 음절만을 정보로 사용할 경우 특별히 자주 사용되는 단어에서 등장한 한 음절이 다른 단어에

사용될 경우 영향을 받을 수 있고, 길이가 긴 단어를 제대로 추출하지 못할 가능성이 있다.

한 예로 '기계도급' 이라는 단어에 대하여 살펴보면 '도매' 라는 단어가 자주 사용되기 때문에

$$P(\mathrm{c_i} = \text{도}, \mathrm{c_{i+1}} = \text{급} \mid c_i = \text{도}) < P(\mathrm{c_i} = \text{도}, \mathrm{c_{i+1}} = \text{매} \mid c_i = \text{도})$$

이고, 2음절 모델로 음절을 예측하면 '기계도' 라는 세 음절 다음에 '급' 이라는 음절이 추가되어 '기계도급' 이 될 가능성보다 '기계도매' 라는 음절이 나타날 가능성을 높게 계산한다. 그렇기 때문에 사전에 정의된 개수의 음절만을 이용하는 것보다 가능한 많은 음절의 정보를 이용하여 다음 음절을 예측하는 것이 바람직하다. 이는 문장에서 기준이 되는 양 옆의 한 단어가 아닌 주변의 많은 단어들을 이용하여 문맥(context)을 예측하는 것과 동일한 원리다. 언어모델의 경우 등장하지 않는 조합의 확률을 추정하기 위하여 스무딩 (smoothing)이나 인터폴레이션 (interpolation)과 같은 방법이 사용되지만, 단어를 추출하려는 데이터의 부분음절의 경우 계산해야 하는 모든 경우가 한 번 이상 데이터에 등장하였으므로, 위와 같은 보정방법이 필요하지 않다.

본 연구에서는 앞에 등장한 모든 음절을 이용하여 연속된 음절의 상관성을 숫자로 계산하는 전방향 점수(forward score)를 제안한다.

$$\text{forward score}(\mathrm{c}_1 \dots c_n) = P(c_2 c_1 | c_1) \times \dots \times P(c_1 \dots c_n | c_1 \dots c_{n-1})$$

기존에 등장한 음절을 이용하여 다음 음절의 조건부 확률을 구한 뒤, 이를 계속 누적하여 곱한다. 위 방법을 음절의 순서를 반대로 적용하여 후방향 점수(backward score)를

$$\text{backward score}(c_1 \dots c_n) = P(c_{n-1}c_n|c_n) \times \dots \times P(c_1 \dots c_n|c_2 \dots c_n)$$

과 같이 응용할 수 있다.

또한 특정한 음절의 영향을 줄이기 위하여 dispersion point normalization 방법(Sasˇa Petrovic´ et al. 2010)을 적용한 확산보정 점수(dispersion score)를 다음과 같이 제안한다.

$$\text{dispersion score}(c_1 \dots c_n) = \frac{1}{n-1}\sum_{i=1}^{n-1} P(c_1 \dots c_n|c_1 \dots c_i)$$

이 점수 계산 방법은 자주 사용되는 음절로 인한 왜곡효과를 줄여준다.

위 과정을 2음절 이상의 모든 단어의 부분음절들에 대하여 수행한다. {‘서비스’, ‘제조도매’}의 경우 {‘서비’, ‘비스’, ‘서비스’, ‘제조’, ‘조도’, ‘도매’, ‘제조도’, ‘조도매’, ‘제조도매’} 네 가지의 연속된 부분음절들로 나눌 수 있고, 모든 부분음절에 대하여 전방향, 후방향, 확산보정 점수를 계산한다. 각각

사전에 정의한 역치보다 높은 점수이면 모든 부분음절들을 단어의 후보로 판단한다.

3.3. 부분음절로부터 단어 후보 추출

모든 부분음절들에 대하여 위 과정을 수행하였기 때문에 후보로 추출된 부분음절들은 서로가 서로의 부분인 경우가 많다. '서비스' 가 후보로 추출되었을 경우 그 부분음절인 '서비', '비스' 모두 역치 이상의 점수를 가질 수 있다. 이 경우 { '서비' + '비스' } 두 부분음절들이 합쳐져 '서비스' 가 되므로, 후보들 중에서 자신보다 길이가 1 작은 부분음절이 모두 존재할 경우, 즉 $c_1 \dots c_p$, $c_1 \dots c_{p-1}$ 그리고 $c_2 \dots c_p$ 이 모두 존재할 경우 $c_1 \dots c_p$ 만을 후보로 사용한다.

모든 부분음절로부터 단어의 후보를 추출하는 과정을 알고리즘으로 기술하면 [그림 3]과 같다. [그림 3]의 S_{final} 은 최종적으로 추출되는 단어의 후보들 집합이다. S_{temp} 는 세 종류의 계산법에 의하여 추출되는 단어 후보를 저장하는 집합이다. (1) m 개의 단어 $W = \{w_1 \dots w_m\}$ 에 대하여 각 단어에서 길이가 2 이상인 부분음절 $c_j \dots c_{j+k}$들을 추출한다. 각 부분음절에 대하여 점수 계산 방법을 이용하여 점수(score)를 계산한 뒤, 사전에 정의한 역치(threshold)보다 클 경우, 후보들을 S_{temp} 에 저장한다. m 개의 모든 단어에 대하여 (1)의 과정을 마치면, S_{temp} 에 속한 모든 단어

후보들에 대하여 (2)의 과정을 수행한다. (2)의 과정에서는 S_{temp}의 단어 중 한 부분음절 $c_1 \dots c_p$와 $c_1 \dots c_p$의 부분인 $c_1 \dots c_{p-1}$, $c_2 \dots c_p$이 모두 존재할 경우 S_{temp} 에서 $c_1 \dots c_{p-1}$, $c_2 \dots c_p$ 을 삭제한다. (3)에서는 S_{temp}의 원소를 모두 S_{final}에 합한다.

Candidate selection Algorithm
Input = $W = \{ w_1, \dots, w_m \}$, ***return*** S_{final}: *final candidate set*

S_{final} = { }
For all *score metrics*, *score metrics = {forward, backward, dispersion}*
    S_{temp} = { }
    (1) For all w_i in $W = \{ w_1, \dots, w_m \}$
        For all $c_j \dots c_{j+k}$ in $w_i = [\, c_1 \dots c_{n_i} \,]$ & $1 \le k \le n_i$
            calculate score($c_j \dots c_{j+k}$)
            if score ≥ threshold :
                add $c_j \dots c_{j+k}$ to S_{temp}
        End For
    End For
    (2) For candidate in each S_{temp}
        if there exist { $c_1 \dots c_p, c_1 \dots c_{p-1}, c_2 \dots c_p$} :
            remove $c_1 \dots c_{p-1}, c_2 \dots c_p$ from S_{temp}
    End For
    (3) S_{final} = $S_{final} \cup S_{temp}$
End For

[그림 3] 단어 후보 선정 알고리즘

3.4. 단어 선택을 통한 사전 보완 및 노이즈 판단

사용자는 위 과정을 통하여 최종적으로 추출된 단어 후보들 중에서 실제 단어를 선택하여 사전에 입력함으로써 외부 데이터 없이도 데이터로부터 사전을 보완할 수 있다. 그 결과 데이터에서 실제로 사용이 되고 있는 단어들을 포함하는 사전을 만들 수 있다.

보완된 사전을 이용하여 데이터에서 사전에 등록된 단어와 그렇지 않은 단어를 분류한다. 사전에 등록되지 않은 단어 중 합성어를 판별하기 위하여 사전에 등록된 두 개 이상의 단어로 조합 가능 여부를 판단하였다.

사전에 등록되어 있지 않으며, 사전의 단어로 조합이 되지 않은 단어는 의도와 다르게 표현된 오탈자와 같은 노이즈로 판단한다.

3.5. 단어 추정을 위한 스트링 매치

‘서빗스’ 와 같이 노이즈가 있는 텍스트는 ‘서비스’라는 옳은 표현을 쉽게 추정할 수 있다. 그 이유는 많은 변형이 일어나지 않기 때문인데 ‘서빗스’ 는 ‘ㅅ’이 추가된 경우이고, ‘써비스’ 는 ‘ㅅ’이 ‘ㅆ”으로 대체된 경우이다. 한글은 한 글자가 초성, 중성, 종성의 조합으로 이루어지기 때문에(Young-Suk Kim, 2010) 입력 과정에서 초성, 중성, 종성이 독립적으로 오류가 날 수 있다. 그러므로 단어 형태의 유사성을 판단하기 위해서 한 음절을 초성,

중성, 종성으로 분리하여 비교한다. ‘서비스’라는 단어의 노이즈인 ‘서빗스’는 3음절 중 1음절이 불일치 하지만 초성, 중성, 종성으로 불리할 경우 6획이 일치하고 1획이 불일치 한다. 음절의 일치하는 비율은 $\frac{2}{3}$이지만, 획의 일치하는 비율은 $\frac{6}{7}$이다.

Levenshtein은 교정 거리(edit distance)로 알려진 Levenshtein distance를 제안하였다(VI Lcvenshtcin, 1966). 한 글자에서 다른 글자로 변하기 위한 최소 수정 행동 횟수를 계산하는 방법으로 수정 행동은 삽입(insertion), 대체(substitution), 삭제(deletion)이다.

‘운소’가 ‘운송’으로 교정되기 위해서 ‘ㅇ’이 삽입 되었고, ‘써빗스’가 ‘서비스’로 교정되기 위해서 ‘ㅆ’이 ‘ㅅ’으로 대체되어 ‘서빗스’로 교정된 된 뒤, ‘ㅅ’이 삭제되어 ‘서빗스’가 ‘서비스’로 교정 된다. 즉 (‘운소’, ‘운송’)의 교정 거리는 1이고, (‘써빗스’, ‘서빗스’)의 교정 거리는 2다. 하지만 긴 글자일수록 서로 다른 글자가 많을 수 있고, 길이에 관계없이 상대비교를 하기 위해서 정규화된 교정 거리(normalized edit distance)가 제안되었다(A. Weigel et al. 1994). 본 연구에서는 두 글자의 길이 중 긴 글자의 길이로 교정 거리를 나누어 정규화된 교정 거리를 계산하였다.

$$\text{normalized ED}(\mathrm{W}_1, \mathrm{W}_2) = \frac{\text{edit distance}(\mathrm{W}_1, \mathrm{W}_2)}{\max(\ length(W_1),\ \ length(W_2)\ )}$$

(‘서빗스’, ‘서비스’)의 교정 거리는 1이지만, 정규화된 교정 거리는 $\frac{1}{7}$이다.

Jaro는 두 문자열에서 자신의 앞 뒤의 $\left\lfloor\frac{\max(l_1,l_2)}{2}\right\rfloor - 1$ 범위 안에 있는 글자 중에서 같은 글자가 있는지를 판단하는 Jaro distance를 제안하였다(Matthew A. Jaro, 1985).

$$Jaro\ distance(s_1, s_2) = \frac{1}{3}\left(\frac{m}{l_1} + \frac{m}{l_2} + \frac{m-t}{m}\right)$$

l_i : i스트링의 길이
m : s_1, s_2 에 모두 존재하는 글자의 수
t : m중에서 서로 다른 위치에 존재하는 글자 수 / 2

[그림 4]는 ‘DIXON’과 ‘DICKSONX’의 Jaro distance의 예시다. ‘DIXON’의 D는 첫 번째 단어이므로, $\left\lfloor\frac{\max(5,8)}{2}\right\rfloor - 1 = 3$ 거리에 있는 단어, ‘DICKSONX’의 첫 번째 단어부터 네 번째 단어인 ‘DICK’에서 일치하는 글자가 있는지 확인하고, ‘DIXON’의 두 번째 단어인 I의 경우, ‘DICKSONX’의 첫 번째 글자부터 네 번째 글자인 ‘DICKS’ 중에서 일치하는 글자가 있는지 확인한다. N은 ‘DICKSONX’의 다섯 번째 글자인 S의 앞 뒤 3칸의 글자인 ICKSONX중에서 N인 글자가 있는지를 판단한다. 즉, DIXON의 각 글자의 위치를 기준으로 ‘DICKSONX’의 앞 뒤 3칸 사이에 있는 부분 중에서 일치하는 글자가 있는지 확인한다. 위 과정은 ‘DIXON’의 각 글자 별로 모두 수행하며 위 과정에서

동일한 글자로 판단되는 횟수를, 즉 [그림 4]의 색칠된 부분에 대하여 행과 열이 같은 글자의 수를 m으로 사용한다. [그림 4]의 예시의 경우 Jaro distance는

$$Jaro\ distance = \frac{1}{3}\left(\frac{4}{5} + \frac{4}{8} + \frac{4-0}{4}\right) = 0.7667$$

이 된다.

	D	I	X	O	N
D	1	0	0	0	0
I	0	1	0	0	0
C	0	0	0	0	0
K	0	0	0	0	0
S	0	0	0	0	0
O	0	0	0	1	0
N	0	0	0	0	1
X	0	0	0	0	0

[그림 4] Jaro distance 예시

한국어의 경우 한 글자마다 의미를 지니며, '제조도소매'와 '도소매제조'가 같은 같은 의미로 쓰일 수 있다. 즉, 한국어의 합성명사의 경우 글자의 위치가 자유로울 수 있으므로 Jaro가 제안한 Jaro distance처럼 비슷한 위치의 글자뿐 아니라 모든

글자의 동일여부를 판단하였다. [그림 5]의 예제에서 본래 Jaro distance는 색칠된 부분에서 동일한 글자가 있는지를 판단하므로 유사도가 0이지만, 변형한 Jaro distance는 모든 범위의 글자에 대하여 동일한 글자 유무를 판단하므로

$$\frac{1}{3}\left(\frac{5}{5}+\frac{5}{5}+\frac{10-10}{2}\right)=0.66$$

와 같이 유사도가 변한다. 이를 통하여 도매가 단어의 가장 앞에 붙은 경우나 가장 뒤에 붙은 경우에 대하여 모두 비슷한 단어로 판단할 수 있고, 두 단어 모두 포함된 글자 중 그 위치가 다른 경우는 불이익(penalty)을 줄 수 있다.

	도	소	매	제	조
제	0	0	0	1	0
조	0	0	0	0	1
도	1	0	0	0	0
소	0	1	0	0	0
매	0	0	1	0	0

[그림 5] 변형한 Jaro distance 예제

‘제조도소매’는 사전에 등록된 단어로 조합 가능한지를 확인함으로써 합성어 여부를 판단할 수 있다. 사전에 등록된 단어가

아니며 합성어가 아닌 모든 단어들에 대하여 교정 거리를 이용하여 사전에 등록된 단어들 중 가장 가까운 단어를 찾는다. 가장 가까운 단어와의 교정 거리가 사전에 설정한 역치보다 작은 경우, 교정 거리가 가장 가까운 단어로 변환한다.

사전에 등록된 단어, 합성어가 아니며 위 과정을 통하여 변환된 단어가 아닌 모든 단어에 대하여 변형된 Jaro distance를 이용하여 가장 가까운 합성어를 찾는다. 이 단어와의 거리가 사전에 설정한 역치보다 작다면 가장 가까운 합성어로 변환한다.

```
String match Algorithm
D = { d : word in dictionary }, W = { w : word in data }, C = { c : compound }
For each w in W
        check whether w is compound
        if w is compound :
                add w to C
End For
For each w ={w :w ∉ W ∩ D & w ∉ C }
        find the closest d using normalized edit distance
        if distance < threshold :
                convert w to the closest word d
End For
For each w ={w :w ∉ W ∩ D & w ∉ C }
        find the closest c using variant of Jaro distance
        if distance < threshold :
                convert w to the closest c
End For
```

[그림 6] 스트링 매치 알고리즘

현재까지 변환되지 않는 단어는 보완된 사전으로 해결할 수 없는

노이즈에 해당한다. 위 모든 과정을 [그림 6]과 같다.

제 4 장 실험

4.1. 데이터 전처리 및 변수 설정

실험에 사용한 데이터는 전자세금계산서에 수기로 입력된 업태의 텍스트 데이터이다. 데이터는 한국의 소프트웨어 개발 및 판매업체로부터 1년 3개월 동안 인터넷을 통하여 입력된 업태 데이터를 수집하였다.

총 236,806 번의 입력 데이터로부터 '제조, 도매'와 같이 중복 입력된 경우 '제조'와 '도매' 두 단어로 분리하였다. 또한 1음절의 단어는 판단이 불가능하므로 제외를 하였다. 그 결과 총 2,719개의 단어가 292,016번 등장하였다. 기존의 사전으로는 [표 3]의 사전을 사용하였다.

[표 4]에 기술된 상위 15개의 업태로 93.74%의 데이터를 설명할 수 있으며, [그림 7]과 같이 롱테일(long-tail) 분포를 따른다. 이는 자주 사용되지 않는 업태와 다양한 형태의 노이즈가 있을 수 있음을 의미한다.

또한 띄어쓰기나 ',' 을 이용하여 '제조, 도매, 부동산' 이나 '서비스 도매' 와 같이 여러 업태를 동시에 기술하는 경우와 '제조도매', '부동산도매' 와 같이 두 개 이상의 업태를 합성하여 기술하는 경우도 있다. '제조도매' 라는 단어가 합성어라는 것은 사람은 쉽게 판단할 수 있지만, 자동으로 합성어임을 인식하기 위해서는 '제조' 와 '도매' 가 각각

업태라는 지식이 필요하다.

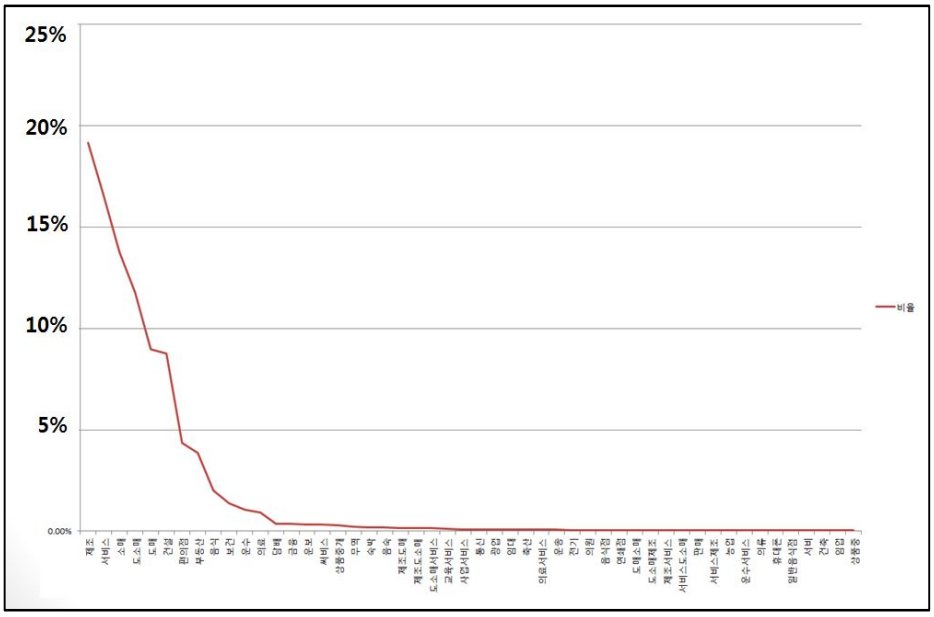
[표 4] 전자세금계산서에 입력된 업태 중 상위 18개의 빈도수 및 비율

빈도수 : 데이터에 단어가 등장한 횟수(frequency)

등장비율 : $\frac{(\text{frequency(word)})}{\sum_{word` \in D} frequeny(word`)}$

등장 누적 비율 : $\frac{\sum_{word \in frequency(word) > frequency(word`)} frequeny(word)}{\sum_{word` \in D} frequeny(word`)}$

순서	단어	빈도수	등장 비율	등장 누적 비율
1	제조	55954	18.49%	18.49%
2	서비스	48316	15.97%	34.46%
3	소매	40260	13.31%	47.77%
4	도소매	34385	11.36%	59.13%
5	도매	26202	8.66%	67.79%
6	건설	25613	8.47%	76.26%
7	편의점	12667	4.19%	80.45%
8	부동산	11281	3.73%	84.18%
9	음식	5897	1.95%	89.45%
10	보건	4018	1.33%	90.78%
11	운수	3117	1.03%	91.81%
12	의료	2730	0.90%	92.71%
13	담배	1109	0.37%	93.08%
14	금융	1010	0.33%	93.41%
15	운보	977	0.32%	93.74%



[그림 7] 전자세금계산서에 사용된 빈도수 기준 상위 50개 단어의 등장 비율

음절 단위의 콜로케이션을 위하여 전방향 점수, 후방향 점수의 경우 0.2 를, 확산보정 점수의 경우 0.65를 역치로 설정하였다. 각 방법의 점수가 설정한 역치보다 클 경우 단어의 후보로 추출한다. 또한 비슷한 단어를 찾기 위한 최대 거리인 정규화된 교정 거리의 역치는 0.3으로, 합성어들간의 유사도의 하한선인 Jaro distance 변형의 역치는 0.9로 설정하였다. 각 변수들은 반복 작업을 통하여 추출되는 후보들 중에서 단어로 인식되는 단어가 나오는 수준에서 경험적으로 선택하였다.

4.2. 실험 결과

세 가지 점수 계산 방법을 이용하여 [표 5]와 같은 194개의 후보를 추출하였으며, 이 중 50개 후보를 삭제한 뒤, [표 6]과 같이 149개 단어를 업태로 사전에 입력하였다. 기존 사전에 포함된 48개의 단어와 중복된 단어를 제외하여 최종적으로 [표 7]과 같이 154개의 단어를 포함하는 사전으로 보완하였다.

이 중에서 '식당'의 경우 실제 6번(331위) 등장하는 단어이고, '컴퓨터'는 18번(174위) 등장하는 단어로, 자주 사용되지는 않지만 업태로 사용할 수 있는 단어들이 추출됨을 확인할 수 있다. 172개 단어 중 154개 단어가 282,512번 (96.75%) 나타났으며, 기존 사전 48개 단어 중 33개가 45.24% 나타남과 비교하여 보완된 사전을 통하여 실제 사용되는 모습을 반영할 수 있음을 확인할 수 있다. 보완된 사전을 통하여 총 5,188번 (1.78%) 사용된 609개의 단어를 합성어로 인식하였으며, 1,295번 (0.44%) 사용된 437개의 노이즈 단어에 대하여 복원을 하였다. 최종적으로 3,342번 (1.14%) 사용된 1,574개 단어에 대하여 인식 및 복원을 할 수 없는 노이즈로 판단하였다.

[표 5] 음절 단위의 콜로케이션으로 추출된 194개의 단어 후보

영리	운송	사업	일반음식점	인터넷	1111
일반	써비스	전기	상품중개	판매대행	금융
위탁판매	조서비스	의원	웨어개발	법무서비스	잡화
운보	약국	휴게	체인	인적용역	화물
타서비스	대여	pc방	휴게음식점	컴퓨터	판매
호텔	음숙	비디오	의류	인적	시설
섬유	PC방	조립금속	전문	조립금속제품제	대리
식당	임대	서비스	어업	디자인	PC
한식	회계관련서비스	화장	조부동산	회계	관련서비스
장품	제조	연쇄	건설	식잡	콘도
보험	문건설	휴대폰	소프트웨어	광업	전자
접객	융서비스	주유소	서적	인서비스	기초화합
보서비스	정보	사업서비스	복제조	개인	출판
사회	병원	슈퍼	초화합물	술서비스	소프트웨
기계	담배	가공	운수관련	건축	계도급
23	전자상거래	종이제품	타화학제품	체인화	자동차
이동통신	농업	자상거래	조경	프트웨어	치킨
산서비스	기타	의복	타전기장비제	컨설팅	창고
요식	용품	도매	매부동산	매써비스	차관련
수출입	교육서비스	운수	음식	제작	정밀
유통	소매	매서비스	정밀기계제조	미용	의료
간이	보건	주택	소메	재조	부동산
통신	일용잡화	의복제조	교육	토목	사업관련
기술서비스	설서비스	사진	철강	편의점	업태
적용역	처리	수서비스	특수	디오대여	포츠
립금속제품제조	상품	스포츠	축산	공사	송서비스
회계서비스	기술	위탁	마트	출입	설기계도급
일반음식	산관련서비스	업부동산	숙박	건설기계도급	무역
가스	탁판매	도소매	장비	섬유제품제조	핸드폰
호프	수업서비스	휴게음식	터넷	알선	세무사
복권	여행	식품접객	학원	연쇄점	기초화합물
료서비스	작물재배	연구개발	양약	수입	치과
종합	비영				

[표 6] 최종적으로 선정한 149개 업태

일반	회계서비스	가공	철강	편의점	호프
위탁판매	제조	종이제품	특수	공사	복권
운보	정보	의복	축산	건설기계도급	위탁판매
호텔	병원	도매	마트	금융	수업서비스
섬유	담배	운수	숙박	잡화	여행
식당	전자상거래	주택	인터넷	화물	작물재배
한식	농업	의복제조	판매대행	판매	도소매
보험	기타	사진	법무서비스	시설	휴게음식
접객	교육서비스	스포츠	인적용역	대리	식품접객
기계	소매	위탁	컴퓨터	PC	연구개발
이동통신	보건	일반음식점	인적	콘도	장비
요식	일용잡화	상품중개	조립금속제품	전자	인터넷
유통	사업	체인	디자인	기초화합	학원
간이	전기	휴게음식점	회계	출판	양약
통신	의원	의류	식잡	소프트웨어	섬유제품
기술서비스	휴게	전문	광업	자동차	알선
회계서비스	pc방	어업	인서비스	치킨	연쇄점
일반음식	비디오	건설	개인	창고	수입
운송	서비스	소프트웨어	건축	정밀	핸드폰
써비스	화장	서적	체인화	의료	세무사
약국	연쇄	운수관련	컨설팅	부동산	기초화합물
대여	휴대폰	조경	제작	사업관련	치과
음숙	주유소	음식	미용	기계도급	종합
PC방	사업서비스	정밀기계제조	재조	무역	비영리
임대	슈퍼	교육	토목	가스	

[표 7] 보완된 사전의 172개 단어

일반	마트	써비스	전자상거래	가구내고용활동	운수
운보	요식	제작	PC	원료재생	주택
위탁판매	숙박	약국	농업	비디오	수리
사업시설관리	인터넷	대여	기타	화장	의복제조
어업	판매대행	음숙	협회	서비스	학원
건설	법무서비스	미용	콘도	영상	사진
증기	인적용역	보건업	교육서비스	무역	양약
소프트웨어	유통	재조	전자	연쇄	알선
섬유	간이	토목	사회복지	가스	스포츠
호텔	컴퓨터	음식점	기초화합	휴대폰	기계도급
식당	인적	공사	출판	호프	위탁
서적	디자인	편의점	소매	주유소	연쇄점
국제	회계	PC방	보건	사업서비스	일반음식점
건설업	식잡	임대	일용잡화	슈퍼	수입
한식	단체	임업	자동차	가공	핸드폰
보험	폐기물처리	건설기계도급	기술	복권	조립금속제품
접객	비영리	과학	자가소비생산활동	종이제품	섬유제품
운수관련	광업	제조	치킨	수업서비스	세무사
제조업	인서비스	외국기관	창고	환경복원	치과
조경	통신	잡화	정밀	여행	기초화합물
기계	개인	금융	의료	의복	종합
음식	기술서비스	화물	사업관련	작물재배	상품중개
정밀기계제조	건축	정보	부동산	도매	체인
교육	방송통신	판매	사업	휴게음식	하수처리
예술	체인화	대리	의원	도소매	휴게음식점
철강	회계서비스	시설	전기	식품접객	의류
특수	일반음식	병원	수도사업	사업지원	전문
축산	운송	여가	휴게	연구개발	장비
이동통신	컨설팅	담배	pc방		

4.3. 성능 평가

본 연구에서 제안하는 방법의 효용성을 설명하기 위하여 사전을 보완하기 전과 후의 단어 인식 및 사전을 이용하여 복원이 가능한 노이즈의 비율을 비교하였다.

[표 8]과 [표 9]에 기술됨과 같이 기존의 사전을 이용한 경우 사전으로 인식할 수 있는 단어는 45.24%이고, 이를 이용하여 106개, 0.5%의 합성어를 인식하였다. 노이즈 중에서 253개, 12.8%의 단어에 대하여 사전에 있는 단어로 추정하여 복원하였지만, 최종적으로 41.46%개의 단어에 대하여 인식 및 복원을 하지 못하였다. 하지만 음절 단위의 콜로케이션을 통하여 사전을 보완한 경우, 144개의 단어를 이용하여 데이터의 96.7%를 인식하였고, 최종적으로 1,574개의 단어에 대하여 1.14%만 수작업을 필요로 하였다. 즉 40.32%의 노이즈에 대하여 수작업으로 처리를 하지 않고 인식 및 추정이 가능하였다.

사용자는 단어 후보로 추출된 194 개 글자 중에서 단어를 선택하는 행동만으로 대부분의 노이즈를 처리할 수 있었다. 즉 이전에는 수작업으로 사용되는 단어를 찾아서 노이즈를 처리하는 과정을 자동화 할 수 있음을 알 수 있고, 이를 통하여 노이즈를 효율적으로 처리할 수 있음을 입증하였다.

[표 8] 사전을 보완하기 전과 후의 노이즈 처리 성능 비교 [단어 개수]

D : 사전에 등록된 단어
$D \cap W$: 사전에 등록된 단어 중 데이터에 나타난 단어
C : 합성어
R : 스트링 매치로 복원된 단어
N : 수작업이 필요한 단어

	D	$D \cap W$	C	R	N
사전 보완 전	48 단어	33 단어	106 단어	253 단어	2,327 단어
사전 보완 후	172 단어	144 단어	564 단어	437 단어	1,574 단어
성능향상	124 단어	111 단어	458 단어	184 단어	**753 단어**

[표 9] 사전을 보완하기 전과 후의 노이즈 처리 성능 비교 [출현 빈도 비율]

D : 사전에 등록된 단어
$D \cap W$: 사전에 등록된 단어 중 데이터에 나타난 단어
C : 합성어
R : 스트링 매치로 복원된 단어
N : 수작업이 필요한 단어

	D	$D \cap W$	C	R	N
사전 보완 전	48 단어	45.24%	0.5%	12.8%	41.46%
사전 보완 후	172 단어	96.7%	1.71%	0.44%	1.14%
성능향상	124 단어	51.46%	1.21%	(-12.36%)	**40.32%**

제 5 장 결 론

5.1. 결론

본 연구를 정리하면 다음과 같다. 단어로 입력되는 텍스트 데이터의 노이즈를 처리하기 위하여 사전을 구축하는 과정에 많은 비용이 들 수 있다. 위와 같은 경우 외부 데이터를 이용하여 자동으로 사전을 만들 수 있지만, 특정 도메인에서만 사용되는 단어가 사전에 포함되어있음을 보장할 수 없다. 그러므로 노이즈를 처리하고자 하는 데이터에서 실제로 사용되는 단어를 상관성이 있는 연속된 음절을 콜로캐이션을 이용하여 추출함으로써 사전을 보완한다. 보완된 사전을 통하여 텍스트 노이즈 중에서 사전의 단어와 가장 유사한 단어를 스트링 매치를 수행하여 의도와 다르게 기술된 텍스트 노이즈를 복원하였다.

전자세금계산서에 입력된 업태 데이터의 노이즈를 제거하기 위하여 제안된 방법을 적용함으로써 본 연구가 실제로 유용함을 입증하였다. 수작업으로 텍스트 노이즈를 처리하는 방법과 비교하여 사용자는 단어의 후보를 추출하기 위한 역치를 설정하고, 후보로 추출된 단어들 중에서 실제 업태로 사용될 단어만 선택함으로써 사전을 보완할 수 있다. 보완된 사전을 이용한 경우, 보완되지 않은 사전을 이용한 경우보다 40.32%의 노이즈를 추가적으로 복원할 수 있었다. 또한 빈도수가 낮은 단어도 추출이 가능하였다. 즉 대량의

외부 데이터와 계산을 하지 않으면서 도메인을 잘 설명할 수 있는 사전을 만들 수 있고, 이를 이용하여 노이즈를 효율적으로 처리할 수 있다.

5.2. 향후 연구 방향

첫째, 본 연구에서 제안한 방법의 성능에 가장 많은 영향을 줄 수 있는 부분은 데이터로부터 단어를 추출하는 부분이다. 이 부분에 대해서 중국어의 자연어처리 분야에서 많은 연구가 이루어지고 있으며, 한국어에서도 단어의 경계를 판단하기 위하여 자동 띄어쓰기에서 많은 연구들이 이루어지고 있다. 위 방법들을 이용하여 단어들을 보다 잘 추출할 수 있도록 하여 노이즈를 처리하는 시스템의 성능을 높일 수 있다.

둘째, 본 연구에서는 명사로 이루어진 단어에 대하여 노이즈를 처리하였다. 언어의 구조적인 특징을 반영하여 특정한 도메인에서 사용되는 명사가 아닌 문장에 대하여 단어를 추출하고, 이를 이용하여 노이즈를 처리하는 방법으로 연구를 진행하려 한다.

셋째, 본 연구에서 사용한 데이터는 약 2,719 개의 단어 중에서 단어를 추출하여 노이즈를 처리하였지만, 문장의 텍스트이거나, 이보다 많은 양의 데이터에서 단어를 추출하고 노이즈를 처리하기 위해서는 많은 계산이 필요하기 때문에 하둡과 같은 분산시스템을 이용하여 데이터를 처리하는 기술을 습득하여야 한다. 그리고 이를

통하여 빠른 시간에 대용량 데이터에 대하여 단어를 추출하고 노이즈를 처리할 수 있는 방법에 대하여 연구하고자 한다.

참고 문헌

Nicola Bertoldi, Mauro Cettolo and Marcello Federico, 2010, *Statistical Machine Translation of Texts with Misspelled Words*, The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Eric Brill, Robert C. Moore, 2000, *An improved error model for noisy channel spelling correction*, ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.

William W. Cohen, Pradeep Ravikumar, Stephen E. Fienberg, 2003, *A Comparison of String Distance Metrics for Name-Matching Tasks*, American Association for Artificial Intelligence.

Danish Contractor, Tanveer A. Faruquie and L. Venkata Subramaniam, 2010, *Unsupervised Cleansing of Noisy Text*, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics.

Lipika Dey, Sk. Mirajul Haque, *Opinion Mining from Noisy Text Data*, 2009, International Journal of Document Analysis and Recognition.

Haodi Feng, Kang Chen, Xiaotie Deng, Weimin Zheng, *Accessor Variety Criteria for Chinese Word Extraction*, 2004,

Stephan Gouws, Dirk Hovy and Donald Metzler, 2011, *Unsupervised Mining of Lexical Variants from Noisy Text*, EMNLP '11 Proceedings of the First Workshop on Unsupervised Learning in NLP.

Matthew A. Jaro, *Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, 1989*, Journal of the American Statistical Association

Daniel Jurafsky and James H. Martin, 2009, *Speech and Language Processing*, Pearson Education, Inc.

Seung-Shik Kang, Chong-Woo Woo, *Automatic Segmentation of Words using Syllable Bigram Statistics*, 2001, Proceedings of NLPRS.

Young-Suk Kim, *Componential Skills in Early Spelling Development in Korean*, 2010, Scientific Studies of Reading.

Craig Knoblock, Daniel Lopresti, Shourya Roy, L. Venkata Subramaniam, *Special Issue on Noisy Text Analytics*, 2007, International Journal of Document Analysis and Recognition.

VI Lcvenshtcin, *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*, 1966, Soviet Physics-Doklady.

Do-Gil Lee, Hae-Chang Rim, and Dongsuk Yook, *Automatic Word Spacing Using Probabilistic Models Based on Character n-grams*, 2007, IEEE Intelligent Systems.

Christopher D. Manning, Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, 1999, MIT Press.

Sasˇa Petrovic´, Jan Sˇnajder, Bojana Dalbelo Basˇic, 2010, *Extending lexical association measures for collocation extraction*, Computer Speech & Language.

Galit Shmueli, Nitin R. Patel and Peter C. Bruce, 2007, ***Data Mining for Business Intelligence,*** John Wiley & Sons ,Inc.

L. Venkata Subramaniam, Shourya Roy, Tanveer A. Faruquie, Sumit Negi, ***A Survey of Types of Text Noise and Techniques to Handle Noisy Text***, 2009, AND '09 Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data.

A. Weigel, E Fein, ***Normalizing the Weighted Edit Distance***, 1994, Pattern Recognition.

통계청, http://kostat.go.kr

Abstract

Cleansing noisy text using corpus extraction and string match

Hyunjoong Kim
Department of Industrial Engineering
The Graduate School
Seoul National University

The text which is not enrolled in a dictionary is defined as noise, and it is called out-of-vocabulary (OOV). Types of noisy texts are divided into two classes. First case is unexpected word that is used actually but not enrolled in a dictionary. Neologism, abbreviations, and jargon are the first case. Second case is unintended words due to many reasons. Spelling error is typical case of them. Usually noisy texts are modified manually. But manually handling noisy texts costs much and sometimes it requires related domain knowledge

This paper introduces efficient process of cleansing noisy

texts. The process consists four phases. In the first phase, correlated sequential syllables are extracted as candidates of words. And for the extraction, this paper proposes three collocation measures; forward score, backward score, and dispersion point normalized score. In the second phase, unusable candidates of words are removed by proposed elimination algorithm. In the third phase, a user determines which words are to be added to dictionary. And all words in data are classified whether enrolled word or not by complemented dictionary. In the last phase, unintended words are transformed to a word of dictionary, if there is a similar word.

This paper includes experiment with real data; words of business type which are putted from the keyboard. Through this experiment, the rate of words recognized using complemented dictionary and converted to clean text is improved by 40.32% comparing the case using the not complemented dictionary. It shows that the proposed process is useful.

Keywords : cleansing noisy text, extraction corpus, syllable collocation, string match

Student Number : 2011-21171