공학석사 학위논문

GMM을 이용한 한국어 말뭉치에서의 오류 탐지

Detecting Errors in Korean Corpus based on GMM

지도교수 김 재 훈

2020 년 2 월

한국해양대학교 대학원

컴퓨터공학과

최 민 석

본 논문을 최민석의 공학석사 학위논문으로 인준함

위 원 장 박 휴 찬 (인)

위 원 류 길 수 (인)

위 원 김 재 훈 (인)

2019년 12월 26일

한국해양대학교 대학원

목 차

List of Tables

List of Figures

Detecting Errors in Korean Corpus Based on GMM

Choi, Min Seok

Department of Computer Engineering
Graduate School of Korea Maritime and Ocean University

Abstract

In computational linguistics, a corpus is a large and structured set of language samples collected from real world text for a specific purpose. There are be various types of errors in the corpus because most corpus are built manually and/or semi-automatically and the errors are caused by human intervention. Such errors make corpus-based learning systems worse in performance. Many studies have therefore been conducted to detect and correct such errors in various ways and most studies have been done from pre-built corpus. Human intervention is, however, still required. In addition, error correction is not only very tedious as well as laborious and cost-expensive.

In this paper, we propose a method for detecting corpus errors using GMM clustering algorithm. The purpose of this paper to detect errors under the small size of corpus. That is, the proposed method can be used in developing corpus by integrating into annotation tools. The proposed method consists of three steps. The first step is to make word embedding vectors of some error-prone context. The second step is to reduce the dimension of the

vectors because clustering with a large dimension of vectors is time-consuming. The third step is to group the reduced vectors and to detect outliers as errors.

For experiments, we have used two kinds of corpora: Korean dependency corpus and Korean semantic role labelling (SRL) corpus of which each one comprises only 1000 sentences. Our results show that the proposed method can serve as a error detector in early stage of corpus development. Our best results achieve recall of 65.15% for Korean dependency corpus and recall of 69.46% for Korean SRL corpus.

In the future, we will do research on representing features for detecting errors and also on correcting errors as well as detecting errors. Motivated by the proposed method, we will start to investigate error detection in case that there is a large tagged corpus.

KEY WORDS: GMM, clustering, error detection, dependency parsing corpus, semantic corpus

GMM을 이용한 한국어 말뭉치에서의 오류 탐지

최민석

한국해양대학교 대학원
컴퓨터공학과

초록

말뭉치란 특정 목적을 가지고 언어 표본을 추출한 집합을 의미한다. 이런 말뭉치에는 목적에 따라 다양한 종류가 있다. 대부분의 말뭉치는 사람의 수작업으로 구축되기 때문에 다양한 오류들이 포함되어 있으며, 오류가 포함된 말뭉치를 사용하는 시스템은 좋은 성능을 기대할 수 없다. 이러한 문제점을 해결하기 위해 다양한 방법으로 오류를 탐지하고 수정하는 연구가 진행되었다. 하지만 대부분의 방법들이 이미 제작된 말뭉치를 학습하여 오류를 탐지하고 수정한다. 이러한 작업을 여러 번 수행하여야 하며 많은 비용이 소요된다. 이 문제를 다소 완화시키기 위해 본 논문에서는 GMM(Gaussian Mixture Model)을 이용한 군집화를 통해 오류 탐지 방법은 제안한다. 군집화는 비지도학습의 한 방법으로 표지가 부착된 학습 데이터가 없거나 적더라도 오류 탐지를 수행할 수 있다. 따라서 이미 제작된 말뭉치가 아니라 말뭉치를 구축하는 과정에도 사용할 수 있다.

본 논문에서 제안하는 방법으로 수행된 오류 탐지를 검증하기 위하여

구문분석 말뭉치와 의미역 말뭉치를 사용하였다. 성능 평가의 척도는 정보검색에서 널리 사용되는 정밀도와 재현율을 사용하였다. 구문분석 말뭉치와 의미역 말뭉치에서 각각 65.15%와 69.46%의 재현율을 보였다. 이와 같은 결과를 바탕으로 제안한 모델을 사용하여 다양한 말뭉치의 오류 탐지를 수행할 수 있음을 알 수 있다.

재현율을 좀 더 향상시킬 수 있도록 자질 확장 등의 연구를 진행할 수 있을 것이다. 또한 말뭉치 구축 도구에 직접 적용하여 제안된 시스템이 얼마나 효율적인지도 평가할 계획이다.

KEY WORDS: GMM, 군집화, 오류탐지, 구문분석 말뭉치, 의미역 말뭉치

제 1 장 서 론

말뭉치란 자연언어처리 연구를 위해 특정 목적을 가지고 언어 표본을 추출한 집합을 의미하며, 그 중에서 품사가 부착된 말뭉치를 품사 부착 말뭉치라고 한다. 언어에 따라 다양한 품사 부착 말뭉치가 구축되었고, 한국어에서도 다양한 품사 부착 말뭉치가 구축되었다(김재훈 & 김길창, 1995; Han & Han 2001; 김홍규, 2007). 이 중에서 다양한 분야에서 널리 이용되는 말뭉치는 세종말뭉치(김홍규, 2007)이다. 세종 말뭉치는 오랜 기간 다양한 사람들이 제작하였으며, 다양한 오류도 포함하고 있다(이미경 외, 2005). 이런 오류들이 많이 포함된 말뭉치를 사용할 경우 자연언어처리 시스템의 성능 저하가 우려된다. 따라서 성능 향상을 위해서는 오류 수정이 필요하다. 하지만 이런 오류를 수정하기 위해서는 많은 인력과 시간이 필요하므로 비용이 많이 든다. 또한, 많은 인력이 수작업을 통해 오류를 수정하기 때문에 일관성을 유지하기가 쉽지 않다. 이와 같은 문제점을 해결하기 위해서 본 논문에서는 수작업 과정에서 발생할 수 있는 오류의 가능성을 줄이고 효율적인 오류 수정에 도움을 주기 위해 말뭉치에서의 오류 탐지 방법에 대해 다룬다.

오류 탐지는 다양한 단계에서 수행될 수 있다. 실시간으로 오류를 탐지하기 위해서는 말뭉치 제작 단계에서 오류 탐지가 수행되어야 한다. 반대로 말뭉치 제작이 완료된 이후에 오류 탐지를 수행하여 말뭉치의 오류를 수정할 수도 있다(이정규 외, 1997; 김영길, 2003; 홍진표, 2013; 최명길 외, 2013). 본 논문에서는 실시간으로 오류를 탐지하는 방법에 관해 연구를 진행한다.

말뭉치의 종류에 따라 탐지해야 할 오류가 다르다. 따라서 오류 탐지

연구를 진행하기 위해서는 대상 말뭉치가 미리 선정되어야 한다. 본 논문에서는 구문분석 말뭉치와 의미역 말뭉치에서의 오류를 탐지하는 방법에 대해 다룬다.

본 논문은 구문분석 말뭉치와 의미역 말뭉치에서의 오류를 실시간으로 탐지하는 시스템을 제안한다. 제안된 시스템은 문맥 표상 제작, 문맥 표상의 차원 축소, GMM(Gaussian Mixture Model)을 이용한 말뭉치에서의 오류 탐지로 구성된다. 문맥 표상은 문맥의 정보를 잘 표현하기 위해 구문분석 말뭉치와 의미역 말뭉치의 여러 정보를 포함하여 각각의 표상을 만든다. 이러한 표상의 크기는 다양한 정보를 포함하므로 커질 수밖에 없다. 표상이 커질수록 오류 탐지를 수행하는데 걸리는 시간은 증가된다. 이러한 소요 시간 증가는 실시간으로 수행되는 시스템에는 적합하지 않다. 이러한 문제점을 해결하기 위해 차원 축소를 진행한다. 차원 축소는 자기부호화기(autoencoder)를 학습시킨 후, 부호부(encoder)를 이용하여 각 표상의 차원을 축소한다. 오류 탐지는 축소된 표상을 바탕으로 비지도학습(unsupervised learning)의 한 종류인 GMM을 이용한 군집화(clustering)를 수행한다. 수행된 결과를 바탕으로 말뭉치 구축 작업 시 오류 발생 가능성을 낮추는 목적에 따라서 재현율을 정밀도보다 우선시하여 기준값을 지정한다. 지정한 기준값을 넘어서는 표상을 오류로 판단한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 오류 탐지 연구의 전반적인 내용과 군집화 알고리즘을 설명한다. 3장에서는 본 논문에서 제안하는 한국어 말뭉치에서의 오류 후보 탐지 시스템을 설명한다. 4장에서는 본 논문에서 제안한 방법으로 구현한 시스템의 성능을 평가하고, 마지막으로 5장에서 결론을 맺고 향후 연구 방향을 제시한다.

제 2 장 관련 연구

2.1 오류 탐지

오류 탐지란 전체 데이터에서 다른 패턴의 데이터를 찾는 것을 말한다. 이런 데이터를 오류(이상, anomaly)라고 정의한다. 오류 탐지는 금융 분야(Aleskerov, 1997), 의학 분야(Spence *et al*., 2001), 사이버 보안 분야(Kumar, 2005), 행동 패턴 분야(Liu *et al*., 2008) 등 다양한 분야에서 활용되고 있다. 금융 분야는 신용카드 데이터를 바탕으로 도난되거나 분실된 신용카드를 탐지할 수 있다. 의학 분야에서는 자기공명 영상장치(MRI) 이미지를 바탕으로 악성 종양을 탐지할 수 있다. 사이버 보안 분야에서는 비정상적인 트래픽을 탐지할 수 있다. 행동 패턴 분석 분야에서는 CCTV 자료를 바탕으로 도난이나 화재와 같은 사고 발생을 탐지할 수 있다. 이와 같이 다양한 분야에서 오류는 악의적 행동이나, 비정상적인 행동, 비일상적인 행동 등과 같은 이유로 발생한다. 이를 자연언어처리 분야에서 생각해 본다면, 말뭉치 제작과정에서 오류를 탐지할 수 있을 것이다. 현재까지 이와 같은 다양한 분야의 오류 탐지를 수행하기 위해 다양한 오류 탐지 방법들이 개발되었다. 이러한 다양한 오류 탐지를 수행하기 위해서는 우선 각 업무 및 분야에 따라 오류를 정의해야 한다. 본 논문은 말뭉치 제작에서 발생되는 표지 부착 오류를 그 대상으로 한다.

오류 탐지 방법은 분류 기반 오류 탐지 방법(Stefano *et al*., 2000), NN(nearest neighbor) 기반 오류 탐지 방법(Byers & Raftery, 1998), 정보 이론 오류 탐지 방법(Li & Vitanyi, 1993), 스펙트럴 이상 탐지 방법(Agovic *et al*., 2007), 군집화 기반 오류 탐지 방법(Yu *et al*., 2002) 등이 있으며 이

하에서 순서대로 간략하게 설명할 것이다.

분류 기반 오류 탐지 방법은 각 데이터가 속하는 클래스 표지가 붙은 데이터를 바탕으로 분류기를 학습하여 새로운 데이터의 클래스를 예측하는 방법이다. 분류 기반 오류 탐지 방법에서는 어느 클래스에도 포함되지 않는 데이터를 오류로 판단한다(Stefano *et al*., 2000; Barbara *et al*., 2001; Fan *et al*., 2004).

NN 기반 오류 탐지 방법은 정상들은 어떤 위치 주위에 밀집되어 있고, 오류들은 각 위치에서 멀리 떨어져 있다고 가정한다. NN기반 오류 탐지 방법에서는 k번째로 가까운 개체와의 거리(Byers & Raftery, 1998) 또는 상대 밀도(Breunig *et al*., 1999) 등을 이용하여 오류 점수를 구하여 오류를 판단한다.

정보 이론 오류 탐지 방법은 자료의 정보량을 분석하는 방법이다. 오류는 정보량의 불규칙을 발생시킨다는 가정을 기반으로 한다. 데이터의 집합 A의 복잡도를 $C(A)$라 한다. $C(A)-C(A-I)$를 크게 하면서 충분히 작은 $I \subset A$를 찾아 I 에 속하는 데이터들을 오류로 판단한다(Li & Vitanyi, 1993; Lee & Xiang, 2001; Lin & Brown, 2006).

스펙트럴 이상 탐지 방법은 변수들의 조합으로 자료의 변동(variability)을 대부분 설명하도록 데이터를 근사한다. 이 방법에서는 데이터를 더 낮은 차원의 부분공간으로 투영한 뒤, 그 공간에서는 오류와 정상이 확연하게 구분된다고 가정한다. 따라서 오류를 쉽게 찾아낼 수 있는 부분공간을 정하는 것이 일반적인 접근 방법이다(Agovic *et al*., 2007). 부분공간을 정하였다면 부분 공간에서 오류를 판단한다.

군집화 기반 오류 탐지 방법은 비슷한 개체들의 군집을 형성하여 오류를 탐지하는 방법이다. 군집화 기반 오류 탐지 방법은 가정에 따라 세 종류로 나누어진다. 첫 번째 가정은 정상들은 군집에 모여 있고, 오류는 군집에 속하지 않는다(Yu *et al*., 2002). 두 번째 가정은 군집의 중심으로부터의 거리가 짧으면 정상이고, 길면 오류라고 가정한다(Smith *et al*.,

2002). 세 번째 가정은 정상은 밀집된(dense) 군집에, 오류는 한산한(sparse) 군집에 속한다는 가정이다(He *et al.*, 2003). 위와 같이 군집화 기반 오류 탐지에서는 각 가정에 따라서 오류를 판단한다.

오류 탐지 방법들마다 각각의 장점과 단점이 존재한다. 정보 이론 기반은 한두 개의 오류도 탐지할 정도의 민감한 척도를 찾아야한다. 분류 기반은 고차원 데이터들을 처리하기 좋지만, 모든 데이터들의 정답 표지를 필요로 한다. 이러한 단점을 생각한다면 일부의 표지만 사용하는 NN 기반이나 군집화 기반의 방법을 사용하는 것이 더 효율적이다. 하지만 NN 기반이나 군집화 기반은 고차원의 데이터를 다루는 것이 분류 기반보다 힘들다는 단점이 존재한다.

따라서 본 논문에서는 모든 데이터들이 정답 표지를 알 수 없는 말뭉치 데이터를 이용하기 때문에 분류 기반 보다는 군집화 기반의 오류 탐지 방법을 사용할 것이다. 하지만 군집화 기반의 단점인 고차원 문제가 발생하는데 이 문제를 보완하기 위해 차원 축소를 이용한다. Figure 2.1은 *k*-means 군집화 오류 탐지의 예를 나타낸다. 군집 밖에 있는 것을 오류로 판단한다.

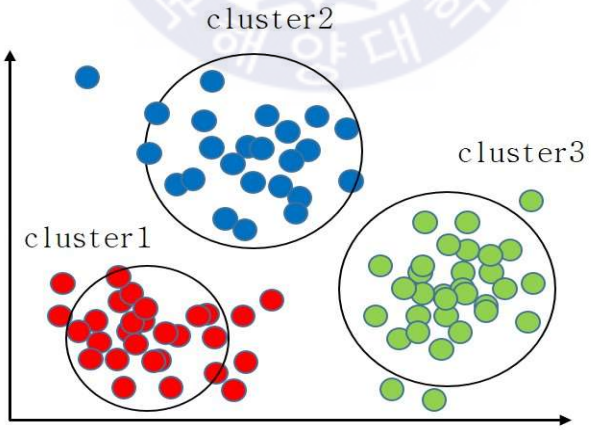

Figure 2.1 An example of anomaly detection with *k*-means clustering.

2.2 GMM 알고리즘

군집화란 데이터를 여러 범주의 군집을 찾는 것을 의미한다. 군집화는 비지도학습의 한 방법이며 데이터 분석을 위한 많은 분야에서 사용되는 방법이다. 이러한 군집화를 수행하면 비슷한 속성 및 특징을 가지는 데이터 요소를 특정 그룹으로 분류할 수 있다. Figure 2.2는 군집이 2개인 군집화의 예를 나타낸다.

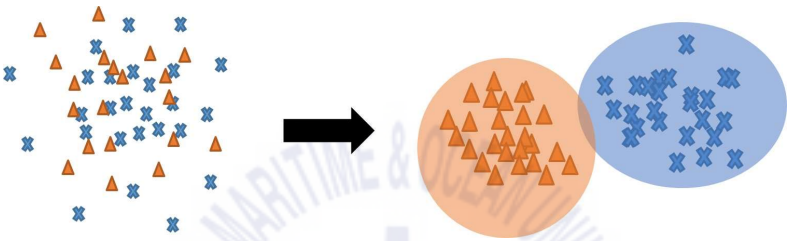

Figure 2.2 An example of clustering with two clusters.

또한, 군집화는 비지도학습의 한 종류인 만큼 학습을 수행하는데 정답 표지가 필요하지 않다. 이러한 특징은 사람이 직접 정답 표지를 부착해야 하는 수작업에 비교하면 효율과 비용 감소에 효과적이다. 뿐만 아니라 많은 정답 데이터가 쌓이기 전인 초기 말뭉치를 구축할 때에도 유용하다. 대표적인 군집화 알고리즘으로는 *K*-Means Clustering(Hartigan & Wong, 1979), Mean-Shift Clustering(Comaniciu & Meer, 2002), DBSCAN(Density-Based Spatial Clustering of Applications with Noise)(Ester *et al*., 1996), GMM(Stan & Anil, 2009) 등이 있다. 본 논문에서는 여러 가지 군집화 알고리즘 중 GMM을 이용하였다.

GMM은 복잡한 형태의 확률 분포를 Figure 2.3과 같이 k개의 가우시안 분포를 혼합하여 표현하는 것이다.

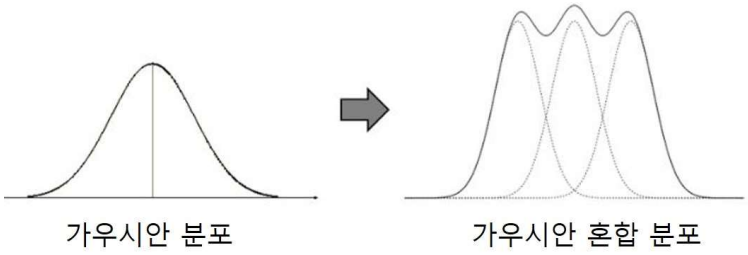

Figure 2.3 An example of a Gaussian mixture with three categories of Gaussian distributions.

k 범주의 확률변수 Z가 있다고 가정하면 확률분포함수는 식 (2.1)과 같다.

$$p(z=k)=\pi_{(k)} \tag{2.1}$$

실수 값을 출력하는 확률변수 X는 확률변수 Z의 표본값 k에 따라 기댓값 μ_k, 분산 Σ_k이 달라진다. 이를 식으로 나타내면 식 (2.2)로 표현된다.

$$p(x|z)=N(x|\mu_k,\Sigma_k) \tag{2.2}$$

즉 x는 z가 주어졌을 때 가우시안 분포를 따른다는 것을 의미한다. 식 (2.1)과 식 (2.2)를 결합하면 식 (2.3)으로 정리된다.

$$p(x)=\sum_{Z}p(z)p(x|z)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k) \tag{2.3}$$

이제 x에 대한 주변 확률을 계산할 수 있다. 단, GMM에서 Z의 값을 알 수가 없으므로, 관측되지 않는다고 가정한다. 따라서 혼합 가우시안 분포는 내부에 숨겨진 확률 변수를 포함하는 잠재변수모형이다.

X가 주어졌을 때의 조건부 확률 $p(z|x)$를 정의해 보면, 식 (2.4)처럼 정리할 수 있다.

$$\begin{aligned}\pi_{ik} &\equiv p(z_i = k|x_i) \\ &= \frac{p(z_i = k)p(x_i|z_i = k)}{\sum_{k=1}^{K} p(z_i = k)p(x_i|z_i = k)} \\ &= \frac{\pi_k N(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)}\end{aligned} \tag{2.4}$$

식 (2.4)는 결과적으로 모수로부터 추정한다.

$$(\pi_k, \mu_k, \Sigma_k) \Rightarrow \pi_{ik}$$

π_{ik}는 i번째 데이터 x_i가 k 범주에서 만들어졌을 확률을 나타내고 k에 대한 조건부확률이라고 한다. 이제 GMM의 모수 추정을 하여야 한다. N개의 데이터에 대한 X의 확률분포는 식 (2.5)로 정의된다.

$$\begin{aligned}p(x) &= \prod_{i=1}^{N} p(x_i) \\ &= \prod_{i=1}^{N} \sum_{z_i} p(z_i)p(x_i|z_i) \\ &= \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)\end{aligned} \tag{2.5}$$

계산의 편의를 위해 식 (2.5)에 로그를 취하면 식 (2.6)으로 정리된다.

$$\log p(x) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)\right) \tag{2.6}$$

만약, x_i가 어떤 범주 z_i에 포함되는지 확인할 수 있다면, 범주 확률분포 π_k와 정규분포의 모수 μ_k, Σ_k도 확인할 수 있다. 하지만 실제로는 z_i를 확인할 수 없기 때문에 확률분포함수 $p(x)$를 최대화하는 π_k, μ_k, Σ_k를 비선형 최적화를 통해 구해야 한다. 식 (2.6)을 μ_k로 미분하여 0이 되도록 방정식을 만들면 식 (2.7)이 되고 이를 정리하면 식 (2.8)과 같이 정리할 수 있다.

$$\frac{\delta \log p(x)}{\delta \mu_{k,}} = 0$$

$$0 = -\sum_{i=1}^{N} \frac{p(z_i = k) p(x_i | z_i = k)}{\sum_{k=1}^{K} p(z_i = k) p(x_i | z_i = k)} \Sigma_k (x_i - \mu_k) \tag{2.7}$$

$$\sum_{i=1}^{N} \pi_{ik} (x_i - \mu_k) = 0$$

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \pi_{ik} x_k \tag{2.8}$$

$$N_k = \sum_{i=1}^{N} \pi_{ik}$$

k 범주에 속하는 데이터의 수와 비슷한 의미를 가진다. 즉 μ_k는 k 범주에 속하는 데이터의 샘플 평관과 같다는 의미를 나타낸다. 마찬가지로 로그-확률분포함수를 Σ_k로 미분하여 계산하면 식 (2.9)로 정리된다.

$$\frac{\delta \log p(x)}{\delta \Sigma_k} = 0$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \pi_{ik} (x_i - \mu_k)(x_i - \mu_k)^T \tag{2.9}$$

마지막으로 로그-확률분포함수를 π_k로 미분하여 계산하여야 하는데, π_k는 혼합 계수이므로 추가적인 제약이 존재한다. 따라서 이러한 제약을 추가하여 라그랑주 승수법(Lagrange multiplier method)을 사용하여 처리한다. 제약을 추가한 식은 식 (2.10)으로 정의되고, 이 식을 미분하여 0이 되는 값을 찾으면 식 (2.11)로 정리된다.

$$\log p(x) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{2.10}$$

$$\frac{\log p(x) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)}{\pi_k}$$

$$\pi_k = \frac{N_k}{N} \tag{2.11}$$

식 (2.8), (2.9), (2.11)은 모두 조건부확률에 영향을 받기 때문에 자명한 (closed-form) 해를 가지지 못한다. 따라서 해를 구하기 위해 반복적인 방

식의 EM(Expectation Maximization) 알고리즘을 사용하여 구한다. EM 알고리즘은 조건부확률을 추정하는 E 단계와 모수를 추정하는 M 단계로 이루어지며 이를 번갈아 추정하여 정확도를 높이는 방법이다. E 단계는 모수가 정확하다고 가정하고 이를 바탕으로 조건부확률을 추정한다.

$$(\pi_k, \mu_k, \Sigma_k) \Rightarrow \pi_{ik}$$

M 단계는 조건부확률이 정확하다고 가정하고 이를 사용하여 모수를 추정한다.

$$\pi_{ik} \Rightarrow (\pi_k, \mu_k, \Sigma_k)$$

위의 두 단계를 반복하면 모수와 조건부확률을 점진적으로 개선하여 해를 구할 수 있다.

2.3 차원 축소

많은 양의 정보를 포함하는 데이터는 변수의 개수가 방대해지므로, 데이터의 차원의 크기가 커지게 된다. 데이터의 차원이 커지면 자동적으로 데이터를 계산하는데 드는 부하가 증가하게 된다. 이러한 계산적 부하는 차원이 커질수록 기하급수적으로 늘어나는데 이러한 현상을 차원의 저주(curse of dimensionality)라고 한다. 이러한 차원의 저주를 피하기 위해 차원 축소를 이용한다. 차원 축소는 여러 가지 방법들이 연구되고 있는데 크게 특징 선택(feature selection), 특징 추출(feature extraction), 심층학습 기반의 차원 축소로 나누어진다.

특징 선택의 목적은 모든 특징의 부분 집합을 선택하거나, 불필요한 특징을 제거하여 작은 특징 집합을 만드는 것이다. 이러한 특징 선택은 사전 지식을 바탕으로 수행할 수도 있지만, 자동 특징 선택(automatic feature selection methods)을 사용하기도 한다. 자동 특징 선택 방법들은 몇 개의 특징들을 제외시킨 후, 성능을 확인하는 방법이며 대부분의 특징 선택 알고리즘의 방식이다. 특징 선택 기법들은 Lasso(Tibshirani, 1996),

SPEC(Zhao & Liu, 2007), Fisher Score(Duda *et al*., 2012) 등이 있다.

특징 추출은 특징 선택과 다르게 원본 특징들의 조합으로 새로운 특징을 생성하는 것이다. 동작 원리는 고차원의 원본 특징 공간을 저차원의 새로운 특징 공간으로 투영시킨다. 새롭게 구성된 특징 공간은 보통 원본 특징 공간의 선형 또는 비선형 결합이다(Jundong *et al*., 2016). 선형 방법에는 Linear Discriminant Analysis(Scholkopft & Mullert, 1999), Principle Component Analysis(Jolliffe, 2002) 등이 있고 비선형 방법에는 ISOMAP(Tenenbaum *et al*., 2000), Locally Linear Embedding(Roweis & Saul, 2000)등이 있다.

심층학습 기반의 차원 축소 방법 중 하나는 자기부호화기이다. 자기부호화기란 들어온 입력의 차원을 은닉층(hidden layer)의 크기만큼 축소하였다가 다시 입력과 같은 크기의 출력층의 크기로 확대하는 방법이다. 이런 축소와 확대를 진행하더라도 입력층(input layer)으로 들어간 입력과 출력층(output layer)으로 출력된 결과가 최대한 유사하도록 학습하는 것이 자기부호화기이다(Baldi, 2012).

2.4 한국어 구문분석 말뭉치

구문분석이란 문장을 구성하는 성분들의 관계를 분석하는 것을 말한다. 구문분석을 통해 문장의 구조를 파악함으로써 문장 내에 존재하는 중의성을 해소하기 위한 연구이다. 구문분석을 하는 방법은 관점에 따라 크게 구 의존 구조분석(dependency parsing)과 구 구조분석(constituency parsing)으로 나뉜다.

의존 구조분석이란 문장을 이루는 각 어절에 대해 지배소와 의존소를 인식하고 둘의 의존 관계를 파악하는 방법이다(Lim, 2015). 의존 구조분석은 단어 간의 의존 관계만으로 문장의 구조를 파악할 수 있어 문장 구성 성분이 생략되거나 도치되어도 분석이 가능하다. 따라서 언어적 특성상 생략과 도치가 빈번한 한국어 구문분석에서도 많이 연구되고 있다(이건

일, 2015; 나승훈, 2017; 최용석 & 이공주. 2019).

현재 공식적으로 구축된 한국어 의존 구조 말뭉치는 없으며, 일반적으로 구 구조 세종 구문분석 말뭉치를 바탕으로 의존 구조 말뭉치로 변환시키는 구축 방법을 많이 사용한다. 보편적으로 의존 구조 말뭉치를 구축할 때 UD(Universal Dependency)의 CoNLL-U(Zeman, 2017) 형식을 많이 사용한다. UD는 언어의 종류와 상관없이 통일된 주석을 바탕으로 의존구조 말뭉치를 구축하기 위한 기준을 말한다. 의존 구조 말뭉치에 적용되는 UD의 CoNLL-U 형식은 Figure 2.4과 같이 10개의 열로 구성된다. 이 중 7열(HEAD 열)은 해당 FORM의 지배소 ID를, 8열(DEPREL 열)은 FORM이 가지는 지배소와의 의존 관계를 표시한다. HEAD 열의 값이 0일 경우, 이는 root를 지배소로 가짐을 뜻한다.

Table 2.1은 (최용석 & 이공주, 2018)에서 구축한 의존 구조 말뭉치 중에서 유연한 중심어 후위 원칙을 적용하여 생성한 말뭉치의 한 예시이다. CoNLL-U 형식을 따르고 있으며, 세종 구문분석 말뭉치를 의존 구조 말뭉치로 변환시킨 것이다. 다만 아직 한국어 구문분석을 위한 지침이 통일되지 않아 최대한 세종 말뭉치의 구조를 유지하도록 변환하였다. 따라서 세종 구문분석 말뭉치와 동일하게 주로 어절을 토큰으로 사용하였으며, 괄호 안에 띄어쓰기가 있는 경우에는 쌍을 이루는 기호나 문장 부호를 분리하여 하나의 토큰으로 사용하였다.

Table 2.1 An example of a sentence in Korean dependency corpus.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	플라스틱으로	플라스틱 으로	NOUN	NNG+JKB	-	2	obl	-	-
2	만든	만들 ㄴ	VERB	VV+ETM	-	3	acl	-	-
3	샤베트기는	샤베트기 는	NOUN	NNG+JX	-	5	nsubj	-	-
4	수입품이	수입품 이	NOUN	NNH+JKS	-	5	nsubj	-	-
5	대부분이다.	대부분 이 다 .	ADJ	NNG+VCP+EF+SF	-	0	root	-	-

2.5 한국어 의미역 말뭉치

의미역은 서술어에 의해 기술되는 행위나 상태에 대한 명사구의 의미역할을 나타내며 서술어에 의해 기술되는 명사구를 논항(argument)이라고 한다(배장성, 2015). 의미역은 문장의 구조가 바뀌어도 논항은 바뀌지 않기 때문에 의미 분석에 필요한 정보를 제공한다. 의미역 결정(Semantic Role Labeling, SRL)이란, 문장의 서술어와 논항들 사이에 적합한 의미 관계를 결정하는 것을 의미한다. 의미역 분석에서 논항의 역할은 서술어의 의미에 따라 결정된다. 각각의 서술어는 활용과 의미에 따라 사용되는 논항이 정해져있다. 이러한 내용을 정리한 것이 격틀 사전이다.

의미역 말뭉치를 구축하는 방법은 일반적으로 격틀 사전 기반(김완수 & 옥철영, 2016)의 방법과 의미역 말뭉치에 기반(박성배 외, 2000)을 둔 방법으로 나뉜다. 격틀 사전 기반의 방법은 입력된 문장과 격틀 사이의 유사도를 계산하여 의미역을 결정한다. 이러한 방법은 빠른 속도와 높은 정확률이라는 장점이 있지만, 격틀 사전을 구축하는 것이 어렵고 격틀 사전에 존재하지 않는 것은 처리하지 못한다는 단점이 있다. 의미역 말뭉치에 기반을 둔 방법은 의미역 말뭉치와 기계학습을 적용하여 의미역을 결정하는 방법이다. 격틀 사전에 비해 쉽게 적용할 수 있다는 장점이 있지만, 의미역 말뭉치를 구축하는데 어려움이 있다는 단점이 있다(김병수 외, 2007).

격틀 사전 기반의 방법과 의미역 말뭉치 기반을 둔 방법 모두 격틀과 의미역 말뭉치를 필요하기 때문에 의미역 결정에 말뭉치는 필수적이라 할 수 있다. PropBank는 의미역 결정에 필요한 말뭉치로써 영어 의미역 말뭉치 중 대중적으로 쓰이는 중요한 말뭉치이다(Palmer *et al.*, 2005). PropBank를 기반으로 만들어진 한국어 의미역 말뭉치가 Korean PropBank이다. 한국어 의미역 말뭉치는 영어 의미역 말뭉치의 양에 비해 상당히 부족하여 연구에 어려움이 있다. 이러한 점을 해결하기 위해 다양한 방법으로 연구가 진행되고 있다(이창기 외, 2014; 배장성 외, 2014; 배장성 외,

2015).

본 논문에서 사용하는 의미역 말뭉치는 전북대 의미역 말뭉치(박광현 & 나승훈, 2017)를 사용하였다. Table 2.2는 본 논문에서 오류 탐지를 위해 사용한 의미역 말뭉치의 구성을 나타낸다. Table 2.1과 비슷한 구성을 가지고 있으며 마지막 열이 의미역 정보를 나타낸다. ARG0, ARG1과 같이 '/' 앞에 오는 정보는 의미역 표지이다. ARG0은 서술어의 동작주, 행위자를 나타내고 ARG1은 서술어의 피동작주, 대상 등을 나타낸다. '/' 뒤에 오는 정보는 의미역 표지의 의존소 ID를 나타낸다.

Table 2.2 An example of a sentence in Korean SRL corpus.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	그는	그 는	UPOS	NP+JX	-	4	-	-	-
2	운전면허	운전면허	UPOS	NNG	-	3	-	-	-
3	시험에	시험 에	UPOS	NNG+JKB	-	4	-	-	-
4	합격하였다.	합격 하 였 다.	UPOS	VV+EP+EF+SF	-	0	-	-	(ARG0 / 1), (ARG1 / 3)

제 3 장 오류 후보 탐지 시스템

본 논문은 GMM 알고리즘 기반의 군집화를 통해 말뭉치에서 오류를 탐지하는 방법을 제안한다.

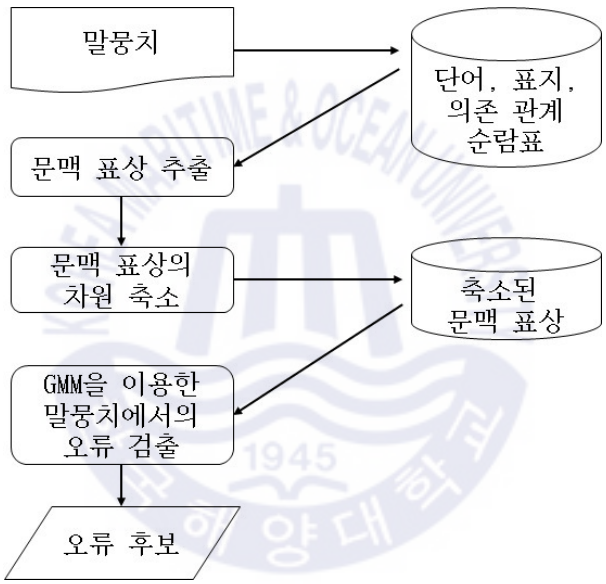

Figure 3.1 The process for detecting error candidates by the proposed method.

Figure 3.1은 본 논문에서 오류 후보를 탐지하는 전체적인 과정과 각 과정의 방법을 보인다. 첫 번째 단계(문맥 표상)는 말뭉치에서 문맥의 뜻을 잘 나타낼 수 있는 문맥 표상을 제작한다. 문맥 표상에 들어갈 특징은 사용되는 말뭉치에 따라 달라진다. 두 번째 단계(차원 축소)로는 제작된 문맥 표상의 차원 크기를 축소한다. 본 논문에서 제안하는 방법은 말뭉치를 제작하는 과정에서 동시에 오류를 탐지하는 방법이기 때문에 소요되는 시간도 중요한 요소이다. 이러한 시간 요소를 충족하기 위해서는 문맥 표상의

차원 축소가 필요하다. 마지막 단계(GMM을 이용한 말뭉치에서의 오류 탐지)로 GMM 알고리즘을 이용하여 축소된 문맥 표상을 군집화 시킨 후, 군집화된 정보를 바탕으로 오류 검출을 수행하고 최종적으로 오류 후보로 판단한다. 각 단계는 이하의 절에서 순서대로 자세히 설명할 것이다.

3.1 문맥 표상

말뭉치에서의 각 토큰[1]의 의미는 독립적이지 않고, 다른 토큰들의 영향을 받는다. 따라서 토큰 하나만으로는 문맥에서의 의미를 파악하기 어려우므로 주위의 여러 가지 정보를 반영한다.

3.1.1 구문분석 말뭉치에서의 문맥 표상

구문분석 말뭉치에서의 문맥 표상은 지배소와 의존소의 관계 정보를 주로 반영하였다. Figure 3.2는 '플라스틱으로 만든 샤베트기는 수입품이 대부분이다.'라는 원문을 바탕으로 의존 트리를 보인다.

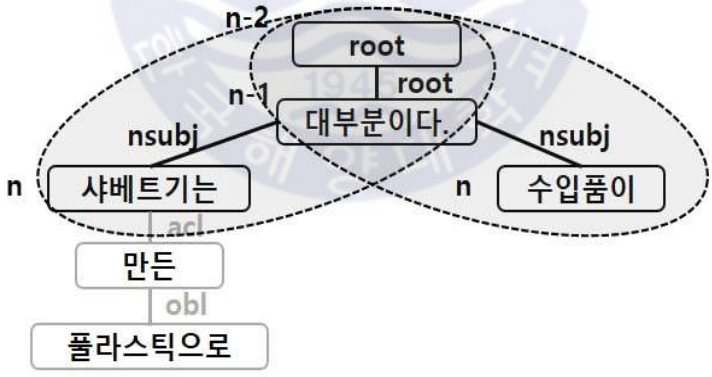

Figure 3.2 The dependency tree of the example sentence.

Figure 3.3은 Figure 3.2를 바탕으로 nsubj 표지를 기준으로 만든 의미 표상을 나타낸다.

1) 각 말뭉치에서 사용되는 기본 단위를 의미: 말뭉치마다 다르다

Figure 3.3 Examples of contextual embedding of the dependency relation "nsubj" tag in the dependency tree.

Figure 3.3과 같이 구문분석 문맥 표상에서는 2단계 위(n-2)의 지배소의 단어($Word_{n-2}$)와 1단계 위(n-1)의 지배소의 단어($Word_{n-1}$), 그리고 두 지배소 간의 의존 관계(dependency realation)($DepRel_{n-2|n-1}$), 현재 단어($Word_n$), 마지막으로 현재 단어와 n-1 단어와의 거리(distance)정보($dis_{n-1|n}$) 총 5개를 포함한다. 각 단어는 200차원의 크기를 가지고 의존 관계와 거리는 30차원의 크기를 가진다. 따라서 구문분석에서 문맥 표상의 크기는 총 660이다.

3.1.2 의미역 말뭉치에서의 문맥 표상

의미역 말뭉치에서는 지배소와 의존소 간의 의미역 정보를 주로 반영하였다. Figure 3.4는 '그는 운전면허 시험에 합격하였다.'라는 원문을 바탕으로 의존 관계를 보인다.

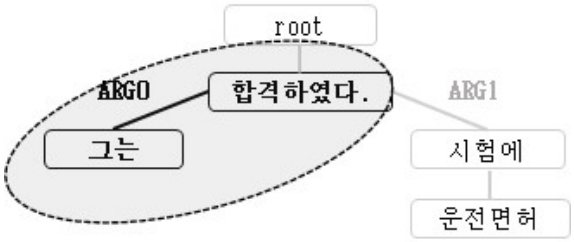

Figure 3.4 A SRL tree of the first example sentence.

원문의 서술어 ‘합격하였다.’를 기준으로 논항 ‘그는’이 ARG0(동작주)이고 논항 ‘시험에’가 ARG1(대상주)이다. 문장으로 나타내면 ‘그는(ARG0) 운전면허 시험에(ARG1) 합격하였다.’가 된다. Figure 3.5는 ‘지혜는 꽃밭을 좋아한다.’라는 원문을 바탕으로 의존 관계를 보인다.

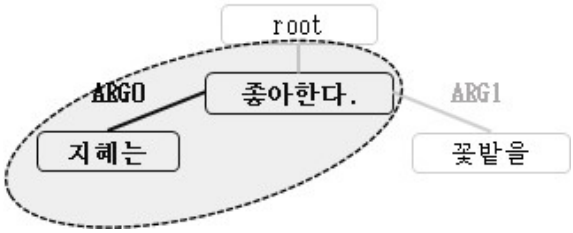

Figure 3.5 A SRL tree of the second example sentence.

원문의 서술어 ‘좋아한다.’를 기준으로 논항 ‘지혜는’이 ARG0(동작주)이고 논항 ‘꽃밭을’이 ARG1(대상)이다. 문장으로 나타내면 ‘지혜는(ARG0) 꽃밭을(ARG1) 좋아한다.’가 된다. Figure 3.6은 Figure 3.4와 Figure 3.5를 바탕으로 의미역 표지 ARG1을 기준으로 만든 의미 표상을 나타낸다.

Figure 3.6 Examples of contextual embedding of the SRL “ARG0” tag in SRL tree.

Figure 3.6과 같이 의미역 문맥 표상에서는 지배소의 단어($Word_{head}$), 지배소의 표지(Tag_{head}), 의존소의 단어($Word_{dep}$), 의존소의 표지(Tag_{dep}), 그리고 지배소와 의존소의 거리($Dis_{head|dep}$) 총 5개를 포함한다. 구문분석과 마찬가지고 각 단어는 200차원의 크기를 가지고 표지와 거리는 30차원의 크기를 가진다. 결과적으로 의미역에서의 문맥 표상의 크기는 총 490이다.

3.2 문맥 표상의 차원 축소

문맥 표상의 차원 크기가 커질수록 오류 후보 탐지를 수행하는데 시간이 오래 걸린다. 이러한 점은 짧은 수행시간을 요하는 작업에서는 치명적이다. 이러한 문제점을 해결하기 위해서는 문맥 표상의 차원 크기를 축소해야한다. 차원 축소의 방법은 2.3절에서 설명한 것과 같이 다양한 방법이 존재하지만 본 논문에서는 자기부호화기를 이용하여 차원 축소를 수행하였다.

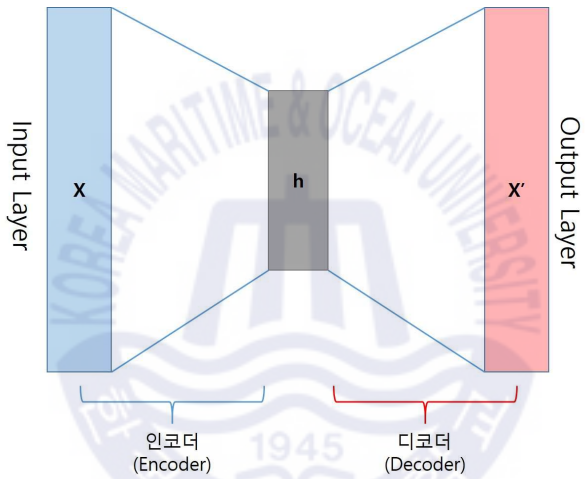

Figure 3.7 The structure of AutoEncoder.

Figure 3.7은 자기부호화기의 구조이다. 실제 구문분석 말뭉치에서 사용한 자기부호화기는 입력층의 크기를 660, 은닉층의 크기를 100, 출력층의 크기를 입력층과 동일한 660으로 설정하여 자기부호화기를 학습시켜 사용하였다. 의미역 말뭉치에서는 은닉층의 크기는 동일한 100으로 설정하고, 입력층과 출력층을 490으로 설정하여 학습하였다. 이렇게 미리 학습한 자기부호화기의 부호부(encoder) 부분을 사용하여 실시간으로 문맥 표상의 차원 크기를 100으로 축소하여 사용하였다.

3.3 GMM을 이용한 말뭉치에서의 오류 검출

GMM을 이용하기 위해서는 우선 각 표지별 k의 값을 사용자가 지정해야 한다. k의 값을 선택하는 방법은 여러 가지가 있지만 본 논문에서는 BIC(Bayesian Information Criterion)를 이용하여 k의 값을 선택한다.

k의 값을 선택한다는 것은 주어진 데이터에 최적화된 복잡도를 가지는 모델을 만드는 k의 값을 선택하는 작업이다. 즉, 입력된 데이터를 가장 잘 나타내는 k를 선택하는 것이다. 전체 데이터 N개를 가지는 데이터 집합 X에 대하여 k개의 가우시안 분포가 있을 때, X_N을 가장 잘 나타낼 수 있는 특정 k가우시안 분포를 선택해야 한다. 본 논문에서는 가우시안 분포를 1개부터 10개까지 변화시켰다. BIC는 다량의 데이터가 있을 때 우도함수 또는 사전확률이 혼합 가우시안 분포로 근사된다는 점에서 유도된다(Chen & Gopalkrishana, 1998).

$$BIC = -2\log(L) + k\log(N) \tag{3.1}$$

첫 번째 항에서 $L = p(x|\theta, M)$이며 후보 모델 M의 우도함수(likelihood function)이고, 여기서 θ는 모델의 매개변수이다. 즉 첫 번째 항은 데이터를 가장 잘 나타낼 수 있는 확률 모델을 찾는 성분이며, 적합도를 나타낸다. 두 번째 항인 $k\log(N)$은 모델 내의 파라미터 개수에 대한 페널티 항이다. 따라서 BIC는 두 항에 서로 조화되는 지점에서 최적의 모델이 만들어진다. 즉 우도와 모델 파라미터 개수가 조화를 나타내는 BIC의 최솟값에서 GMM을 선택한다.

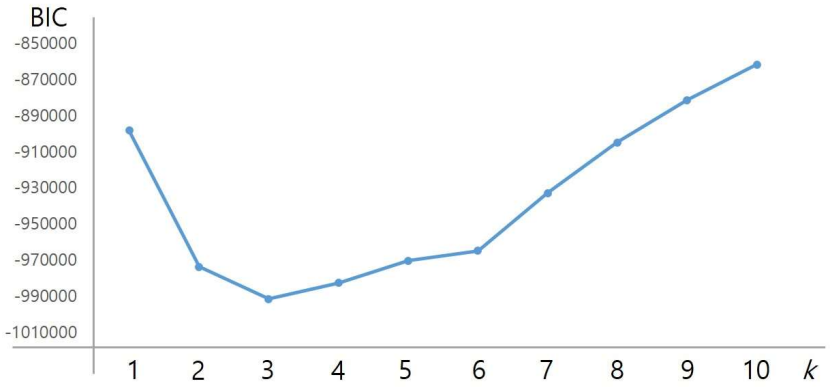

Figure 3.8 The graph of BIC scores of the "nsubj" tag according to the GMM k value changes from 1 to 10.

Figure 3.8은 본 논문에서 진행한 한 예로 구문분석 말뭉치 내의 존재하는 여러 표지 중 nsubj 표지에서의 1부터 10까지 각 GMM에 따른 BIC값 변화를 보여준다. 동일한 방법으로 구문분석 말뭉치, 의미역 분석 말뭉치에 존재하는 표지들의 k값을 각각 선택하였다.

k의 값이 선택된 이후 각 말뭉치에 존재하는 표지별로 선택된 k의 값을 바탕으로 군집화를 수행한다. 같은 군집 안에 속하는 데이터들이라도 각 데이터별로 특정 가우시안 분포에 속할 확률에는 차이가 존재한다. 본 논문에서는 기준값(threshold)을 정해 그 이하를 오류로 판단한다. 각 표지별로 기준값의 범위를 0.3부터 0.7까지 0.05씩 증가시켜 실험할 것이다. Figure 3.9 ~ 3.11은 각 기준값에 따라 달라지는 군집화 결과를 나타낸다. 실제 문맥 표상으로 예를 보여주기는 어려우므로 예시 단어를 통해 군집화 결과를 표현한다. 세 가지 경우 모두 가위표(×)는 정상(normal)이고 삼각형(△)은 오류(anomaly)이다. 또한 원의 경계를 벗어난 데이터들을 오류라고 판단한다. 따라서 가위표는 경계 내부에 존재해야 정답이고, 경계 외부에 존재하면 오답이다. 반대로 삼각형은 경계 외부에 존재해야 정답이고, 경계 내부에 존재하면 오답이다. 그 중 Figure 3.9는 기준값이 너무 높은 예제를 나타낸다. 오류 이외에도 대부분의 정답 데이터들이 경계를 벗어나 오류로 판단된 경우를 보여준다. 경계 내부에 데이터들이 모두 정답

이기 때문에 정밀도는 가장 높게 측정되지만, 많은 정답 데이터들이 경계 외부에 존재하기 때문에 재현율은 가장 낮게 측정된다.

Figure 3.9 An example of clustering with a high threshold value.

Figure 3.10은 기준값이 가장 적합한 예제이다. 대부분의 정답 데이터들은 경계 내부에 존재하고 반대로 대부분의 오류 데이터들은 경계 외부에 존재하는 가장 이상적인 경우를 나타낸다. 대부분의 정답 데이터들은 경계 내부에 분포하고 대부분의 오류 데이터들은 경계 외부에 분포하여 재현율과 정밀도가 이상적인 경우이다.

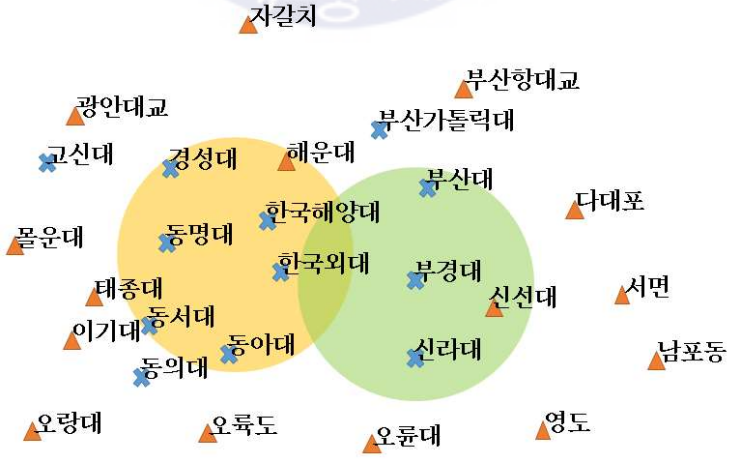

Figure 3.10 An example of clustering with the most proper threshold value.

Figure 3.11은 Figure 3.8과는 반대의 경우인 기준값이 가장 낮은 예제이다. 정답 데이터들뿐만 아니라 대부분의 오류 데이터들까지 경계 내부에 존재하여 오류를 제대로 탐지하지 못하는 경우를 나타낸다. 모든 정답 데이터들이 경계 내부에 분포하여 재현율은 가장 높게 측정되지만, 대부분의 오류 데이터들도 경계 내부에 분포하기 때문에 정밀도는 가장 낮게 측정된다.

Figure 3.11 An example of clustering with a low threshold value.

실제 GMM을 이용한 군집화에서 가위표는 표지가 제대로 부착된 문맥 표상을 의미한다. 반대로 삼각형은 정상적인 표지가 아닌 표지가 부착된 문맥 표상을 의미한다. 문맥 표상은 3.1절과 같이 제작하여 3.2절과 같이 축소하여 사용하였다. 이러한 문맥 표상들을 3.3절의 방법으로 각 표지별로 정해진 k의 값을 이용하여 군집화를 수행한다. 수행된 결과를 바탕으로 정상과 오류로 나눌 기준값을 정하였다. 실제 실험에서 사용한 데이터의 정보들과 실험을 통해 정해진 k, 기준값은 다음 장인 4장에서 자세하게 설명할 것이다.

제 4 장 실험 및 평가

이 장에서는 제안한 GMM을 이용한 군집화를 통해 수행한 오류 탐지의 성능 및 효용성 검증을 진행한다. 실험은 구문분석 말뭉치와 의미역 말뭉치를 이용하여 각각 진행하였으며, 그 결과를 바탕으로 제안한 방법의 효용성을 검증해 보고자 한다.

4.1 실험 데이터

본 실험의 목적은 말뭉치가 구축된 후 오류 탐지를 하는 것이 아닌, 말뭉치를 구축하는 과정에서 오류 탐지를 수행하는 것이다. 하지만 실험 데이터의 양이 너무 많은 경우 실험이 적합하지 않다. 따라서 실험의 목적에 맞게 실험 데이터는 일정 양을 초과하지 않도록 각 말뭉치 별로 1,000개의 문장을 사용하도록 전제 조건을 설정하였다.

실험에 필요한 오류 데이터는 각 말뭉치에서 표지별로 5%를 다른 표지로 임의 변경하여 생성하였다. 만약 표지별 5%가 1개보다 적을 경우 1개의 오류를 생성하였다. 전체 말뭉치에서 오류 데이터로 변경할 토큰이 선택되어 오류 데이터를 골고루 분포시켰다. 또한 표지별 오류 데이터의 비율을 정답 데이터의 비율에 맞춰 원본 데이터의 비율을 최대한 유지하였다. Table 4.1과 Table 4.2는 각각 구문분석 말뭉치와 의미역 말뭉치에서 오류 데이터를 생성하는 예시를 나타낸다. Table 4.1에서는 ID 2와 헤드의 원래 의존 관계(DEPREL<전>)가 acl이였지만 advcl로 변경되어 오류로 생성되었다. Table 4.2에서는 ID 4의 ARG1/3을 ARG2/3로 변경하여 오류를 생성하였다.

Table 4.1 An example of errors generated in Korean dependency corpus for testing.

ID	FORM	LEMMA	…	DEPREL<전>	DEPREL<후>	…
1	플라스틱으로	플라스틱 으로	…	obl	obl	…
2	만든	만들 ㄴ	…	acl	**advcl**	…
3	샤베트기는	샤베트기 는	…	nsubj	nsubj	…
4	수입품이	수입품 이	…	nsubj	nsubj	…
5	대부분이다.	대부분 이 다 .	…	root	root	…

Table 4.2 An example of errors generated in Korean SRL corpus for testing.

ID	FORM	LEMMA	…	MISC<전>	MISC<후>
1	그는	그 는	…	-	-
2	운전면허	운전면허	…	-	-
3	시험에	시험 에	…	-	-
4	합격하였다.	합격 하 였 다 .	…	(ARGO/1), (ARG1/3)	(ARGO/1), **(ARG2/3)**

Table 4.3은 사용한 말뭉치의 정보를 나타낸다. 구문분석 말뭉치는 2.4.2 절에서 언급한 바와 같이 (최용성 & 이공주, 2018)의 유연한 중심어 후위 원칙에 따라 의존구조로 변환한 말뭉치를 사용하였다. 데이터의 양은 미리 정한 전제 조건에 따라 1,000 문장 총 13,183개의 토큰을 사용하였다. Table 4.4의 정답 개수는 구문분석 말뭉치에서 정답 데이터로 사용한 개수를 각 표지별로 나타낸다. 의미역 말뭉치는 2.5.2 절에서 언급한 바와 같이 전북대 의미역 말뭉치(박광현 & 나승훈, 2017)를 사용하였다. 구문분석 말뭉치와 마찬가지로 전제 조건에 따라 1,000 문장 총 9,803개의 토큰을 사용하였다. Table 4.5의 정답 개수는 의미역 말뭉치에서 정답 데이터로 사용한 개수를 각 표지별로 나타낸다.

Table 4.4의 오류 개수는 Table 4.1과 같은 오류 생성 방법을 바탕으로 구문분석 말뭉치에서 생성한 각 표지별 오류의 개수이다. Table 4.5의 오류 개수는 Table 4.2와 같은 오류 생성 방법을 바탕으로 의미역 말뭉치에서 생성한 각 표지별 오류의 개수이다.

Table 4.3 Statistics of corpus used in test.

	구문분석 말뭉치	의미역 말뭉치
제작한곳	최용석, 이공주(충남대)	박광현, 나승훈(전북대)
문장 수	1,000	
토큰 수	13,183	9,803

Table 4.4 The distribution of each tag in dependency corpus.

표지명	정답 개수	오류 개수	표지명	정답 개수	오류 개수
nmod	2895	144	case	121	6
obl	1671	83	det	120	6
nsubj	1380	69	xcomp	65	3
acl	1346	67	appos	43	2
obj	1076	53	dep	43	2
advcl	991	49	amod	37	1
punct	716	35	mark	30	1
aux	601	30	csubj	24	1
conj	516	25	cc	12	1
advmod	410	20	parataxis	5	1
nummod	314	15	vocative	2	1
ccomp	184	9			

Table 4.5 The distribution of each tag in SRL corpus.

표지명	정답 개수	오답 개수	표지명	정답 개수	오답 개수
ARG0	1727	86	ARGM-CND	99	4
ARG1	3995	199	ARGM-MNR	297	14
ARG2	1056	52	ARGM-INS	181	9
ARG3	82	4	ARGM-TMP	1095	54
ARGM-LOC	274	13	ARGM-CAU	243	12
ARGM-DIR	12	1	ARGM-EXT	280	14

4.2 실험 결과

이 절에서는 구문분석 말뭉치와 의미역 말뭉치에서 실험한 각 표지별 k의 값과 기준값을 기술하고 k의 값과 기준값을 바탕으로 측정한 정밀도와

재현율을 바탕으로 평가하였다. 본 논문에서는 말뭉치 구축 작업 시 오류를 탐지하여 최대한 오류 발생 가능성을 낮추는 것을 목적으로 한다. 따라서 재현율을 정밀도보다 우선으로 설정하였다. Table 4.6은 구문분석 말뭉치에서 각 표지별 k의 값, 기준값, 정밀도와 재현율을 측정한 것이다. Table 4.7은 의미역 말뭉치에서 각 표지별 k의 값, 기준값, 정밀도와 재현율을 측정한 것이다.

Table 4.6 The Gaussian k value, threshold value, recall and precision of each dependency corpus tag according to the test result.

표지명	k	기준값	재현율	정밀도
nmod	5	0.65	0.64	0.53
obl	3	0.70	0.61	0.56
nsubj	3	0.60	0.65	0.54
acl	3	0.65	0.69	0.56
obj	3	0.65	0.64	0.53
advcl	3	0.6	0.59	0.51
punct	3	0.75	0.63	0.61
aux	2	0.75	0.63	0.54
conj	2	0.65	0.58	0.51
advmod	1	0.60	0.65	0.61
nummod	1	0.60	0.66	0.54
ccomp	1	0.75	0.61	0.53
case	1	0.65	0.63	0.57
det	1	0.75	0.63	0.58
xcomp	1	0.65	0.62	0.59
appos	1	0.70	0.57	0.51
dep	1	0.75	0.60	0.55
amod	1	0.70	1.0	0.53
mark	1	0.85	0	0.69
csubj	1	0.80	1.0	0.58
cc	1	0.80	1.0	0.58
parataxis	1	0.80	1.0	0.6
vocative	1	0.75	1.0	0.5
micro-average			65.15	56.39

Table 4.7 The Gaussian k value, threshold value, recall and precision of each SRL corpus tag according to the test result.

표지명	k	기준값	재현율	정밀도
ARG0	3	0.70	0.71	0.65
ARG1	9	0.60	0.68	0.64
ARG2	3	0.65	0.69	0.62
ARG3	1	0.55	0.62	0.57
ARGM-LOC	2	0.60	0.64	0.62
ARGM-DIR	1	0.80	1.0	0.61
ARGM-CND	1	0.55	0.58	0.53
ARGM-MNR	2	0.60	0.59	0.51
ARGM-INS	2	0.55	0.57	0.50
ARGM-TMP	3	0.65	0.60	0.53
ARGM-CAU	2	0.65	0.58	0.53
ARGM-EXT	2	0.60	0.56	0.51
micro-average			69.46	62.64

미시 평균(micro-average)의 결과 구문 말뭉치는 65.15%, 의미역 말뭉치는 69.46%의 재현율을 보였다. 구문 말뭉치와 의미역 말뭉치에서 각각의 표지를 비교해 보면 비교적 숫자가 낮은 표지들은 수치가 낮게 측정되었다. 이 결과를 확인하기 위하여 실험의 사용한 데이터의 양을 바꿔 실험하였다.

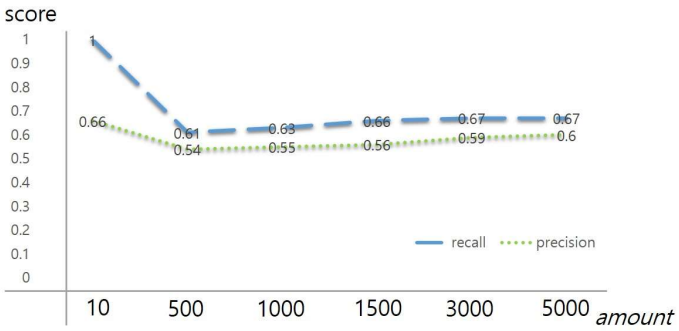

Figure 4.1 The recall and precision for “nusbj” tag on dependency corpus of different amount.

Figure 4.1은 구문분석 말뭉치의 “nusbj” 표지의 전체 양을 바꿔서 실험한 결과이다. 오류 생성 방법이나 개수는 이전 실험과 동일하게 설정하였다. 전체 개수가 너무 적은 경우는 오류의 개수가 1개이므로 재현율이 0 또는 1이 나온다. 이런 경우 결과의 신뢰성이 떨어진다. 반대로 너무 많은 경우 k의 값과 군집화를 수행하는 시간이 오래 걸린다. 또한 재현율과 정밀도 값의 변화가 기준값의 영향을 받을 뿐 큰 격차가 존재하지 않는다. 따라서 어느 정도 전체 개수의 일정 범위 내에서 효율을 보인다.

제 5 장 결론 및 향후 연구

본 논문에서는 말뭉치의 오류를 효율적으로 탐지하기 위해 말뭉치를 구축하는 단계에서 오류를 탐지할 수 있는 방안을 제시하였다. 말뭉치에서 오류를 탐지하는 기존의 연구들은 이미 구축된 말뭉치에서 오류를 찾는 방식이 일반적이다. 이러한 방식은 말뭉치가 완전히 구축된 후에야 오류를 탐지할 수 있으며, 시간 및 비용적 측면에서 비효율적이다.

따라서 본 논문에서는 이러한 문제점을 해결하기 위해 말뭉치를 구축하는 단계에서 오류를 탐지할 수 있도록 GMM 알고리즘을 이용한 군집화를 제안하였다. 군집화는 비지도학습의 한 종류로 표지를 부착하여 학습데이터를 구축하는 시간을 단축할 수 있다. 또한 비지도학습의 특징을 가지고 있어 말뭉치를 구축하는 단계에서 바로 오류 탐지를 수행할 수 있는 특징을 가진다.

이러한 특징을 최대한 유지하기 위해서 본 논문에서는 문맥 정보를 반영하는 문맥 표상을 만들고 실시간으로 문맥 표상을 이용하기 위해 차원 축소를 수행하였다. 이렇게 축소한 문맥 표상을 이용하여 GMM 알고리즘을 이용하여 군집화를 수행한다. 군집화는 비슷한 특성을 가지는 데이터끼리 공통된 군집으로 모으는 것을 의미하며, 이 과정을 통해 하나의 군집으로 모이지 않거나 군집에 모였으나 경계 근처에 위치한 데이터를 확인할 수 있다. 이러한 데이터를 본 연구에서는 표지 부착 오류로 판단한다. 이러한 과정을 통해 오류 탐지를 수행하였고 검증을 위해 구문분석 말뭉치와 의미역 말뭉치를 사용하여 재현율과 정밀도를 확인하였다.

미시 평균 결과 구문분석 말뭉치에서는 65.15%, 의미역 말뭉치에서는

69.46%의 재현율을 보였다. 이는 초기 말뭉치 구축 작업에서 어느 정도의 효율을 가진다고 판단된다. 또한, 각 말뭉치에서 문맥 표상을 제작하는 방법 이외에 전체적인 모델은 구문분석 말뭉치와 의미역 말뭉치가 동일하다. 이는 하나의 모델을 가지고 여러 가지 말뭉치의 오류를 탐지할 수 있는 것을 의미한다.

다만, 그럼에도 불구하고 성능 향상은 필요할 것으로 보인다. 본 연구에서 문맥 표상을 만드는데 이용한 정보 이외에 다른 정보를 추가하여 문맥 표상을 제작한다면 어느 정도의 성능 향상이 있을 것으로 예상된다. 또한 말뭉치의 양에 따른 성능 변화의 연구도 필요하다고 생각한다.

참고문헌

Abney, S. P. (1996). “Part-of-speech and partial parsing” *Corpus-Based methods in language and Speech Processing, eds*. Young, S and Bloothooft, G., Kluwer Academic Publishers, pp. 118-173.

Agovic, A., Banerjee, A., Ganguly, A. R., and Protopescu, V. (2007). “Anomaly detection in transportation corridors using manifold embedding”, *Proceeding of the 1st International Workshop on Knowledge Discovery from Sensor Data*, pp. 435-455

Aleskerov, E,. Freisleben, B., and Rao, B. (1997). “Cardwatch: A neural network based database mining system for credit card fraud detection”, *Proceeding of IEEE/IAFE 1997*, pp. 220-226.

Baldi, P. (2012). “Autoencoders, Unsupervised Learning, and Deep Architectures”, *Proceeding of the 2011 International Conference on Unsupervised and Transfer Learning workshop*, vol. 27, pp. 37-50.

Barbara, D., Couto, J., Jajodia, S., and Wu, N. (2001). “ADAM: A testbed for exploring the use of data mining in intrusion detection”, *ACM Sigmod Record*, vol. 30, no. 4, pp. 15-24.

Breunig, M. M., Kreigel, H., Ng, R. T., and Sander, J. (1999). “Optics-of: Identifying local outliers”, *Proceeding of the European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 262-270.

Bybers, S., and Raftery, A. E. (1998). “Nearest-neighbor clutter removal for estimating features in spatial point processes”, *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 572-584.

Chen. S. S and Gopalkrishana. P. S (1998) “Speaker, enviroment, and channel change detection and clustering via the Bayesian information criterion,” *Proceedings of the IEEE Interational Conference*, vol. 2, pp. 645-648.

Chinchor, N. and Sundheim, B. (1993) “MUC-5 Evaluation Metrics”, *Proceedings of the 5th Message Understanding Conference*, pp. 69-78.

Choi, K., Han, Y., Han, Y., and Kwon. O. (1994). “KAIST Tree Bank Project for Korean: Present and Future Development”, *Proceedings of the International Workshop on Sharable Natural Language Resources*, pp. 7-14.

Comaniciu, D., Meer, O. (2002). “Mean Shift: A Robust Approach Toward Feature Space Analysis”, *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 24, no. 5. pp. 603-619.

De Stefano, C., Sansone, C., and Vento, M. (2000). “To reject or not to reject: That is the question-an answer in case of neural classifiers”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 30, no. 1, pp. 84-94.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). Pattern Classification, John Wiley & Sons

Edward, L. and Steven, B. (2002). “NLTK: the natural language toolkit”, *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.* vol. 1, pp. 63-70.

Eisner, J**.** (1996), “Three new probabilistic models for dependency parsing: An exploration”, *Proceeding of The 16th International Conference on Computational Linguistics*. pp. 340-345.

Ester, M., Kreiegel, H., Sander, J. and Xu, X. (1996). “A density-based algorithm for discovering clustering in large spatial databases with noise”, *Proceeding of the Knowledge Discovery and Data mining 1996*, vol. 96, no. 34, pp. 226-231,

Fan, W., Miller, M., Stolfo, S., Lee, W., and Chan, P. (2004). “Using artificial anomalies to detect unknown and known network intrusions”, *Knowledge and*

Information Systems, vol. 6, no. 5, pp. 507-527.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). "Pulse: Mining customer opinions from free text", *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis.* pp. 121-132.

Han, C., Han, N., Ko, E., Palmer, M. and Yi, H. (2002). "PENN Korean Treebank: Development and Evaluation", *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation.* pp. 69-78.

Hartigan. J. A., and Wong, M. A. (1979). "Algorithm AS 136: A K-means clustering algorithms", *Journal of the Royal Statistical Society. Series C*, vol. 28, no. 1, pp. 100-108.

He, Z., Deng. S., and Xu, X. (2003). "An optimization model for outlier detection in categorical data", *Proceeding of the International Conference on Intelligent Computing*, pp. 400-409.

Jundong Li, Xia Hu, Liang Wu., and Liu. H. (2016). "Robust unsupervised feature selection on networked data", *Proceeding of SDM*, pp. 387–395.

Kumar, V. (2005). "Parallel and distributed computing for cybersecurity", *IEEE Distributed Systems Online*, vol. 6, no. 10.

Lee, W., and Xiang, D. (2001). "Information-theoretic measures for anomaly detection". *Proceedings of IEEE Symposium On Security and Privacy*, pp. 130-143.

Lin, S., and Brown, D. E. (2006). "An outlier-based data association method for linking cirminal incidents", *Decision Support Systems*, vol. 41, no. 3, pp. 604-615.

Nivre, J. (2003). "An efficient algorithm for projective dependency parsing", *Proceedings of the 8th International Workshop on Parsing Technologies*, pp. 149-160.

Nivre, J. (2004). "Incrementalityh in deterministic dependency parsing", *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering*

and Cognition Together. pp. 50-57.

Palmer, M., Daniel, G., Paul, K. (2005). "The proposition bank: An annotated corpus of semantic roles", *Computational Linguistics* vol. 31, no. 1, pp. 71-106.

Smith, R., Bivens, A., Embrechts, M., Palagiri, C., and Szymanski. B. (2002). "Clustering approaches for anomaly based intrusion detection", *Proceedings of Intelligent Engineering Systems through Artificial Neural Networks*, pp. 579-584.

Spence, C., Parra, L., and Sajda, P. (2001). "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model", *Proceeding of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, vol. 3.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society. Series B*, pp. 267–288.

Mikolov, T., Chen, J., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space", *Proceedings of International Conference Learning Representations*, arXiv:1301.3781.

Yu, D., Sheikholeslami, G., and Zhang, A. (2002). "Findout: Finding outliers in very large datasets", *Knowledge and Information Systems*, vol. 4, no. 4, pp. 387-412.

Yutaka, S. (2007). "The truth of the F-measure", *Teach Tutor mater.* vol. 1, no. 5, pp. 1-5.

Zhao, Z., and Liu, H. (2007). "Spectral feature selection for supervised and unsupervised learning", *Proceeding of International Conference on Machine Learning*, pp. 1151–1157.

김병수, 이용훈, 이종혁. (2007). "비지도 학습을 기반으로 한 한국어 부사격의 의미역 결정", *정보과학회논문지: 소프트웨어 및 응용*, vol. 34, no. 2, pp. 112-122.

김영길, 양성일, 홍문표, 박상규. (2003). "형태소 어휘 문맥에 기반한 태깅 오

류 정정", *제15회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 63-68.

김홍규, 강범모, 홍정하. (2007). "21세기 세종계획 현대국어 기초말뭉치: 성과와 전망", *한글 및 한국어 정보처리 학술대회 발표논문집*. pp. 311-316.

김재훈. (2000b). "한국어 부분 구문분석의 단위와 그 표지", *한국해양대학교 컴퓨터공학과*, 기술문서, KMU-NLP-TR-2000-006.

김재훈, 김길창. (1995). "한국어에서의 품사 부착 말뭉치의 작성 요령: KAIST 말뭉치", *한국과학기술원 전산학과*, 기술문서, CS-TR-95-99.

나승훈, 이건일, 신종훈, 김강일. (2017). "Deep Biaffine Attention을 이용한 한국어 의존 파싱", *한국정보과학회 학술발표논문집*, pp. 584-586.

박은정. (2015). "한국어와 NLTK, Gensim의 만남", *PyCon Korea 2015*.

박광현, 나승훈. (2017). "Attention 기반 한국어 의미역 결정", *한국정보과학회 학술발표논문집*, pp. 634-636.

배장성, 이창기. (2015). "한국어 의미역 결정의 위한 Korean Propbank 확장 및 도메인 적응 기술 적용", *인지과학*, vol. 26, no. 4, pp. 377-392.

배장성, 이창기, 임수종. (2015). "딥 러닝을 이용한 한국어 의미역 결정", *한국정보과학회 학술발표논문집*, pp. 690-692.

배장성, 오준호, 박천음, 최경호, 이창기. (2014). "한국어 의미역 말뭉치 구축을 위한 반자동 태깅 도구 개발", *한국정보과학회 학술발표논문집*, pp. 592-594.

서형원, 이공주, 류길수, 김재훈. (2010). "뉴스 댓글의 감정 분류를 위한 자질 가중치 설정", *한국마린엔지니어링학회지*. vol. 34, no. 6, pp. 871-879.

신준철, 옥철영. (2012). "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기", *정보과학회 논문지, 소프트웨어 및 응용*. vol. 39, no. 5, pp. 415-424.

양준호, 배용진, 김현기, 김윤정, 이규철. (2015). "의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치", *제27회 한국어 정보처리 학술대회 논문집*, pp. 234-239.

이건일, 이종혁. (2015). “순환 신경망을 이용한 전이 기반 한국어 의존 구문 분석”, *정보과학회 컴퓨팅의 실제 논문지*, vol. 21, no. 8, pp. 567-571.

이미경, 정한민, 성원경, 박동인. (2005). “품사 표지 부착 말뭉치 검증”, *제17회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 145-150.

이정규, 이상주, 임희석, 임해창. (1997). “규칙 기반 한국어 품사 태깅을 위한 어휘 규칙 획득의 수작업 최소한 방안”, *한국정보과학회 학술발표논문집*, vol. 24, no. 1B, pp. 479-482.

이창기, 임수종, 김현기. (2014). “Structural SVM 기반의 한국어 의미역 결정”, *한국정보과학회 학술발표논문집*, pp. 574-576.

최명길, 서형원, 권홍석, 김재훈. (2013). “한국어 품사 부착 말뭉치의 오류 검출 및 수정”, *한국마린엔지니어링학회*, vol. 37, no. 2, pp. 227-235.

최용석, 이공주. (2018). “한국어 구절 구문 코퍼스의 의존 구문 구조 트리로의 변환에서 중심어 전파 규칙”, *한국정보과학회 언어공학연구회 학술발표논문집*, pp. 514-519.

최용석, 이공주. (2019). “고차원 정보와 스택-포인터 네트워크를 이용한 한국어 의존 구문 파서”, *정보과학회논문지*, vol. 46, no. 7, pp. 636-643.

홍진표 (2013). “품사 태거와 빈도 정보를 활용한 세종 형태 분석 말뭉치 오류 수정”, *정보과학회논문지*, vol. 40, no. 7, pp. 417-428.

감사의 글

먼저 학부 생활부터 석사 과정을 지내는 동안 많은 부족함을 가지고 있는 제가 아낌없는 격려와 가르침으로 지도해주신 김재훈 지도교수님께 진심으로 감사드립니다. 또한 바쁘신 와중에도 논문지도 신경을 써주시고 가르침을 주신 박휴찬 교수님과 류길수 교수님께도 감사드립니다.

학부 생활부터 대학원 생활까지 항상 신경 써주시고 챙겨주신 강군호 조교님, 김경언 조교님께도 감사의 말씀을 드립니다.

석사 과정 동안 같이 자연어처리연구실에서 생활을 했던 연구실의 기둥인 천민아님, 박호민님, 같이 들어왔지만 동생인 저를 잘 챙겨주신 남궁영님, 윤호님, 늦게 들어왔지만 선배로서 잘 챙겨주신 김재균님, 그리고 같이 생활하며 즐겁게 어울려준 신영진님께 감사드립니다. 같은 연구실은 아니였지만 수학을 함께 하며 어려울 때마다 도와주신 김정래님, 여동규과 이정욱님께도 감사의 말씀을 드립니다.

마지막으로 석사 과정을 마칠 때까지 묵묵히 지켜봐주시고 응원해주신 사랑하는 부모님과 누나에게 감사의 말을 전합니다.