

# 언어의 최소 의미단위 결합분석을 통한 언어사용의 본질적 이해

오세인, 김민경

선문대학교 AI소프트웨어학과

{ohshane71, minkyounkim}@sunmoon.ac.kr

## Analysis of Combinatory Semantic Units for Understanding Language Usage

Shane Oh and Minkyoun Kim

Department of Artificial Intelligence and Software Engineering, Sunmoon University

### 요약

최근 중국어 기계번역 분야에서 한자 분해 방법을 적용하여 번역 성능을 향상시키는 연구가 진행되었으나, 분해 수준이 부수 및 획수에 그쳐 표의문자의 특성을 충분히 반영하지 못하고 있다. 이에 본 연구는 표의문자의 그림적 특성을 고려하여 각 글자의 구성 요소와 배치 형태 중에서 최소 의미 단위를 도출하고, 글자의 의미 해석이 어떻게 이루어지는가를 분석함으로써 언어 사용에 대한 보다 본질적 이해에 접근하고자 한다.

### I. 서론

동아시아 문화권은 한자를 자국의 언어 체계에 도입하거나 과거에 차용한 적이 있어서 현재에 한자를 직접적으로 사용하지 않더라도 한자와의 연관성이 깊다. 한자를 정의한 유니코드 상에서도 CJK(Chinese Japanese Korean)로 표기하고 있다는 사실은, 한자가 중국어에만 국한된 언어가 아님을 확인할 수 있다. 한국어는 조선 시대 전기까지만 해도 한자를 원형의 형태로 표의문자 및 표음문자로도 사용하였으나, 현재에는 한자와 한글 혼용의 형태로 발전하였다. 또한 일본어의 경우 표의문자인 한자와 함께 표음문자인 히라가나와 가타카나를 동시에 사용하여 언어를 구성하고 있다. 이 중 히라가나와 가타카나 또한 9세기경 한자의 초서체 형태가 발전하여 표음문자가 되어 지금까지 사용되고 있다. 이렇게 한자에 기반하여 각국의 언어가 현재의 형태로 발전하였음을 알 수 있다.



또한 한자는 여러 개의 작은 한자가 결합하여 생긴 문자기도 하다. 예를 들어 수풀 림(林)이라는 한자는 나무 목(木)이 두 개 결합한 모습이며 수풀 삼(森)은 나무 목(木)과 수풀 림(林)이 결합된 모양을 하고 있다. 이 중에서 뜻 부분이 비슷한 구성 요소를 추출하는 것이 부수(Radical)이다. 위의 한자를 예시로 보면 林과 森은 공통적으로 ‘빼곡한 나무’라는 뜻이 있으며 이에 따라 부수 또한 ‘나무 목’(木)이다. 이와 같은 특성으로 인하여 모든 한자는 여러 구성 요소의 조합으로 나타낼 수 있으므로, 시리얼라이징(Serializing)을 적용한 기계번역 방법론이 대두되고 있다. 기계번역(Machine Translation)은 기계학습(Machine Learning)에 기반한 대표적인 연구 분야이며 다양한 언어들 간의 번역을 위해 활용되고 있다. 최근 신경망을 이용한 기계번역은 신경망 네트워크의 입력 값으로 문장뿐만 아니라, 각 문장을 구성하고 있는 글자들의 부수 시퀀스를 이용함으로써 번역 성능의 향상을 가져왔다[1].

하지만, 기존 연구에서는 표의문자를 구성하는 원소들의 결합적 특성을 반영하지 못해 글자 본연의 뜻을 해석하기에는 부족함이 있다. 이에, 본 연구는 한자의 각 글자를 구성하는 최소 의미단위를 도출하고, 이들 간의 결합적 분석을 통해 기계번역의 정확성 향상을 위한 새로운 특성 제시 및 언어사용에 대한 본질적 이해에 접근하고자 한다. 이 연구를 통해서, 비단

한자뿐만 아니라 결합적 특성을 가진 언어들의 본질적 이해에 도움을 줄 수 있을 것이라 기대한다.

### II. 선행연구

한자로 구성된 문장을 분해할 수 있는 구성적 단위는, ‘단어 단위’(Character level)에서부터 ‘부수 단위’(Radical level), 또는 ‘획 단위’(Stroke level)로 세분화할 수 있다. 이러한 한자의 성분 분해는 기계번역에서 주요 관심사가 되어 왔으며, 한자 부수를 이용한 기계번역을 중·영, 일·중 번역에 적용한 결과 문장 번역의 정확도가 향상되었다[1]. 또한, [2]에서는 단어+글자, 단어+부수, 단어+글자+부수의 조합 중에서 단어+글자+부수의 조합에서 가장 높은 정확도를 도출하였다. 이에, [3]은 일·중 번역에서 글자와 부수의 조합으로 성능을 향상시켰고, 비슷하게 [4]는 중·영 번역에서는 부수 단위, 일·영 번역에서는 획 단위 분해를 적용하여 최고의 성능을 보여주었다. 이는, 한자의 성분 분해 수준이 기계번역 성능에 영향을 미치며, 따라서 기계학습에서 중요한 특성으로 적용 가능성을 시사한다. 하지만, 한자의 본질적 구성 방식인 ‘그림 결합’적 측면에서 성분 분해를 바라본다면 부수는 한자의 정리 및 배열의 효율성을 위해 사용되고 있기 때문에 한자의 근본적인 뜻을 내포하고 있다고 보기 어렵다. 따라서 한 글자당 하나의 부수로 해당 글자의 특성을 정의한다면 정보손실(Information loss)의 우려가 있다. 예를 들어 ‘아닐 불/부(不)’의 경우 부수가 ‘한 일(一)’이지만 글자의 뜻에 비추어 보았을 때 크게 연관성이 있지 않다. 또한 획 기반 분해법을 사용하였을 때도 한자의 근본적인 뜻을 파악하기 힘들 정도로 분해가 되어 해석 가능성이 사라진다.

한자 분해 방법에 관한 연구도 진행되어왔다. 대표적으로 한자의 부수를 중심으로 분해한 시퀀스인 IDS(Ideographic Description Sequences)가 있다[1]. IDS는 유니코드 중 결합 모양을 나타내는 IDC(Ideographic Description Characters)를 이용한 시퀀스로 정의하며, 각 시퀀스는 결합 모양(IDC), 부수, 해당 부수를 제거한 나머지 한자의 모양으로 구성되어 있다. 예를 들어 ‘옴을 가(可)’의 경우 IDS는 ‘口丁’가 되며, 이의 IDC는 이며 부수는 ‘입구(口)’, 나머지 한자는 ‘고무래 정(丁)’이 됨을 의미한

다. 또한 더 이상 분해가 되지 않는 한자의 IDS는 IDC를 사용하지 않고 해당 한자가 IDS가 된다. 예를 들어 ‘말마(馬)’의 경우 자신이 부수이므로 IDS는 최종적으로 ‘馬’가 된다. 다른 방법으로는 획 기반 분해법으로, 한자의 글자를 구성하는 모든 획을 필획 순으로 순차적으로 나열한 모양이다. 예를 들어 ‘옴을 가(可)’의 경우 ‘一丨ㄣ丨’의 형태로 분해한다.

기계번역은 통계 기반[1]에서 인공지능경망 기반[1]의 기계학습으로 발전하면서 월등한 성능적 향상을 가져왔으나, 표의문자의 그림적 특성을 반영한 의미 해석은 아직 매우 어려운 문제이다. 이에, 본 연구에서는 한 글자를 구성하는 구성 요소의 집합 및 배치 형태 중에서 의미를 가질 수 있는 최소 단위를 도출하여 해석 방법을 제안하고자 한다.

### III. 방법

위에서 살펴본 바와 같이, 한자의 IDS는 IDC를 비롯한 유니코드(부수 또는 한자)로 구성되어 있다. 이때, IDS를 구성하는 한자에 대한 IDS가 정의되어있는 경우에는 기존의 한자대신 IDS로 대체함으로써, 각 한자의 시퀀스를 재구성할 수 있는 RIDS(Recursive IDS)를 제안하고자 한다(그림1 참조). 이는, 한자의 유니코드를 더 이상 분해할 수 없는 IDS들로 구성함으로써, 의미를 내포하고 있는 최소 단위 간의 결합 관계를 다양한 관점에서 분석할 수 있는 가능성을 제공하기 위함이다.



그림1. IDS와 RIDS

그림1과 같이, 수풀 삼(森)의 IDS는 ‘木木木’이지만 이 중 ‘林’은 IDS가 정의되어있으므로 이를 ‘木木’으로 치환하면 ‘木木木’이 된다. 해당 시퀀스에서는 더 이상 분해 가능한 유니코드가 없으므로 재귀적 분해를 종료하고 최종적으로 ‘森’의 RIDS는 ‘木木木’가 된다. 이런 방법으로 전체 한자에 대해서 RIDS를 재구성한다.

다음으로 CC-CEDICT 한자 사전 데이터셋[5]에 정의되어있는 한자의 속성 중 상형문자(Ideographic character)기반 한자들이 들어있는 집합(CJK)과 이들을 구성하고 있는 RIDS의 유니코드들을 한 집합에 넣은 RIDS 집합을 구성한다. 이렇게 구성된 두 개의 집합에서, 한자(CJK) 집합 원소와 RIDS 유니코드 집합 원소 간의 연결 관계를 가중치를 적용한 이분 그래프(Weighted bipartite graph)로 정의한다(그림2 참조).

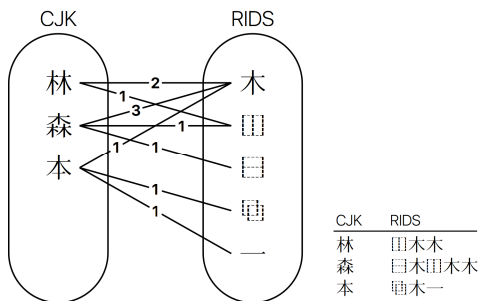


그림2. CJK원소와 RIDS원소의 가중치를 적용한 이분 그래프

그림2에서 ‘林’의 경우 RIDS는 ‘木木’이며 ‘木’은 한 번 사용되었으므로, ‘林’과 ‘木’의 연결 가중치는 1, ‘林’과 ‘木’의 연결 가중치는 2가 된다. 한자 집합(CJK)과 RIDS 집합 간의 연관관계를 인접행렬로 표현하고 투영(Projection)시킴으로써, RIDS 집합 원소들 간의 연결 관계 그래프  $G_{RIDS}$ 를 도출한다.

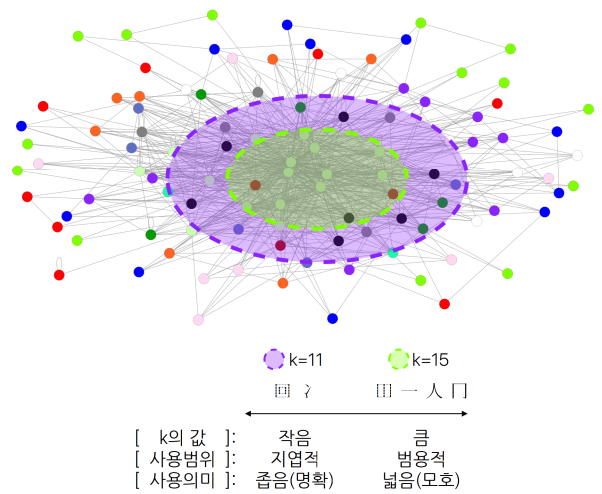


그림3.  $G_{RIDS}$ 의 최대로 연결된 그래프에 대한 k-core 분석

그림3은 위에서 도출한  $G_{RIDS}$ 에서 최대로 연결된 그래프(Giant Component)에 대해 k-core 분석 결과를 나타낸다. 최댓값 k=15 클러스터에 속하는 원소들은 ‘田’, ‘一’, ‘人’, ‘口’가 있으며, 차상위 k=11 클러스터에 속하는 한자는 ‘口’, ‘?’ 등이 있다. 즉, k 값이 커질수록 보편적으로 많이 사용되는 원소들이 등장하였고, 각 RIDS 원소들이 가진 본래의 뜻으로 사용되는 한자들을 도출하기는 어려웠다. 반면에 k의 값이 작아질수록 특정 한자들에 사용되는 원소들이 발견되었으며, 이들 원소들로 구성된 한자들은 비교적 비슷한 의미를 공유하고 있음을 알 수 있었다. 예를 들어 ‘口’, ‘?’와 같은 원소를 가진 한자의 경우 테두리/틀의 뜻, 차가움의 뜻을 지니는 경우가 많았다. 이는, 사용 범위가 넓은 원소일수록 본래의 뜻을 유지하기보다는 다른 의미로 파생되는 경우가 많다는 것을 알 수 있다.

### IV. 결론

본 연구에서는 한자 시퀀스의 재구성(RIDS)을 통해, 기존에 정의된 한자 부수 이외에 중심에 있는 새로운 의미적 구성 요소를 도출할 수 있었다. 한자의 정리나 배열을 위해 사용되는 부수에 비해 보다 본질적인 뜻을 내포하고 있는 구성 요소로서 가치가 있다. 또한, 신경망 기계번역에서 기존의 획 수준(Stroke level) 분해에 기반한 입력값 대신 본 연구의 결과를 토대로 해석 가능한 특성을 입력 값으로 활용할 수 있는 가능성을 제시하였다. 더 나아가, 한자문화권을 포함한 다른 지역의 언어 중에서 표의문자의 음가가 남아 발전된 표음문자 형태의 언어들에 대해서도 확장·적용한다면 번역의 성능향상을 기대해볼 수 있다.

### 참 고 문 헌

- [1] L. Han, et al., 2021. Chinese Character Decomposition for Neural MT with Multi-Word Expressions. NoDaLiDa.
- [2] L. Han, et al., 2018. Incorporating chinese radicals into neural machine translation: Deeper than character level. FoLLI.
- [3] J. Zhang, et al., 2017. Improving character-level japanese-chinese neural machine translation with radicals as an additional input feature. IALP.
- [4] L. Zhang, et al., 2018. Neural machine translation of logographic language using sub-character level information. Assoc. for Computational Linguistics.
- [5] <https://cc-cedict.org/>