지능형 한글 편집기의 구현

이규은, 김황수
경북대학교 전자계산학과

An Implementation of Intelligent Korean Editor

Kyueun Yi, Hwangsoo Kim
Department of Computer Science, Kyungpook National University

요 약

본 논문에서는 한글 입력시 중성 다음에 입력되는 자음이 현재 글자의 종성인지 다음 글자의 초성인지를 예측하는 능력을 가지는 지능형 한글 편집기를 구현하고 성능을 검사하였다. 지능형 한글 편집기는 한글 입력시 어색한 단어를 화면에 나타내지 않고, 마치 사람이 쓰는 것과 같이 화면에 글자를 나타내므로 사용자의 생각과 화면에 나타난 글자와의 차이를 없게 하여 사용자에게 편안함과 자연스러움을 느끼게 하는 한글 편집기이다. 지능형 한글 편집기는 확률, 사전, 조사표, 어미표, 그리고 문법 지식을 이용한다.

Ⅰ. 서론

컴퓨터의 사용 범위가 넓어 지고, 개인용 컴퓨터의 보급이 대중화되면서 컴퓨터를 보다 편리하게 사용하기 위한 노력이 날로 증가하고 있다. 그러한 노력중의 하나가 컴퓨터에서 문서를 보다 편리하게 입력하기 위한 노력이다. 영어에서는 입력 속도를 증가시키고 신체적 장애를 가진 사람들이 컴퓨터를 보다 편리하게 사용할 수 있도록 도와 주는 편집기들이 개발중이다[1]. 그러나 컴퓨터에서 한글을 사용하기 위한 노력은 주로 기계 번역을 위한 형태소 분석등에 치중되었고[5,6,7], 여러 한글 편집기가 개발되었지만 이들은 한글의 특성중의 하나인 모아쓰기를 효과적으로 지원하지는 않는다. 한글은 영어의 풀어 쓰기와는 달리 모아쓰기를 하므로 이를 효과적으로 지원하는 한글 편집기는 사용자에게 보다 자연스러움과 친밀함을 느끼게 할 것이다[15].

본 논문에서는 한글 입력시 중성 다음에 입력되는 자음이 현재 글자의 종성인지 다음 글자의 초성인지를 예측하는 능력을 가지는 지능형 한글 편집기를 구현하고 성능을 검사하였다. 지능형 한글 편집기는 한글 입력시 어색한 단어를 화면에 나타내지 않고, 마치 사람이 쓰는 것과 같이 화면에 글자를 나타내므로 사용자의 생각과 화면에 나타난 글자와의 차이를 없게 하여 사용자에게 편안함과 자연스러움을 느끼게 하는 한글 편집기이다.

앞으로 2절에서는 지능형 한글 편집기가 사용할 수 있는 국어의 특성들을 설명하고 3절에서는 지능형 한글 편집기 설계와 구현 방법을 설명하며, 4절에서는 구

본 연구는 1991년도 한국학술진흥재단의 자유공모과제 학술 연구 조성비에 의하여 연구되었음

현된 편집기의 성능을 검사한 결과를 설명하고 5절에서는 결론과 지능형 한글 편집기의 성능 향상을 위한 방안과 앞으로의 개발 방향을 제시하겠다.

Ⅱ. 국어에 대한 고찰

이 절에서는 국어의 특성을 설명하고 지능형 한글 편집기를 위한 품사의 분류 등을 설명한다.

2.1 국어의 특성

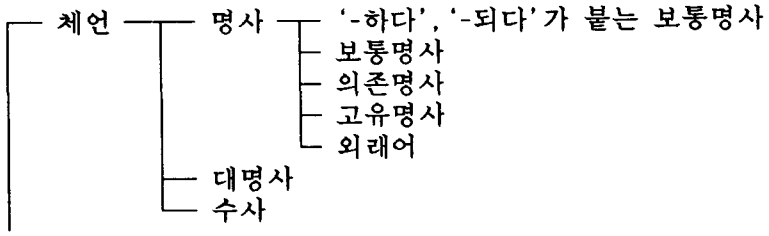
국어는 우랄 알타이 계통의 교착어로서, 어근을 중심으로 뜻을 더하거나 품사를 바꾸는 접사나 어미가 여러 개 붙어서 단어가 이루어지는 첨가적 성격을 띤 언어이다[8]. 그러므로 지능형 한글 편집기를 구현하기 위해 고려해야 할 국어의 특성은 다음과 같다.

① 한글의 최소 단위는 자음과 모음이며 자모음이 조합되어 단음절을 형성하고, 단음절이 모여 단어를 구성한다.
따라서, 컴퓨터에서 한글 처리시 한글 코드는 조합형이 바람직하고 본 편집기가 가정하는 입력 문서는 상용 조합형 한글 코드이다.
② 조사는 중복하여 사용이 가능하고 그 쓰임 또한 다양하다.
조사 부분의 종성을 예측하기 위해 여러 조사의 결합형태를 고려해야 한다.
③ 용언과 어미의 활용이 있다.
용언은 뒤에 붙는 어미에 따라 어간이나 어미, 또는 어간과 어미가 모두 변화한다. 그러므로 활용하는 부분의 종성을 예측하기 위해 활용 규칙을 이용한다.

문법적 형태소는 반드시 어근이나 어간 뒤에 오므로 (조사는 체언뒤에 붙어 쓰이며 활용어미는 용언의 어간뒤에 쓰인다) 한 어절이 입력중일 때 철자의 변화가 없는 어근이나 용언의 어간 부분은 사전에 담고, 비교적 많이 사용되고 복잡한 규칙에 따라 결합하는 조사나 활용을 하는 어미에 관한 정보는 분리하여 둔다. 그래서 어근이나 어간부분의 종성은 사전을 이용하여 예측하고, 조사와 어미 부분의 종성은 조사표와 어미표를 이용하여 예측한다.

2.2 품사의 분류

국어에 대한 국어 학자들의 품사 분류도 다양하지만 컴퓨터에서 국어를 처리하기 위한 기존의 논문들에서도 각각 독자적인 품사 분류를 하고 있다 [5,6,7]. 국어는 각 품사가 가지는 고유의 성질에 따라 조사의 결합과 어미의 활용등 어절의 형성에 영향을 미치므로, 지능형 한글 편집기를 위한 사전 구성을 위하여 새로운 품사 분류가 필요하고 이에 따라 품사 분류를 그림 1과 같이 한다. 이 분류는 단어의 쓰임에 따라 세분화한 것이 특징이다.



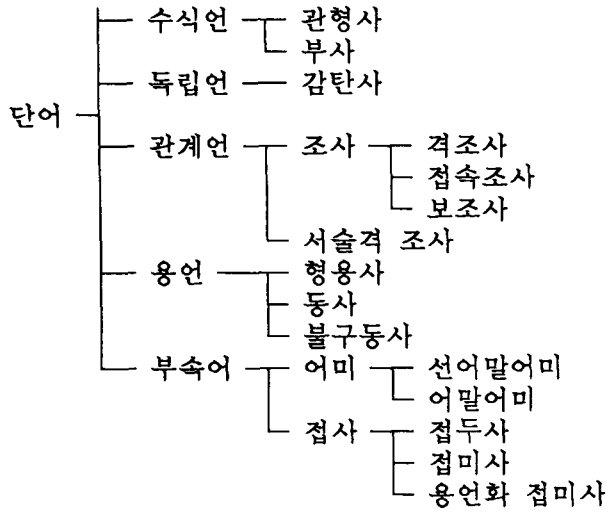


그림 1 지능형 한글 편집기를 위한 품사표

Ⅲ. 지능형 한글 편집기의 설계 및 구현

3.1 지능형 한글 편집기의 구성과 기본 동작

전체적인 구조는 그림 2와 같이 편집기, 제어기, 블랙보오드(Blackboard), 어간 처리, 조사·접미사 처리, 어미 처리, 초기화 Knowledge Source(KS)와 확률 탐색 함수등으로 이루어져 있다.

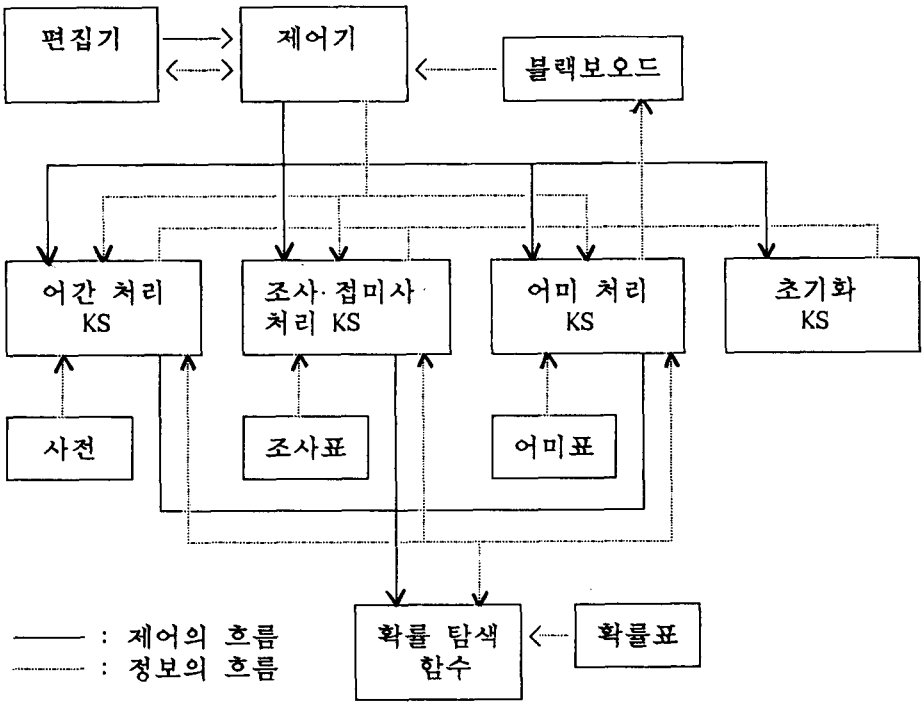


그림 2 지능형 한글 편집기의 구조

3.1.1 편집기의 역할과 동작

편집기는 일반적인 편집기와 같은 기능을 가지고 한글 오토마타에 의해 동작한다. 편집기는 다음과 같은 경우에 제어기를 호출하여 제어기가 되돌려 준 값으로 종성의 가능성을 판별하며, 판별할 수 없으면 일반 편집기와 같이 동작한다.

① Space 혹은 Return 키의 입력으로 하나의 어절이 생성될 때
⇒ 블랙보오드 혹은 어간처리, 조사·접미사, 어미 처리 Knowledge Source등이 이용하는 여러 변수들을 초기화하기 위해
② 중성 다음에 입력 된 자음이 종성이 될 가능성을 예측할 때
⇒ 종성의 가능성 여부를 판별하기 위해
③ 종성을 가지지 않는 글자가 생성될 때
⇒ 하나의 형태소가 완성되었는가를 알기 위해

3.1.2 제어기의 역할과 동작

제어기는 현재 입력된 글자와 블랙보오드의 어절 정보를 이용하여 어간 처리, 조사·접미사 처리, 어미 처리, 초기화 Knowledge Source중 하나를 선택한다. 각 Knowledge Source가 결정한 종성가능성을 편집기에 되돌려 준다. 각 Knowledge Source를 선택하는 조건은 다음과 같다.

① 어간 처리 Knowledge Source
제어기에 의해 초기값으로 설정되어 있어서 어절이 시작할 때 가장 먼저 호출된다.
② 조사·접미사 Knowledge Source
체언, 부사, 명사형 어미가 입력된 후에 호출된다.
③ 어미 처리 Knowledge Source
용언, 선어말 어미, 용언화 접미사, 서술격 접미사가 입력된 후 호출된다.
④ 초기화 Knowledge Source
Space 혹은 Return 키의 입력으로 하나의 어절이 생성된 후 호출된다.

3.1.3 블랙보오드

현재까지 입력된 어절의 정보를 가지고 있다. 예를 들면 "세상에는"이 입력되면 블랙보오드는 그림 3과 같다.

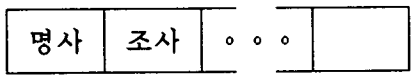


그림 3 블랙보오드에 실린 정보의 예

3.2 확률 탐색 함수

어간 처리, 조사·접미사 처리, 어미 처리 Knowledge Source가 확률 탐색 함수를 이용한다. 입력된 글자에서 종성이 될 가능성을 탐색하고 종성이 되는 경우, 종성이 되지 않는 경우, 판별할 수 없는 경우로 나누어 호출한 Knowledge Source로 그 결과를 되돌려 준다. 확률표는 하나의 Slot이 하나의 Bucket으로 이루어져 있고, 각 Slot은 표제어 부분과 확률 부분으로 이루어져 있다. 오버플로우를 해결하기 위해 선형탐색법(Linear probing)을 이용하고[2], 해쉬함수는 균일 해쉬함수의 나눗셈법을 변형한 것으로 다음과 같다[14].

Key = (초성 + 중성 × 19 + 종성 × 19 × 21) mod PRIME

3.3 어간 처리 Knowledge Source

어절에서 어간 부분의 종성을 판별하기 위해 제어기에 의해 호출되고 먼저 확률표를 이용하여 종성이 될 가능성을 알아낸다. 확률로 판별할 수 없으면 사전의 지식을 이용하여 종성이 될 가능성을 알아낸다. 현재까지 입력된 글자들이 하나의 단어를 이루면 사전에 존재하는 그 단어의 정보를 시스템의 어절정보에 기록한다. 어간 처리 Knowledge Source의 동작은 그림 4의 흐름도와 같이 동작한다.

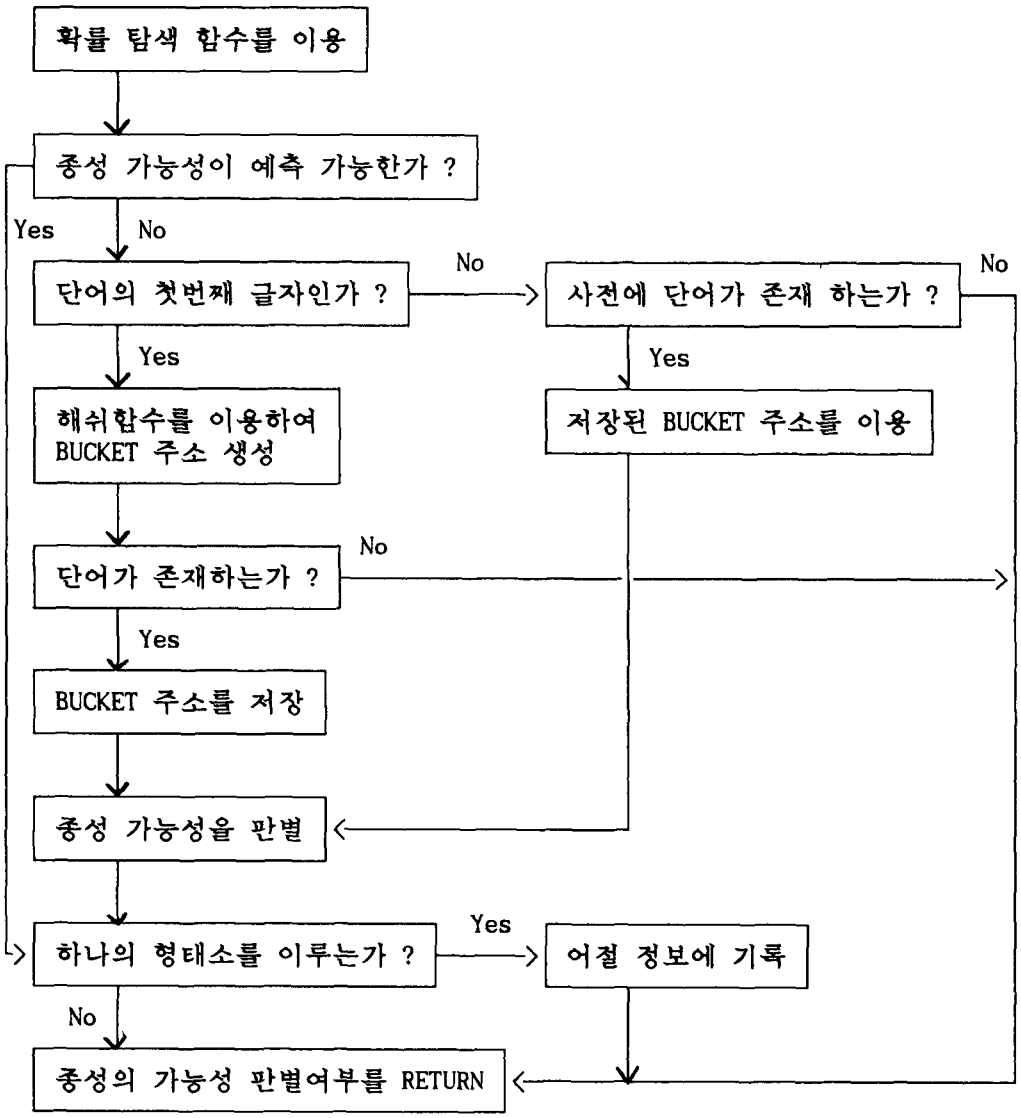


그림 4 어간 처리 Knowledge Source의 흐름도

3.3.1 사전의 구현

1) 표제어의 선정

편집기의 구현을 위한 사전의 크기와 규칙의 수와의 관계를 살펴보면 사전이 작으면 많은 규칙을 필요로 하고 사전이 크면 필요로 하는 규칙의 수가 적어진

다. 편집기는 실시간에 동작하므로 가급적이면 규칙의 수를 줄이고 실생활에 사용되는 단어들만 사전의 표제어로 삼아서 실시간 처리의 필수 조건인 적은 수의 규칙과 작은 크기의 사전을 구현할 수 있다. 표제어의 선정조건은 다음과 같다.

① 일어 일표제어 방식을 택하며 복합어(합성어, 파생어)도 표제어로 한다.
② 준말은 사전에 그대로 싣는다.
③ 접사와 결합한 단어와 접사와 결합하지 않은 단어 모두를 표제어로 함으로써 규칙을 간소화 한다.
④ 두개의 용언이 어울려 한 개의 용언으로 쓰이는 것을 표제어로 한다
⑤ 가변어는 어간만 저장한다.
⑥ 어간이 변하는 불규칙 용언은 변화가 발생한 경우 기본형을 알기 위하여 많은 규칙을 필요로 한다. 그러므로 기본형과 대표적인 변화형을 함께 실어서 예측 확률을 높이고 규칙의 수를 줄인다.
⑦ 불구동사는 변화형을 그대로 사전의 표제어로 실어서 규칙을 간소화 한다.

2) 사전의 형태와 탐색 방법

사전은 확률표와 같은 해쉬함수를 이용하여 탐색하고 동일한 방법으로 오버플로우를 해결한다. 그러나 사전은 한 개의 Bucket이 4개의 Slot으로 이루어져 있다. 사전의 Slot은 표제어, 품사, 형태 부분으로 이루어져 있다. 여기서 품사 부분에는 2.2절에서 분류한 품사중의 하나가 저장되고 형태 부분에는 불변어, 가변어 혹은 가변어에서 불규칙 변화를 하는 경우 불규칙 정보가 저장된다. 한 개의 Bucket에는단어의 첫번째 글자에서의 종성 예측을 쉽게 하고 기억 장치의 낭비를 줄이기 위해 종성을 가지는 글자와 종성을 가지지 않는 경우 두번째 글자의 초성이 종성을 가지는 첫번째 글자의 종성과 같은 경우 동일한 Bucket에 두었다. 예를 들면 "가게"와 "각선미"는 동일한 Bucket에 저장된다.

3.3.2 사용규칙

1) 사전을 이용하여 종성을 예측하는 경우의 규칙

```
If 표제어에서 종성으로 된 경우만 존재
Then 종성으로 예측
Else if 표제어에서 다음자의 초성으로 된 경우만 존재
Then 초성으로 예측
Else
     예측 불능.
```

2) 표제어가 마지막 글자인 경우 적용되는 규칙

```
If (품사가 용언) and (표제어의 마지막 글자가 종성이 없음)
   and ("ㄴ","ㄹ","ㅂ","ㅆ" 이 입력)
Then 종성으로 두고 품사와 형태정보를 기록
Else if (품사가 용언) and (표제어의 마지막 글자가 종성이 없음)
        and ("ㅁ" 이 입력)
Then 종성으로 두고 품사와 명사형 어미라는 형태정보를 기록
Else if 현재까지 입력된 글자가 표제어와 일치
Then 품사와 형태정보를 기록
```

3.4 조사·접미사 처리 Knowledge Source

어절에서 조사 부분의 종성을 판별하기 위해 제어기에 의해 호출된다. 어미가 활용하여 어절의 뒷 부분의 생성에 영향을 미치는 서술격 조사와 용언화 접미사는 규칙을 가지고 종성을 예측한다. 조사표를 사용하는 것과 어간처리 Knowledge Source와는 다른 규칙을 사용하는 것을 제외하고는 어간 처리 Knowledge Source와 동일하게 동작한다.

3.4.1 조사표

조사는 조사 상호간의 결합이 가능하므로 빈번하게 사용되는 결합형태는 조사표에 저장하여 이용한다. 조사표는 확률표와 같은 해쉬함수를 이용하여 탐색하고 동일한 방법으로 오버플로우를 해결한다. 첫번째 탐색에서 같은 글자로 시작되는 모든 조사의 주소를 탐색하여 저장한다.

3.4.2 규칙

```
If ("-하","-되","-이"가 입력) AND (종성으로 "ㄴ""ㄹ""ㅂ"이 입력)
Then 종성으로 두고 어절에 정보를 기록
Else if ("-하","-되"가 입력) AND (종성으로 "ㅁ"이 입력)
Then 종성으로 두고 "명사형 어미"라는 정보를 어절에 정보를 기록
Else {
  If 표제어에서 종성으로 된 경우만 존재
  Then 종성으로 예측
  Else if 표제어에서 다음자의 초성으로 된 경우만 존재
  Then 초성으로 예측
  Else
       예측 불능
}
```

3.5 어미 처리 Knowledge Source

어절에서 어미 부분의 종성을 판별하기 위해 제어기에 의해 호출되고 어미표를 이용하여 어미 부분의 종성을 예측한다. 어미표를 사용하는 것과 어간처리 Knowledge Source와는 다른 규칙을 사용하는 것을 제외하고는 어간 처리 Knowledge Source와 동일하게 동작한다.

3.5.1 어미표

어미표도 확률표와 같은 해쉬함수를 이용하여 탐색하고 동일한 방법으로 오버플로우를 해결한다. 첫번째 탐색에서 같은 글자로 시작되는 모든 어미의 주소를 탐색하여 저장한다.

3.5.2 규칙

```
If X 불규칙 활용
Then X가 어미에 나타남
Else {
  If 표제어에서 종성으로 된 경우만 존재
```

```
  Then 종성으로 예측
  Else if 표제어에서 다음자의 초성으로 된 경우만 존재
  Then 초성으로 예측
  Else
       예측 불능
}
```

3.6 초기화 Knowledge Source

Space 혹은 Return 키의 입력으로 하나의 어절이 생성된 후 어절 정보와 어간 처리, 조사·접미사, 어미 처리 Knowledge Source등이 이용하는 여러 변수들을 초기화한다.

Ⅳ. 성능 평가

평이하고 보편적인 단어와 문장을 선택하기 위하여 중학교 2학년 국어 교과서에서 단어를 추출하여 사전을 구성하고 문장을 선택하였다.

4.1 성능 평가 방법

"세상에서"를 고려할 때 입력된 글자의 수는 4자이고 그 중 종성 가능 여부를 측정한 것은 "셋", "상", "엣" 3자이며 이 중에서 확률로 예측한 것은 1자이고 ("엣"의 경우) 지식을 이용한 경우는 2자("셋","상"의 경우)이다. 그러므로 확률로 예측한 글자는 전체의 1/3(33.3%)이고, 지식을 이용한 글자는 2/3(66.7%)이며, 종성을 예측할 수 있는 능력은 3/3(100%)이다.

4.2 평가의 결과

전체 1415자 중에서 확률이 1인 226자와 0인 769자를 가지는 확률표, 205개의 단어를 가지는 사전, 88개의 조사와 접미사를 가지는 조사표, 56개의 선어말어미와 어미를 가지는 어미표를 성능 평가에 사용했다. 총 713자에 대해서 성능 평가를 한 결과는 그림 5와 같다. 그림 5에서 보는 바와 같이 확률에 의하여 예측한 경우는 21.6%, 지식을 이용한 경우는 67.6%로 올바르게 예측할 수 있는 능력은 88.9%였으며, 판별이 불가능하여 종성으로 둔 경우는 11.1%였다.

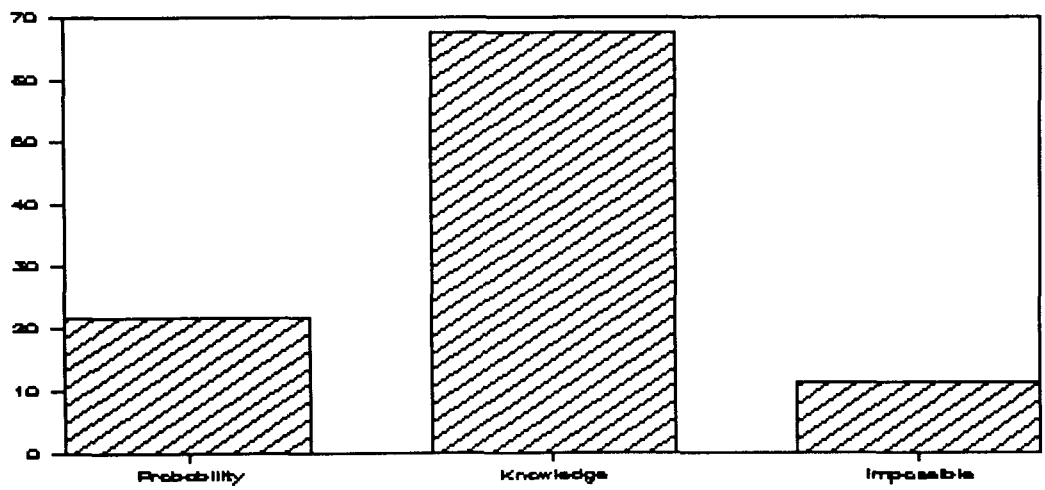


그림 5 예측 결과

V. 결론

그림 5 에서 보는 바와 같이 확률과 문법 지식을 이용하여 종성을 예측할 수 있음을 보았다. 그러나 지능형 한글 편집기는 몇가지 문제점을 가지고 있다. 문제점중의 하나는 사전에 단어가 없으면 어간 부분도 사전을 이용하여 예측이 불가능하지만 어절의 뒷 부분도 예측이 불가능하다는 것이다. 예를 들면 "세상에서"를 입력할 때 "세상"이라는 단어가 사전에 없으면 제어기는 "세상" 다음의 글자들이 조사인지 알 수 없으므로 조사표의 지식을 이용하지 못한다. 다른 하나는 사전에 그 단어가 있어도 예측을 못하는 경우로서 대체적으로 예측이 불가능한 경우는 첫 글자의 경우인데 사전에 "가게", "각각"이 있을 때 첫 글자의 경우 "가" 다음에 "ㄱ"이 입력되면 확률로도 예측이 불가능하고 사전에도 종성이 된 경우와 초성이 된 두가지 경우가 모두 존재하므로 예측이 불가능하다. 그러므로 보다 높은 수준의 예측을 위하여 현재는 어절의 정보만 이용하고 있지만 문장 구조의 정보도 함께 이용하는 방안과 문법 지식뿐만아니라 의미 지식을 이용하는 방안을 연구중이다. 그리고 지능형 한글 편집기는 어절의 입력과 동시에 형태소 분석이 이루어지므로 형태소 분석의 결과를 이용하여 입력중에 띄어쓰기를 검사하고, 잘못된 맞춤법의 단어가 입력되면 올바른 철자법의 단어를 입력하도록 사용자를 유도하는 기능을 가지는 한글 편집기로 개발할 예정이다.

참고문헌

[1] John J. Darragh, Ian H. Witten, and Mark L. James. "The Reactive Keyboard : A Predictive Typing", Computer , Nov . 1990, pp. 41-49

[2] E. Horowitz and S. Sahni, Fundamental of data Structure in Pascal, Computer Science Press Inc., 1987

[3] Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, D. Raj Reddy, "The Hearsy-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty", Computing Surveys, Vol. 12, No. 2, June 1980

[4] Robert Engelmore, Tony Morgan, Blackboard Systems, Addision Wesley, 1988

[5] 김영웅, 한글 철자법 교정 시스템, 한국과학기술원 석사학위논문, 1984

[6] 손우형, 한국어 기계번역을 위한 형태소 분석에 관한 연구, 한국과학기술원 석사학위논문, 1986

[7] 강재우, 접속정보를 이용한 한글 철자 및 띄어쓰기 검사기의 설계 및 구현, 한국과학기술원 석사학위논문, 1990

[8] 남기심, 고영근, "표준 국어 문법론", 탑출판사, 1991

[9] 미승우, 새 맞춤법과 교정의 실제, 어문각, 1988

[10] 장유미, 최윤철, "한글 텍스트의 띄어쓰기를 위한 소프트웨어 개발에 관한 연구", 1989년도 한국정보과학회 가을 학술발표논문집, 제16권, 제2호, pp 457-460, 1989

[11] 송춘환, 강재우, 김연배, 최기선, 권용래, 김길창, "한글 철자 및 띄어쓰기 검사기", 1989년도 한국정보과학회 가을 학술발표논문집, 제16권, 제2호, pp 595-598, 1989

[12] 채영숙, 김재원, 권혁철, "도움말 기능을 가진 문서 철자 검색/교정 시스템", 1990년도 한국정보과학회 가을 학술발표논문집, 제17권, 제2호, pp 815-818, 1990

[13] 조영환, 김덕봉, 최기선, 김길창, "한글 맞춤법 오류 및 교정 시스템", 1990년도 한국정보과학회 가을 학술발표논문집, 제17권, 제2호, pp 823-826, 1990

[14] 한부형, 조유근, "한국어 정보 검색 시스템에서의 한글 해싱 기법에 관한 연구", 1991년도 한국정보과학회 가을 학술발표논문집, 제18권, 제2호, pp 837-840, 1991
[15] 이규은, 김황수, "지능형 한글 편집기 개발에 관한 연구", 1991년도 한국정보과학회 가을 학술발표논문집, 제18권, 제2호, pp 845-848, 1991
[16] 금성출판사 편집부, 국어사전, 금성 교과서, 1990