통계적 기법에 의한 한-영 문자열의 자동 전환

안영훈[0], 강승식
국민대학교 컴퓨터학부
{xfilecom, sskang}@cs.kookmin.ac.kr

Statistical Approach to the Automatic Korean-English String Conversion

Young-Hoon Ahn, Seung-Shik Kang
School of Computer Science, Kookmin University

요 약

한글 혹은 영어 문자열을 입력할 때 입력 모드를 수동으로 전환하지 않더라도 입력된 문자열이 한글인지, 영어인지를 자동으로 판단하여 해당 문자열로 변환하는 방법을 제안한다. 한글 문자열일 확률을 계산하기 위해 음절 구성 요건과 음절 빈도 정보를 이용하고, 영어 문자열일 확률을 계산하기 위해 영어 bigram 및 trigram 정보를 이용한다. 또한, 한글과 영어가 혼합된 문자열은 한글일 확률과 영어일 확률이 교차되는 경계 위치를 인식함으로써 혼합 문자열을 생성한다.

1. 서론

한글과 영문이 혼합된 문서를 입력할 때 한-영 전환 키(한-영 키 또는 Shift+Space 키)를 이용하여 입력 모드를 한글 모드 또는 영문 모드로 전환하게 된다. 그런데 한글과 영어가 혼합된 문서를 작성할 때 매번 입력 모드를 전환하는 것은 매우 불편하다. 또한, 잘못된 입력 모드에서 데이터를 입력했을 경우에 데이터를 삭제하고 다시 입력해야 하는 번거로움이 있다. 이처럼 잘못된 입력 모드에서 문자열을 입력하는 근본적인 원인은 사용자가 현재의 입력 모드를 잘못 알고 있거나 인지하지 못하기 때문이다.

이러한 불편함을 해결하는 방법의 하나는 사용자가 현재 입력 모드를 쉽게 인지할 수 있도록 "*입력 모드에 따라 마우스 모양을 다르게 하는 방법*"이 있다. 즉, 사용자가 입력 모드를 전환하면 영문 모드와 한글 모드의 마우스 모양을 다르게 보여주는 방법이다. 이 방법은 마우스를 텍스트 창으로 이동하여 문자열을 입력할 때 유용할 것으로 판단된다. 그런데 윈도98 이후의 윈도 운영체제는 입력 모드가 응용 프로그램마다 개별적으로 유지되기 때문에 모든 응용 프로그램에 적용하려면 운영체제 차원에서 이 기능을 지원해 줘야 하는 문제점이 있다.

한글 문서편집기에서는 한-영 자동 전환으로 인한 사용자의 불편함을 해소하기 위해 입력 모드를 자동으로 전환해 주는 기능을 사용하고 있다[1]. 이러한 자동 전환 기능은 입력 모드 오류에 의한 문자열이 한글인지, 영어인지를 자동으로 판별할 수 있다는 특성을 이용한 것이다. 입력 모드에 따른 문자열 입력 오류는 "한글 모드에서 영문을 입력"하는 오류('한영 오류')와 "영문 모드에서 한글을 입력"하는 오류('영한 오류')가 있다. 아래 예는 "automatic conversion system"과 "한영 자동 전환 시스템"을 잘못된 입력 모드에서 입력한 예이다.[1)]

며샤ㅐㅡ며셔	채ㅜㅍㄷㄱ냐ㅐㅜ	뇟ㅅ드		(한영오류)
automatic	conversion	system		
gksdud	wkehd	wjsghks	tltmxpa	(영한오류)
한영	자동	전환	시스템	

위 예와 같이 영문 모드에서 한글을 입력하거나 한글 모드에서 영문을 입력했을 때 일반적으로 전혀 엉뚱한 문자열로 표시된다. 따라서 문자열의 특성을 분석함으로써 입력 모드 오류를 판단할 수 있으며, 입력 오류로 판단된 경우에 자동으로 해당 문자열로 변환해 준다. 특히, 영어 문자열을 한글 모드에서 입력한 경우는 한글의 음절 구성 요건을 검사하는 단순한 방법만으로 대부분의 입력 오류가 발견된다[2].

본 논문에서는 입력된 문자열을 영어 문자열과 한글 문자열일 두 가지 가능성에 대해 확률을 계산하여 한글 문자열인지, 영어 문자열인지를 판단하고 해당 문자열로 전환해 주는 방법을 제안한다. 또한, "system은"과 같은 한-영 혼합 문자열은 한글일 확률과 영어일 확률이 교차되는 문자의 위치를 인식하여 혼합 문자열을 생성한다.

2. 관련 연구

유승목(1995)은 한글 모드에서 입력되는 영어 문자열을 한글로 전환해 주는 방법을 제안하였는데, 영어 문자열인지를 판단하기 위하여 한글의 음절 구성 오토마타를 이용하였다. 이 방법은 한글 문자열이라고 가정했을 때 음절 구성이 안 되는 것을 영어로 간주한다. 영어의 경우 첫 3문자만으로 89.1%의 정확도로 한글이 아니라는

1) 입력 모드 오류에 의한 오류 문자열은 컴퓨터 자판의 배열에 따라 달라지지만 자판 배열에 의한 편차는 크지 않을 것으로 추정된다.

사실을 판단할 수 있다. 이 방법에서는 down(애주), work(재가) 등과 같이 한글의 음절 구성 요건을 만족하는 중의적인 문자열을 처리하기 어려운 문제점이 있다.

이긍해(2000)는 한-영 자동 전환 방법으로 어절 구분자가 입력된 후에 자동 전환을 하는 어절 단위 방법(word- phrase mode algorithm)과 어절이 입력되는 중에 실시간으로 자동 전환을 수행하는 예측 방법(predictive mode algorithm)을 제안하였다. 이 방법에서는 "PC와 TV는"처럼 한영 혼합 문자열에서 조사 인식 문제를 해결하기 위해 어절 끝부분의 형식 형태소(조사, 접미사 등)를 인식하는 방법을 이용한다. 즉, 실질 형태소와 형식 형태소를 구분하고 실질 형태소에 대해 한글인지 영어인지를 판단한다.

3. 한글 및 영어 빈도 계산

임의의 문자열이 한글 문자열일 확률을 계산하기 위하여 한글 문자열의 음절 구성 요건 및 음절 빈도 정보를 이용한다. 또한, 영어 문자열일 확률을 계산하기 위하여 영어 bigram 및 trigram 통계 정보를 이용한다. 입력 문자열이 한글인지 영문인지를 판단하기 위해 한글 음절 빈도와 영문 bigram 및 trigram 빈도를 조사하였다.

3.1 한글 음절 빈도

한글 음절 빈도수를 추출하기 위해 1200만 어절 규모의 말뭉치를 수집하였으며 말뭉치는 표 1과 같이 구성되어 있다.

표 1. 말뭉치의 구성

말뭉치 유형	어절수
신문기사	540만 어절
Krist Collection	370만 어절
KTSET	80만 어절
기 타	210만 어절
합 계	1200만 어절

표 1의 말뭉치에서 한글 음절 빈도를 조사하여 각 빈도수를 최다 출현 빈도로 나누어 음절 확률 빈도를 계산하였다. 최다 출현 빈도 음절을 S_{max}라 할 때 음절 S_i의 출현 확률 $P_K(S_i)$는 식(1)과 같다.[2)]

$$P_K(S_i) = freq(S_i) / freq(S_{max}) \quad \text{식(1)}$$

3.2 영문 bigram과 trigram 정보

영문 bigram과 trigram 음절과 그 빈도수를 추출하기 위하여 Unix 영어 사전 68,000여 개의 단어에 대해 bigram과 trigram 빈도를 추출하였다. 이 때 bigram 및 trigram 빈도는 위치에 따라 <시작 위치, 중간 위치, 끝 위치>로 구분하여 빈도를 추출하였다. 예를 들어, 'study', 'student', 'enter'에 대한 3가지 bigram 빈도는 아래와 같다.

예) study, student, enter

- 단어 시작 위치 : {st}, {st}, {en}
- 단어 중간 위치 : {tu, ud}, {tu, ud, de, en}, {nt, te}
- 단어 끝 위치 : {dy}, {nt}, {er}

- 'st'의 빈도 : <2, 0, 0>
- 'tu'의 빈도 : <0, 2, 0>
- 'nt'의 빈도 : <0, 1, 1>

위치에 따른 bigram 혹은 trigram 확률은 식(2)와 같이 해당 빈도수를 최대 빈도수로 나눈 값으로 계산한다.

$$P_E(T_i) = freq(T_i) / freq(T_{max}) \quad \text{식(2)}$$

4. 한영 자동 전환 알고리즘

입력 모드에 따른 문자열 입력 오류를 자동으로 교정하기 위한 방법으로는 (1) 입력 문자열에 대해 한글일 확률, 영어일 확률을 계산하여 가능성이 높은 쪽으로 판단하는 방법과, (2) 기본적으로 한글 문자열로 간주하고 한글일 확률이 낮고 영어일 확률이 높은 문자열만 영어로 변환하는 방법이 있다.

한글 음절의 특성인 자음과 모음을 초성, 중성, 종성의 3개 항목으로 구성하여 하나의 음절을 구성하는 것과 자주 사용되는 음절의 빈도와 자주 사용되지 않는 음절 빈도를 조사한 정보를 이용하여 입력 음절이 한글일 가능성을 계산하고, 영어 사전에 수록된 68,000여 개의 단어들에서 얻은 bigram 및 trigram을 이용하여 영문일 가능성을 판단한다.

4.1 자료구조

자동 전환을 위한 자료구조로 engstr, korstr, ekmap, ekmark 4개의 배열이 사용된다.

```
char engstr[MAXCHAR];  // 영어 문자열
char korstr[MAXSYLL];  // 한글 문자열
int ekmap[MAXCHAR];
    // engstr[]과 korstr[]의 대응관계
char ekmark[MAXCHAR];
    // 입력 문자들에 대한 한-영 mark 표시
```

engstr과 korstr은 입력된 문자열에 대해 각각 영어 문자열과 이에 대응하는 한글 문자열을 저장한다. engstr은 입력 문자열에 대한 영어 문자열이 저장된다. 영문 모드에서 입력한 경우는 입력 문자열을 저장하고, 한글 모드에서 입력한 경우는 입력 문자열을 이에 대응하는 영어 문자열로 변환하여 저장한다.

ekmap은 engstr과 korstr의 대응 관계를 입력된 문자

2) 빈도가 높은 음절은 순서대로 '이/다/의/는/에/을/하/한/고/가/로/기'이고, 가장 빈도가 높은 음절이 '이'이므로 S_{max}는 '이'이다.

단위로 저장한다. ekmap[i]는 영문자 engstr[i]에 대응하는 한글 음절의 위치로서 korstr[]의 인덱스값이다. 예를 들어, 영어 문자열 "study"에 대응되는 한글 문자열"ㄴ셔요"에 대해 ekmap의 값은 0, 1, 1, 2, 2이다.[3] 즉, ekmap은 engstr의 substring에 해당하는 한글 음절의 위치정보를 저장한다.

ekmark는 입력된 각 문자들이 영문인지 한글인지를 나타내는 배열로서 자동 전환 알고리즘에 의해 입력된 키 단위로 영문 또는 한글 표시를 하기 위한 목적으로 사용된다. ekmark[i]는 i번째 입력된 키에 대한 한-영 표시로서 아래 4가지 값 중 하나를 갖는다.

MARK_K : 한글 음절의 일부로 판정된 경우
MARK_k : 한글 음절의 일부로 추정된 경우
MARK_E : 영문자로 판정된 경우
MARK_e : 영문자로 추정된 경우

4.2 초기화

입력 문자열에 대해 engstr, korstr 및 ekmap을 초기화한다. 입력 모드가 '한글'이면 korstr에 입력 문자열을 저장하고 engstr에는 korstr에 대응하는 영어 문자열을 저장한다. 입력 모드가 '영문'일 때는 engstr에 입력 문자열을 저장하고 korstr에는 engstr에 대응하는 한글문자열을 저장한다. 이와 같이 대응 문자열을 생성하는 과정에서 한글 음절 korstr[i]에 대해 대응하는 영어 문자열의 시작 위치가 engstr[j]라면 ekmap[j] = i를 저장한다.

대응 문자열이 생성되면 ekmark에 입력 문자가 한글 혹은 영문일 가능성을 추정하여 ekmark[]를 MARK_k (한글로 추정) 또는 MARK_e(영어로 추정)로 초기화한다. 한글로 추정되는 경우는 "음절 구성 요건을 만족하고 음절 빈도가 1 이상인 경우"이며, 그렇지 않은 것은 영문으로 추정한다.

```
void initialize() {
  int i, j=0;

  if (입력 모드 == 한글)
     Korstr ← 입력 문자열;
  else engstr ← 입력 문자열;
  set_ekmap(korstr, engstr, ekmap);
     // 한영 대응 문자열 초기화 및
     // ekmap[]에 대응 index 저장
  while (i < strlen(engstr)) {
     i = get_index_syl_start(j);
     j = get_index_syl_end(i);
     syl = get_one_syl(korstr, i, j);
     if (freq(syl) > 0)
  ekmark[i~j] = MARK_k;
     else ekmark[i~j] = MARK_e;
  }
}
```

알고리즘 1. 초기화 알고리즘

4.3 자동 전환 알고리즘

초기화 단계에서 적용된 ekmark[]의 내용에 따라 입력 문자열의 유형을 아래와 같이 3가지로 구분한다.

유형-1. 모두 영어 문자열로 추정된 경우
유형-2. 모두 한글 문자열로 추정된 경우
유형-3. 한글과 영문이 혼합된 경우

4.3.1 모두 영어로 추정된 경우

입력 문자들이 모두 영어로 추정된 경우는 입력 문자열이 음절 구성이 되지 않거나 음절 구성이 되더라도 빈도가 매우 낮으므로 한글 문자열일 가능성이 매우 낮다. 따라서 영어 문자열로 판정한다.

4.3.2 모두 한글로 추정된 경우

입력 문자들이 모두 한글로 추정된 경우는 입력 문자열이 모두 한글 음절로 구성될 수 있는 문자열이다. 그런데 이 경우에 'and는'과 같은 영어+조사 유형이 '뭉는'이라는 한글로 추정되는 오류가 발생한다. 따라서 영어일 확률이 높은 문자열을 찾아서 해당 문자열을 영문으로 추정한 후에 유형-3의 "한글 및 영문이 혼합된 경우"의 처리 방법을 적용한다.
영문으로 추정되는 문자열을 찾는 방법은 다음과 같다. 한글 음절의 누적 확률 PK와 영문의 누적 확률 PE를 계산하여 영문 누적 확률이 한글 누적 확률보다 커지는 위치를 영문 문자열의 끝 위치로 한다.

한글 i번째 음절부터 m번째 음절까지 한글 음절의 누적 확률은 식(3)과 같이 계산한다.

$$PK = \sum_{i}^{m} \ln P_K(S_i) \qquad \text{식(3)}$$

한글 i번째 음절 위치부터 m번째 음절에 해당하는 문자열까지 영문 tigram 개수가 n개일 때 영문 trigram의 누적 확률[4]은 식(4)와 같다.

$$PE = \sum_{j} \ln P_E(T_j) \qquad \text{식(4)}$$

PE와 PK를 계산할 때 bigram_trigram 개수와 음절수가 다를 수 있으므로 PE값을 아래와 같이 수정한다.

$$PE = PE \times \text{음절수} \;/\; (\text{bigram\_trigram 개수})$$

PK 값과 PE 값을 비교하여 "PE가 PK보다 큰 경우"와 "현재 음절의 확률이 매우 낮은 경우"가 발생하는 문자를 영문 문자열의 끝 위치로 한다. 영문 문자열의 시작 위치는 초기에 입력 문자열의 첫 문자 위치로 하고, 첫 문자부터 한글 2 음절 이상인 문자까지의 한글일 확률이 높고 영문일 확률이 매우 낮은 경우는 PE와 PK

3) korstr[i]에는 한글 음절이 저장된다고 가정한다. 한글 음절이 2 바이트를 차지한다고 가정할 때는 ekmap의 값이 0, 2, 2, 4, 4 이다.

4) 영문자 개수가 2개이면 bigram 확률을 사용한다.

의 계산 시작 문자 위치를 다음 문자 위치로 한다.

이러한 방법으로 영어 문자열의 시작 및 끝 위치를 찾아 해당 문자열 부분의 ekmark를 모두 MARK_E로 수정한다. 이 과정에 의해 한글 문자열로 추정된 것은 "영어 문자열" 또는 "한글 및 영문이 혼합된 문자열"로 수정될 수 있다. 영어 문자열이 발견되지 않은 경우는 한글로 판정한다.

4.3.3 한글과 영문이 혼합된 경우

입력 문자열에서 '영어'+'한글' 유형인 부분에 대해서만 '한글'로 추정된 부분을 영어로 수정해 줄 것인지를 판단하여 수정한다. '한글' 부분을 영어로 수정할 것인지를 판단하기 위한 방법은 다음과 같다. '영어'의 마지막 문자 위치부터 시작하여 식(3), 식(4)의 영어 문자열 확률과 한글 문자열 확률을 계산하여 4.3.2와 동일한 방법으로 영문 끝 문자 위치를 설정하여 그 위치가 달라지면 영어 문자열 부분을 확장한다.

4.4 후처리

한영 자동 전환을 위하여 위와 같은 과정을 수행한 후 후처리 과정을 적용한다. 후처리 단계는 빈도 정보를 이용하여 확률로 처리할 때 나타나는 오류를 보정하기 위해서 예외 사전을 도입하여 그에 해당하는 단어나 substring이 나타나면 사전에 정의된 내용으로 변환하는 과정이다.

자동 전환 과정에서 발생하는 오류를 수정하기 위하여 각 오류 유형에 따라 오류어 사전을 구성한다. 이 사전은 특정 substring 또는 어절 자체에 대해 해당 문자열이 항상 영문인 경우, 항상 한글인 경우, 그리고 입력 모드에 의해 결정되는 정보로 구성된다.

<word/substring, 영문/한글/IME>

입력 문자열에 대해 오류어 사전을 검색하여 해당 문자열이 발견된 경우는 해당 문자열 부분을 사전에 기술된 정보에 따라 한글 혹은 영문으로 ekmark를 변경한다. 본 논문에서 사용된 오류어 사전은 어절 자체가 수록된 것이 194개, substring이 805개로 구성되어 있다.

5. 실험 결과

본 논문에서 시스템을 개발하는데 사용된 데이터에서는 99.99%의 높은 성능을 보이고 있다. 시스템의 정확도를 평가하기 위해 네 가지 형태의 실험 데이터를 구축하였으며 표2와 같다. 표2에서 '영어문서'는 Brown Corpus에서 태그를 제거한 원시 말뭉치이고, '한글 문서'는 신문기사에서 한글 어절을 추출한 것이다. '일반 문서'는 초등학교 교과서에서 문장 부호를 제거한 것이고, '영한 혼합 문서'는 대용량 말뭉치에서 '영어' + '한글' 유형만 추출한 것이다.[5)]

5) '한글' + '영어' 유형의 어절은 실험 대상에서 제외하였다. 그 이유는 현재까지 구현된 알고리즘에서 한글 뒤에 영문이 오는 경우 첫문자부터 영어로 판단된 위치까지 모두 영문으로 처리했기 때문이다.

표 2. 자동 전환 실험결과

문서명	실험 데이터	어절수	정확도
영어 문서	Brown Corpus	1,012,930	99.40%
한글 문서	신문기사	98,764	99.93%
일반 문서	초등학교 교과서	6,274	98.86%
영한 혼합 문서	영한 혼합 어절	14,135	98.46%

영한 혼합 문서에서 발생한 오류 유형을 검토한 결과로 cm('츠')로 인한 오류가 31개, rk('가') 오류 44개, ek('다') 오류 2개, go('해') 오류 11개로서 총 오류어 218개 중 80개를 차지하고 있다. 이처럼 자주 발생하는 오류를 처리한다면 정확도는 99% 이상으로 향상될 수 있을 것으로 추정된다. 또한, 일반 문서(초등학교 교과서)에서도 이와 유사한 오류가 빈번하게 발생함으로 인하여 정확도가 저하됨을 알 수 있었다.

6. 결론

한글 음절 빈도와 영어 bigram 및 trigram 빈도를 이용하는 한-영 자동 전환 방법을 제안하였다. 한글 및 영어 문서에 대해 실험한 결과, 순수한 영어 단어에 대한 정확도가 99.4%, 순수 한글에 대한 정확도가 99.93%였다. 'abc는'과 같이 영어 뒤에 한글이 오는 경우는 98.46%로 정확도가 낮아졌으나, 한글 및 영어 빈도가 모두 높은 소수의 특정 문자열로 인해 빈번하게 발생하는 오류이므로 이러한 유형의 오류를 처리한다면 정확도가 향상될 것으로 기대된다.

본 논문에서는 한글 혹은 영어를 판정하는 문제를 중심으로 알고리즘을 구현하면서 '한글' 뒤에 '영어'가 오는 경우는 드물기 때문에 이를 고려하지 않았다. 또한, 현재까지 구현된 시스템을 대화식으로 입력되는 문자열에 대해 실험할 경우에 한 어절 내에서 2번 이상 입력 모드가 전환되는 경우가 발생할 수 있다.

7. 참고 문헌

[1] 이긍해, "한영 자동 전환 시스템 AIMS의 성능 평가", 한국정보과학회 가을 학술발표논문집, Vol.22, No.2, pp.585-588, 1995.

[2] 유승목, 이정훈, 김삼묘, "한/영 키 모드 자동 전환기", 한국정보과학회 가을 학술발표논문집, Vol.22, No.2, pp.589-592, 1995.

[3] K. H. Lee, "Recognizing Korean Postpositions in Mixed-mode Strings for the Automatic Korean/English Input Mode Switching", 19th International Conference on Computer Processing of Oriental Languages(ICCPOL'2001), pp.339-344, 2001.