□ 기술해설 □

음성인식에서의 언어모델링

시스템공학연구소 최환진·박재득·박동인*

1. 서 론

음성인식을 통한 인간과 기계간의 의사 소통이라는 주제는 지난 수십년 동안에 음성연구 관련 공학자와 과학자들에게 있어서 커다란 꿈이자 목표였다[1]. 이러한 목표이자 꿈이 이제 서서히 구현되어가고 있다. 컴퓨터의 데스크톱 상에서 음성을 이용하여 응용 프로그램을 구동시키고, 대화형 음성인식 시스템을 이용하여 열차나 비행기의 티켓 등을 예매하는 등 음성인식이 이제는 우리 생활의 일부로써 활용되기 시작하고 있다[2, 3].

일반적으로 "음성인식(speech recognition)"이란 마이크나 전화기를 통해 화자에 의해서 발성된 음향적인 신호를 인간이 이해할 수 있는 단어나 구문들로써 표현하는 일련의 과정을 말하며, 최종적으로 인식된 단어나 구문 등을 컴퓨터나 기계 상의 명령이나 제어, 자료입력, 그리고 문서의 준비 등을 위한 용도로써 사용되게 된다. 음성인식은 해부학, 음성학, 음운학, 음향학, 전산학, 언어학, 그리고 전자 공학 등 다 분야의 지식들이 결합되어 이루어지는 분야로 어느 한 분야의 발전만으로는 음성인식의 궁극적인 발전을 도모하기는 어렵다. 현재의 음성인식 연구는 크게 이론적인 배경과 계산적인 분야의 발전으로 인해서 가능하게 되었다[4, 5].

일반적으로 음성인식을 위해서 "음향모델"과 "언어모델"[6, 7]이라는 두 개의 지식원이 사용된다. "음성인식기"란 이러한 지식원들을 이용하여 입력된 음성 신호적 특성을 나타내는 특징 파라미터와 음향 모델 및 언어 모델들을 비교하는 일종의 정합기로써의 역할을 하게 된다. "음향모델"이란 음성신호의 신호적인 특성을 모델링하는 것으로, 입력 신호와 직접적으로 정합을 이루게될 표준적인 패턴들에 해당되며, "언어모델"은 음향적인 신호 수준에서가 아닌 언어적인 단위, 즉 인식 어휘에 해당되는 단어나 음절 등에 해당되는 단위들간의 언어적인 순서 관계를 나타내게 된다. 특히, 이러한 언어 모델은 인식 어휘가 단어나 음절인 경우에는 무의미하나, 인식 대상이 문장이나 구, 절을 대상으로 할 경우 언어 모델의 필요성은 매우 커지게 된다.

언어라는 것이 언어를 구성하는 심벌들의 유한적인 조합의 형태로 표시되므로, 언어를 구성하는 심벌들간의 순서관계가 존재하게 되는데, 그러한 순서 관계를 음성인식에서 얻어진 단어 열들에 적용함으로써 음향적인 모호성으로 인해서 오식될 가능성이 있는 단어들을 언어 모델을 통해서 교정함으로써 음성인식의 오류를 줄일 수 있게 된다.

본 고에서는 음성인식을 위한 처리 과정중에서 음성인식의 오류를 줄이며, 구나 절, 그리고 문장 등의 음성인식을 수행하기 위해서 사용되는 언어 모델링들을 중심으로 기술하고자 한다.

본 고의 구성은 다음과 같다. 2장에서는 언어 모델의 필요성에 대해서 살펴보도록 하고, 3장에서는 음성언어를 위한 통계적인 모델들과 오토마타형식의 언어모델들에 대해서 기술하고자 한다. 마지막 4장에서는 결론을 기술하고자 한다.

* 종신회원

2. 음성인식과 언어모델

2.1 음성 언어 처리의 이해

음성처리를 위한 언어모델을 이해하기에 앞서서 "음성(speech)"과 "음성 언어(spoken language)"라는 용어에 대해서 살펴보기로 하자. "음성"이란 화자에 의해서 발성된 소리에 해당되는 것으로, 언어적인 정보와 화자의 개인성을 나타낼 수 있는 화자 정보가 혼합되어 있다. 반면, 음성 언어는 텍스트 언어와는 대별되는 것으로 텍스트 언어가 나름대로의 구문 규칙, 문법, 의미 처리 등을 필요로 하듯이, 음성 언어에도 그러한 요소들이 필요하다. 음성 언어는 음성을 하나의 독립적인 언어로 규정하고, 그러한 언어 현상을 규명하기 위한 환경들을 함께 정의한다고 볼 수 있다. 단순히 "음성 처리"라는 의미는 음성이라는 신호적 특성만을 중심적으로 다루는 반면, "음성 언어 처리"는 음성이 갖는 언어적인 특성들-즉, 음성의 물리적 특성, 음성인식, 음성합성, 음성변환, 음성 언어 모델 등-을 고찰하는 것을 그 대상으로 하고 있다. 이러한 측면에서 보면, 음성인식이란 음성 언어 처리의 한 분야로, 음성이라는 음향적인 신호로부터 언어적인 정보로의 변환을 위해서 신호 처리 및 음향 해독 기술, 이와 아울러 인식된 결과로부터 음성 언어의 구문이나 의미 구조 분석을 위한 언어 처리 기술이 사용되게 된다. 언어모델은 입력 신호에 대한 음향적 인식 어휘모델의 출력 확률과 그들의 순서 관계에 적용되는 언어모델의 출력 확률을 함께 최대화하는 인식 어휘의 열을 얻기 위해서 사용된다. 다음으로 언어모델이 갖는 중요성과 그 필요성에 대해서 살펴보고자 한다.

2.2 음성인식에 있어서의 언어모델의 필요성

음성인식에 있어서 언어모델은 앞서 음성 언어 처리과정에서 살펴 보았듯이, 음성의 음향적인 해독 결과로써 얻어진 음성인식을 위한 어휘 사전에 등록된 어휘의 열에 대한 언어적인 수준에서의 제약을 가하기 위해서 사용된다. 이러한 언어모델이 음성인식에 적용됨으로써 인식기가 얻는 장점들[8, 9]을 요약하면 다음과 같다.

① 다중 지식을 사용함으로써 인식 오류를 줄일 수 있다.

② 인식 어휘들간의 언어 수준에서의 제약은 문장 오류율를 줄인다.

첫번째로, 다중 지식을 활용함으로써 인식 오류를 줄인다는 점은 음성인식에 다양한 지식원들이 사용되며, 그러한 지식들의 결합이 인식 오류를 줄일 수 있다는 점에서 중요하다고 볼 수 있다. 음성인식은 크게 음향적인 지식과 언어적인 지식이 사용되는데, 이러한 지식들은 음성인식 과정에서 서로 다른 부분에서 적용됨으로써, 다른 지식원을 필요로 하는 부분에서의 오류를 줄이는데 사용된다는 점에서 유효하다고 할 수 있다. 음성 신호의 음향적인 특성을 표시하는 음향적인 지식을 모델화한 음향모델은 음성 신호의 신호적인 특성의 오류율을 줄이기 위해서 사용되며, 언어모델은 음성인식기에 의해서 인식된 인식 어휘의 열들에 대한 오류를 줄이므로써 음성인식시에 수반되는 다양한 오류들을 단계별로 줄여준다는 점에서 음성인식을 위한 다중 지식의 사용은 필요하다고 생각된다.

두번째로, 인식 어휘들간의 언어 수준에서의 제약은 문장 오류를 줄인다는 점에서 언어모델은 필요하다. 음성인식이 단순히 화자에 의해서 발성된 음성에 대한 음향적인 해독과정이라고 본다면, 음성인식의 궁극적인 목적인 음성인식을 통한 기계에 의한 자동적인 명령 수행 혹은 자료 입력 등을 위해서 음성인식된 내용으로부터 사용자가 원하는 내용을 처리할 수 있어야 한다. 이를 위해서는 문장 혹은 구나 절 형태의 음성인식을 수행해야 한다. 구나 절은 음성인식 대상 어휘들로 이루어진 열로써 이러한 어휘들간에는 순서적인 제약적인 존재하게 된다. 이러한 인식 어휘들 사이의 구문적인 위치적 제약 관계를 규정하기 위해서 언어모델이 사용되는데, 이러한 언어모델은 음성 신호의 음향적인 해독 결과에 오류를 보정하기 위해서 유효하다.

지금까지 언어모델이 음성인식에 있어서 유효한 점들에 대해서 살펴보았다. 다음 3장에서

는 구체적으로 음성인식에 적용되는 언어모델에 대해서 살펴보고자 한다.

3. 음성 언어 모델링

3.1 음성 언어 모델링이란?

소규모 어휘를 대상으로 한 음성인식 시스템들은 인식 대상 어휘가 음성인식 단위에 해당되며, 주로 명령 및 제어를 위한 응용 프로그램들을 위해서 사용된다. 이러한 시스템들의 경우, 일반적으로 언어 모델들에 크게 의존하지 않는다. 단어에 대응되는 어휘가 하나의 명령에 바로 사상되기 때문이다. 반면, 대어휘를 대상으로 한 음성인식의 경우는 입력 음성에 내재한 언어적인 지식에 크게 의존하게 된다. 따라서, 대어휘를 대상으로 음성인식을 할 경우, 언어 모델의 형태로 이러한 언어적인 지식을 결합하는 것이 대단히 중요하게 된다. 음성인식에 있어서 언어 모델링은 음성인식된 결과에 대한 언어적인 제약을 가하기 위해서 사용되는 지식으로써, 화자에 의해서 발성된 음성으로부터 음향적인 해독과정을 통해 얻어진 인식 어휘들의 열에 대해 구문적인 수준에서의 제약을 주기 위해서 사용된다. 음성인식에 언어 모델을 적용한 연구는 IBM의 연속 음성인식 그룹의 초기 연구인 source-channel 이론에 기반하여 음성 신호와 단어 열간의 사상 관계를 통계적인 방법으로 모델링하는 연구[1]를 시작으로, 지금까지 언어 모델링과 음성인식을 결합하는 다양한 연구들이 수행되고 있다. 화자가 발성한 음성 신호와 인식될 인식 어휘 열간의 관계는 아래와 같은 관계를 갖는다[8, 10].

$$\hat{W}=\mathrm{argmax}_{w} P(W \mid O)=\mathrm{argmax}_{w} \frac{P(W)P(O \mid W)}{P(O)} \quad (1)$$

여기서, O는 음성 신호에 대응되는 음향적인 특성을 나타내는 특징 파라미터의 열을 나타내며, W는 음성 신호에 대응되는 단어 열을 나타낸다. 위의 식에 의하면, 음성인식의 궁극적인 목표란 입력된 음성 신호 O에 대하여 인식 가능한 단어 열들 가운데에서 최적으로 정합하는 단어 열인 W를 구하는 문제 P(W | O)로 볼 수 있다. 이때, P(W | O)은 Baysian규칙에 의해서 위의 식과 같이, 단어 열의 출력 확률 P(W)과 단어 열에 대한 음성 신호의 출력 확률 P(O | W)의 곱을 음성 신호의 출력 확률 P(O)로 나누는 값으로 추정할 수 있다.

3.2 음성인식을 위한 언어 모델링

음성인식을 위해서 사용되는 언어 모델로는 크게 통계에 기반한 언어 모델과 FSA(finite state automata)형태에 기반한 모델링 방법들이 널리 사용되고 있다. 통계 기반한 모델링은 대량의 코퍼스를 기반으로 수집된 음성인식 대상의 텍스트 코퍼스를 대상으로 수집된 자료로부터 텍스트를 구성하는 단어들간의 전이 관계를 모델링하는 것으로, unigram, bigram, trigram 등이 이에 해당된다. 이러한 언어 모델들은 1,000~20,000단어를 대상으로 한 대용량 음성인식을 수행하는 경우에 일반적으로 적용된다. 이에 반해서, FSA에 기반한 경우는 인식 대상 영역이 제한적이며, 사용되는 어휘 수가 그리 크지 않고, 발성되는 문장의 형태가 패턴화되어 있는 경우에 사용된다. 로봇을 제어하고나 컴퓨터 상의 테스크톱에서 응용 프로그램의 수행이나 간단한 명령어 등을 수행하는 경우에 활용된다. 이 장에서는 음성인식을 위해서 사용되는 다양한 언어 모델링 방법들을 중심으로 기술하고자 한다.

3.2.1 FSN(finite state network) language Model[11, 12]

형식 언어인 FSN에 기반한 언어모델은 소규모 어휘들을 대상으로 명령이나 제어를 위한 짧은 문장 형식의 고정된 구문형태를 갖는 문장들을 인식하게 위해서 사용된다. CFG(context free grammar)형태로 표시된 문장이나 구, 절 등을 구성하는 단어 문법을 FSN형태로 표시하여 인식을 수행하게 된다. 이러한 방법은 단어의 앞/뒤 순서관계가 명확함으로써 인식의 오류를 줄일 수 있게 된다. FSN을 이용한 방법의 경우, 인식대상 어휘들로 구성된 문법들을 이용해서 하나의 커다란 어휘 탐색 네트워크를 만들고, 이러한 어휘 탐색 네트워크

를 통해서 입력된 신호에 대한 정합을 수행하게 된다. 이때, 어휘 탐색 네트워크가 인식 어휘들간의 순서 관계를 나타내는 언어모델로 사용되게 된다. FSN형태의 언어 모델로는 하나의 인식 어휘에 올 수 있는 어휘 수가 인식기에서 인식할 총 어휘수와 동일한 nogram, 그리고 단어간에 올수 있는 제약을 이진값으로 제약하는 word pair가 있다. 이들 각각에 대해서 살펴보기로 하자.

● **nogram** nogram이란 말 그대로 문법적인 제약이 가해지지 않은 것으로, 한 인식어휘의 다음에서 인식기에서 인식할 어떠한 어휘도 뒤이어 올 수 있다. 예를 들어서, DARPA의 RM(resource management)시스템에서 사용되는 991단어를 대상으로 인식을 수행할 경우, nogram의 경우 인식기의 복잡도는 991이다. 이 언어모델은 타당한 문장들의 커버리지가 완벽하나, 인식할 적합한 문장 이상을 커버하는 초과로 인해서 인식기의 성능이 저하될 수 있는 특징을 갖는다. nogram을 이용한 FSN의 예는 아래의 그림 1과 같다.

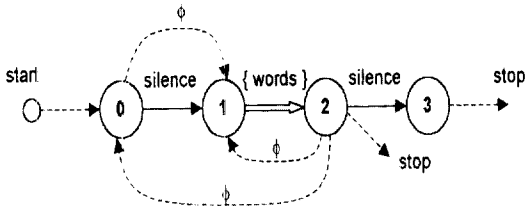

그림 1 nogram문법을 갖는 FSN

위의 그림은 발성 가능한 단어들의 그룹 다음에 휴지가 오는 경우를 모델링하고 있다. 이러한 언어모델은 다른 언어모델에 비해서 널리 사용되지 않고 있으나, 인식기를 구성하는 음향모델의 성능을 측정하기 위해서 사용되며, 다른 언어모델과의 비교 항목으로 종종 사용된다.

● **word pair** 음성인식을 위한 언어 모델링에서 FSN을 갖는 두번째 형태로는 단어쌍(WP : word pair)이 있다. 이 문법은 하나의 인식 대상 어휘 다음에 올 수 있는 어휘를 이진값의 형태로 표시한 것으로, nogram보다는 낮은 복잡도를 갖는다. 예를 들어서, RM영역에서의 991 단어에 대해서 단어쌍을 사용할 경우 복잡도는 60으로, nogram보다는 낮지만 여전히 높은 초과 커버리지를 갖는다. 단어쌍을 언어 모델로 사용하는 경우는 nogram에 비해서 보다 효과적인 탐색구조를 가진다.

3.2.2 Stochastic language Model[8, 9, 13]

대어휘 음성인식의 경우, 단어 열 W는 화자가 의도하는 바를 전달하며, 언어모델은 음성인식율을 향상시키기 위해서 필요하다. 언어 모델인 n개의 단어들로 구성된 단어열 $W=\{w_1, w_2, \cdots, w_3\}$가 주어졌을 때, 언어 모델의 확률을 다음과 같이 표시할 수 있다.

$$p(W)=p(w_1, w_2, \cdots, w_n)=\prod_{i=1}^{n} P(w_i \mid w_{i_0}, \cdots, w_{i-1}) \quad (2)$$

여기서, w_0은 초기확률을 나타낸다. 다음 단어인 w_i는 지금까지 발성된 단어들로 구성된 이력(history) h_i에 의존하게 된다. 만일, 이력의 크기가 커진다면, 모델의 복잡도는 지수적으로 커지게 된다. 따라서, 이러한 문제들을 실제적으로 다루기 위해서, 이력의 일부만을 다음 단어의 출력확률의 추정에 적용하는 방법이 등장하게 되었다. 이러한 것을 위해서 가능한 방법의 하나는 이력들의 공간을 K개의 클래스들로 분할하는 사상함수 $\Psi()$를 사용하는 것이다. 모델은 다음과 같다.

$$p(w_i \mid h_i) \approx p(w_i \mid \Psi(h_i)) \quad (3)$$

지난 20년동안 가장 성공적인 모델들의 하나가 n-gram이며, 특히 현재 단어의 예측을 위해서 이전의 한 단어를 고려하는 bigram과 이전의 두단어를 이력으로 사용하는 trigram이 널리 사용되고 있다. bigram언어 모델을 이용한 주어진 단어열에 대한 출력확률은 다음과 같다.

$$p(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \quad (4)$$

trigram언어 모델을 이용한 주어진 단어열에 대한 출력확률은 다음과 같다.

$$p(W) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1}) \quad (5)$$

trigram의 확률을 추정하기 위해서, 학습 코

퍼스라 불리우는 다량의 텍스트 코퍼스를 이용해서 trigram의 빈도를 추정한다.

$$f_3(w_3 \mid w_1, w_2) = \frac{c_{123}}{c_{12}} \quad (6)$$

여기서, c_{123}은 단어열 $\{w_1, w_2, w_3\}$의 출현 횟수를 나타내며, c_{12}는 단어열 $\{w_1, w_2\}$의 출현 횟수를 나타낸다. 단어수가 V일 경우, V^3가지 수의 trigram이 가능하다. 그러나, 실제적으로 이러한 조합수가 텍스트 코퍼스에 존재하지 않으므로, 많은 수의 trigram확률이 zero의 값을 가지게 된다. 따라서, 이러한 문제를 해결하기 위해서 학습 코퍼스에 존재하지 않은 trigram 확률을 평활화하기 위한 방법이 사용된다. 이를 위해서 일반적으로 어휘들에 대한 trigram, bigram, unigram 그리고 unform 분포의 확률을 선형결합하는 방법이 사용된다.

$$p(w_3 \mid w_1, w_2) = \lambda_3 f_3(w_3 \mid w_1, w_2) + \lambda_2 f_2(w_3 \mid w_2) + \lambda_1 f_1(w_3) + \lambda_0 \frac{1}{V} \quad (7)$$

여기서, $f_1()$와 $f_2()$은 적절한 bigram과 unigram의 빈도의 비율로써 추정된다. 선형 결합을 위해서 사용되는 가중치들은 n-gram의 추정에 사용된 자료와는 다른 다른 자료들의 확률을 최대화하도록 추정된다. 이러한 최대 출력확률의 추정을 위해서 전향-후향(forward-backward) 알고리즘이 사용된다. 이외에 평활화를 위해서 사용되는 방법으로는, deleted interpolation, backing-off, co-occurrence, count re-estimation 등이 있다.

4. 결 론

앞서 음성인식을 위해서 사용되는 언어 모델들에 대해서 살펴보았다. 인간이 언어를 이해하는 전체과정 측면에서 살펴보면 음성인식이란 전처리부에 해당된다. 화자에 의해서 발성된 음성으로부터 그러한 음성을 인간이 내부적으로 처리하기 위한 내부적인 포맷으로 변환하기 이전 단계까지에 해당되는 것으로, 여기에 음성인식과 언어 처리 기술의 접합이 필요하게 된다. 음성인식 자체만으로는 인식에 한계가 있다. 즉, 음성인식의 대상인 음성이 포함하는 정보들이 다양한 계층의 지식들이 복합적으로 포함되어 있으므로, 단순히 음성이라는 신호를 인식 어휘의 열로 변환하는 것만으로는 음성인식의 목표를 달성할 수 없다. 인식된 인식 어휘들의 열을 인간의 언어적인 특성을 고려하여 이해할 수 있는 열로써 변환하는 것이 필요하게 된다. 이러한 변환을 위해서 언어모델이 사용되며, 음성인식의 음향적 해독에 따른 오류를 보정하기 위해서 유효하다. 보다 향상된 음성인식을 위해서 정교한 언어모델이 요구되며, 음성신호의 해독과정인 음향모델과의 정합과정과 잘 결합되는 언어모델의 사용이야말로 대어휘 음성인식을 위한 궁극적인 해결책이 될 것으로 판단된다.

참고문헌

[1] R. A. Cole, J. Mariani, et al, "Survey of the State of the Art in Human Language Technology," *Techincal Report*, NSF, 1995.

[2] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine.

[3] S. Young, G. Bloothooft, *Corpus-Based Methods in Language and Speech Processing*, Kluwer Academic Publishers, 1997.

[4] K. F. Lee, *Automatic Speech Recognition-The Development of the SPHNIX System*, Kluwer Academic Publisher, Boston, 1989.

[5] C. H. Lee, L. R. Rabiner, et al., "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, Vol. 4, pp. 137~1,165, Jan., 1990.

[6] M. Weintrabub et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. of ICASSP 89*, pp. 699~702, May 1989.

[7] R. Pieraccini, C. H. Lee, "Factorization of Language Constraints in Speech Recognition," Proc. 29th Annual Meet-

ing of the Association for Computational Linguistics, June, 1991.

[8] F. Andry, S. Thornton, "A Parser for speech lattices using a UCG grammar," *Proc. 2nd European Conference on Speech Communication and Technology*, pp. 219~222, 1991.

[9] S. Austin, R. Schwartz, et al., "The Forward-Backward Search Algorithm," *Proc. ICASSP 91*, pp. 697~700, 1991.

[10] F. Jelinek, L. R. Bahl, et al., "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," *IEEE trans. on Information Theory*, Vol. 21, No. 3, pp. 250~256, 1975.

[11] P. Baggia, E. Gerbino, et al, "Efficient representation of linguistic knowledge in continuous speech understanding," *Proc. of IJCAI 91*, 1991.

[12] A. M. Derouault, B. Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription," *IEEE trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, pp. 742~749, 1984.

[13] E. Giachin, "Automatic Training of stochasitc funite-state language models for speech understanding," *Proc. of ICASSP*, 1992.

[14] T. Booth, Probabilistic Representation of Formal Language, *In Tenth annual IEEE symposium on Switching and Automata Theory.*

[15] R. Rosenfeld, "The CMU Statistical Language Modeling Toolkit for Language Modeling and its Use in the 1994 ARPA CSR Evaluation," Proc. of Spoken Language Systems Technology Workshop, pp. 47~50, 1995.

최 환 진

1990 고려대학교 전산학과(학사)
1992 한국과학기술원 전산학과(석사)
1997 한국과학기술원 전산학과(박사)
1997~현재 시스템공학연구소 자연어정보처리연구부(선임연구원)
관심분야 : 음성인식, 화자인식, 신경회로망, 퍼지이론, 사용자 모델링

박 재 득

1983 서울대학교 계산통계학과(학사)
1985 한국과학기술원 전산학과(석사)
1989~1991 삼성종합기술원 정보시스템연구소 선임연구원
1992~1995 삼성전자 멀티미디어연구소 선임연구원
1994 한국과학기술원 전산학과(박사)
1995~현재 시스템공학연구소 자연어 정보처리연구부 실장
관심분야 : 언어처리, 사용자 모델링, 인공지능

박 동 인

1979 서강대학교 전자공학과(학사)
1979~현재 시스템공학연구소 자연어정보처리연구부 부장
1994 공업진흥청 산업표준 심의회 위원
1995 국어정보학회 이사, 문화체육부 국어심의회 위원
1996 한글정보과학회 한국어정보처리 연구회 위원장, 한국정보과학회 평의원
관심분야 : 자연언어이해, 기계번역, 정보검색, 한국어정보처리