韓國컴퓨터情報學會 論文誌
第14卷 第8號, 2009. 8.

2009-14-8-3-2

정보 검색을 위한 숫자의 해석에 관한 구문적·의미적 판별 기법

문 유 진*

Syntactic and Semantic Disambiguation for Interpretation of Numerals in the Information Retrieval

Moon, Yoo-Jin*

요 약

월드 와이드 웹의 정보 검색에서 산출되어지는 수많은 정보를 효율적으로 검색하기 위해서 자연어 정보처리가 필수적이다. 이 논문은 텍스트에서 숫자의 의미 파악을 위한 판별기법을 제안한 것이다. 숫자 의미 판별기법은 챠트 파싱 기법과 함께 문맥자유 문법을 활용하여 숫자 스트링과 연관된 접사를 해석하였으며, N-그램 기반의 단어에 의거하여 조직화된 의미 파악을 하도록 설계되었다. 그리고 POS 태거를 사용하여 트라이그램 단어의 제한조건이 자동 인식되도록 시스템을 구성하여, 점진적으로 효율적인 숫자의 의미 파악을 하도록 하였다. 이 논문에서 제안한 숫자 해석 시스템을 실험한 결과, 빈도수 비례 방법은 86.3%의 정확률을 나타냈고 조건수 비례 방법은 82.8%의 정확률을 나타냈다.

Abstract

Natural language processing is necessary in order to efficiently perform filtering tremendous information produced in information retrieval of world wide web. This paper suggested an algorithm for meaning of numerals in the text. The algorithm for meaning of numerals utilized context-free grammars with the chart parsing technique, interpreted affixes connected with the numerals and was designed to disambiguate their meanings systematically supported by the n-gram based words. And the algorithm was designed to use POS (part-of-speech) taggers, to automatically recognize restriction conditions of trigram words, and to gradually disambiguate the meaning of the numerals. This research performed experiment for the suggested system of the numeral interpretation. The result showed that the frequency-proportional method recognized the numerals with 86.3% accuracy and the condition-proportional method with 82.8% accuracy.

• 제1저자 : 문유진
• 투고일 : 2009. 06. 23, 심사일 : 2009. 06. 24, 게재확정일 : 2009. 08. 04.
* 한국외국어대학교 경영정보학과 교수

▸ Keyword : 의미적 판별(semantic disambiguation), 숫자(numeral), 트라이그램 단어(trigram word), 토큰화 처리(tokenization), 문맥자유 문법(context-free grammar)

Ⅰ. 서 론

인터넷의 웹검색(web surfing)은 지식정보화 사회에서 비즈니스 뿐만 아니라 실생활에 매우 큰 영향을 미치고 있다. 웹검색에서 제공하는 결과물은 유용한 정보 뿐만 아니라 수많은 가비지(garbage)를 포함하고 있어서 지능형 웹검색(intelligent web surfing)의 필요성이 강조되고 있다. 이를 위해서는 자연어 정보처리가 기반이 되어서 정보 검색이 수행되어져야 한다. 이 정보 검색은 문장의미의 파악을 포함하는 것으로, 이 논문에서는 문장에 나오는 숫자 스트링의 의미 파악을 위한 방법을 제시하고자 한다.

사용자가 웹검색창에 "Benz speed"를 입력하여 정보 검색을 하고자 할 때, 기대하는 산출물은 'Benz'와 'speed'란 글자가 포함되어 있는 문장들뿐만 아니라 'Benz'와 '140 km/h' 혹은 '80 mile/h' 등이 포함된 문장들이다. 사용자가 기대하는 지능형 산출물은 정보 검색에서 자연어의 숫자 스트링에 대한 의미분석이 수행되어져야 가능한 것이다[1][2]. 이러한 지능형 정보 검색은 자연어 문장의 구문적 판별과 의미적 판별을 수행하여야 가능한 것으로, 이 논문에서는 자연어 문장에 포함된 숫자 스트링과 관련 접사의 의미 판별을 수행하고자 한다.

숫자 스트링의 의미 파악을 위한 연구가 현재까지는 국내에서 거의 이뤄지지 않았고 해외에서도 별로 이뤄지지 않았다. 숫자 스트링에 관한 기존의 연구를 살펴보면 다음과 같다. ICE-GB 문법[3]은 기수, 서수, 분수, 하이픈으로 연결한 수, 단수와 복수의 특성을 갖는 수 가운데 하나로 숫자를 처리했다. Maynard[4]는 숫자의 명명된 개체 인식을 위해서 의미론적 범주를 사용했고, Wang[5]은 숫자의 명명된 개체 인식을 위하여 한자의 의미론적 분류 시스템을 사용했다. Asahara[6]는 어휘분석, POS 태깅과 청킹(chunking) 기법 등을 제한된 형태의 숫자 스트링을 해석하는데 적용하였던 것이다. 일본어에서 돈과 온도 등을 해석하는 데 사용하기 위하여 규칙기반 기법으로 숫자 분류기가 연구되었다[7]. Polanyi와 van den Berg[8]는 숫자와 한정 기호의 지시어 해결을 하기 위하여 한정어 논리 구조(constraint logic framework)을 사용하였다. Zhen과 Su[9]는 HMM 기반으로 청킹 태거를 이용하여 숫자, 이름, 연도, 양을 4 개의 의미자질(semantic feature)과 11 개의 표면 하위자질(surface subfeature)로 인식하고 분류하였다. MUC의 FACILE[10]은 챠트 파싱 기법과 함께 규칙기반을 사용하여 명명된 개체 인식 시스템을 구성하였으며, 의미 범주(예, 사람, 조직, 날짜, 시간)를 활용하였다. 하지만 본 논문에서 제안하는 숫자 스트링 판별 기법에서 수행하는 의미적 판별 기법을 FACILE[10]에서는 수행하지 않아서 의미 해석의 성능이 뒤떨어진다.

기존의 연구들은 제한된 형태의 숫자 스트링을 해석하는데 중점을 두었으나, 본 논문은 광범위하고 일반적인 형태의 숫자 스트링을 대상으로 하여 기존의 연구보다 더 많은 문법규칙과 의미범주를 설정함으로써 광범위한 숫자 스트링에 대하여 구문적 판별을 효과적으로 수행하도록 한다. 그리고 기존의 연구에서는 구문적 판별에 중점을 두었으나 본 논문에서는 구문적 판별과 의미적 판별을 차례로 수행하여 정확성과 정교함을 향상시키도록 한다. 또한 이 논문에서 제안한 판별 기법은 숫자 스트링에 대한 인간의 의미 분석방법을 시뮬레이션(simulation)하여 정보시스템에 적용한 것이다. 즉, 구문 문법 규칙과 트라이그램 단어 제한조건을 사용하여 숫자 해석 시스템을 구현하며, 이 때 토큰나이저, 규칙기반 숫자 스트링 처리, 트라이그램 기반 의미 분석 등을 활용한다. 규칙기반 숫자 처리 시스템은 각 숫자 스트링을 그의 형태와 범주에 따라서 어휘/구문론적으로 분석한다.

문장 내에 있는 숫자 스트링의 구문을 해석하기 위해서는 숫자 스트링과 결합된 접사를 분석하는 것이 중요하다[11][12][13]. 예를 들어, '12m' 이라는 접사가 있을 때, 어휘처리 단계에서 숫자 스트링을 분석한다. 그 다음에, 숫자 스트링과 결합된 접사를 분석하는데 구문 분석에 기반한 규칙기반 처리를 수행한다. 숫자 스트링은 의미있는 접사와 함께 사용될 경우가 많기 때문이다. 예를 들면, '12m'은 LENGTH, NUMBER 혹은 MONTH 등의 의미 범주에 속할 수 있다. 위와 같이 어휘/구문적 분석 단계를 수행한 후에, 의미 분석 단계를 의미 분석 제한 조건을 활용하여 수행한다[14][15]. 이 단계에서 가능한 여러 의미 범주 중 하나를 선택한다.

이 논문에서는 숫자 스트링을 효율적으로 해석하는 데 필요한 판별기법을 제안하고자 한다. II 장에서는 이 논문에서 활용하는 지식베이스와 제안하는 시스템의 전체 개요를 설계하고자 한다. III 장에서는 숫자 스트링과 함께 사용된 접사의 해석 과정과 분류 규칙을 설명하고자 한다. IV 장은 시스템의 실험결과를 기술하고 V 장은 결론 및 향후연구를 기술하고자 한다.

II. 지식베이스 및 시스템 개요

1. 지식베이스를 활용한 구문적 판별 방법

숫자 스트링의 구문적 처리를 위하여 기본적 파싱 처리단계와 고급 토큰화 처리단계의 두 단계가 수행된다. 숫자 스트링과 접사를 파싱할 경우에, 접미어 혹은 접두어로서의 구조적 패턴 즉 구문적 기능어 관계가 숫자 스트링의 해석에 유용할 수 있다. 기본적 파싱 처리단계는 기본적으로 숫자, 알파벳 그리고 다른 기호들을 분리시키는 단계이다. 예를 들면, "20 ~ 30 min"의 스트링을 "20 (숫자)" + "~ (기호)" + "30 (숫자)" + "min (알파벳)" 으로 분리시키는 것이다. 기호에는 ".", "/", "(", ")", "~" 등이 있다.

고급 토큰화 처리 단계에서는 숫자 스트링과 접사의 작업을 위하여 구문 규칙이 사용된다. 구문 규칙은 64 개의 문맥자유 문법으로 제한조건을 기술한다. 예를 들면, SPEED의 구문 규칙은 다음과 같다. SPEED는 NUMBER, LU, SLASH, TU 의 4개 구성요소를 취하며, NUMBER는 속도량을 표현하고 LU는 길이 단위로 km, m, cm, mm 등을 표현하고 SLASH는 "/" 이며 TU는 시간 단위로 h, m, s, ms, hr, min, sec 등을 나타낸다.

(NUMBER LU SLASH TU)
Constraints: ((SPEED-AMOUNT-P NUMBER))
Where LU (Length Unit - km/m/cm/mm)
and TU (Time Unit - h/m/s/ms/hr/min/sec)

숫자 스트링 판별을 위한 64 개의 구문 규칙은 40 개의 숫자 스트링 의미 범주에 속하게 된다. 40 개의 숫자 스트링 의미 범주는 〈표 1〉에서 제시하듯이 Quantity, Range, Temperature, Daytime, Date, Age, Year, Money, Score 등이다. 그리고 기본적 파싱 처리 작업과 고급 토큰화 처리 작업에서 어휘 정보를 활용하기 위하여 유저 사전을 구축한다.

표 1. 숫자 스트링을 위한 의미 범주의 예
Table 1. Example of Semantic Categories

Semantic Categories	Semantic Categories
Quantity	Daytime
Money	Formatnumber
Date	Score
Year	Ordinal
Number	Length
Floatnumber	Name
Age	Range
Century	Temperature

유저 사전은 POS, 수 (number), 성 (gender), 격 (case) 등의 구문 정보를 활용한다. 유저 사전의 한 예는, 다음과 같이 단어 "₩"은 통화 단위(Currency Unit) 로서"WON"의 의미를 표현하는 것이다.

("₩" ((:POS CUR :SEM WON)))
where CUR is Currency Unit.

2. 지식베이스를 활용한 시스템 개요

숫자 스트링의 의미 파악을 위하여 〈그림 1〉과 같은 정보시스템을 설계한다. 〈그림 1〉에서 도시하듯이, 자연어 정보시스템에 문서가 입력되면 각 문장별로 숫자 스트링과 접사를 중심으로 토큰화와 추출 작업을 수행하며, 그 후 숫자 스트링과 접사를 분석하는 구문적 판별 작업을 진행하게 된다[16]. 이 때 챠트파싱 기법을 사용하여 문장을 파싱하고 고급 토큰화 작업을 수행하게 되는데, 문맥자유 문법 규칙과 유저 사전을 활용하게 된다. 그 다음 단계는 숫자 스트링과 접사의 의미적 판별 작업을 진행하는데, 이 때 유저사전과 트라이그램 단어에 기반한 제한조건들을 활용한다. 이러한 처리를 수행하여 숫자 스트링의 의미적 판별 결과가 두 개 이상일 경우에, 숫자 스트링과 접사에 대한 의미를 휴리스틱 선택 기법에 의하여 결정하도록 한다. 이 논문의 IV 장에서는 두 가지의 휴리스틱 선택 기법을 제안하고 비교 분석한다.

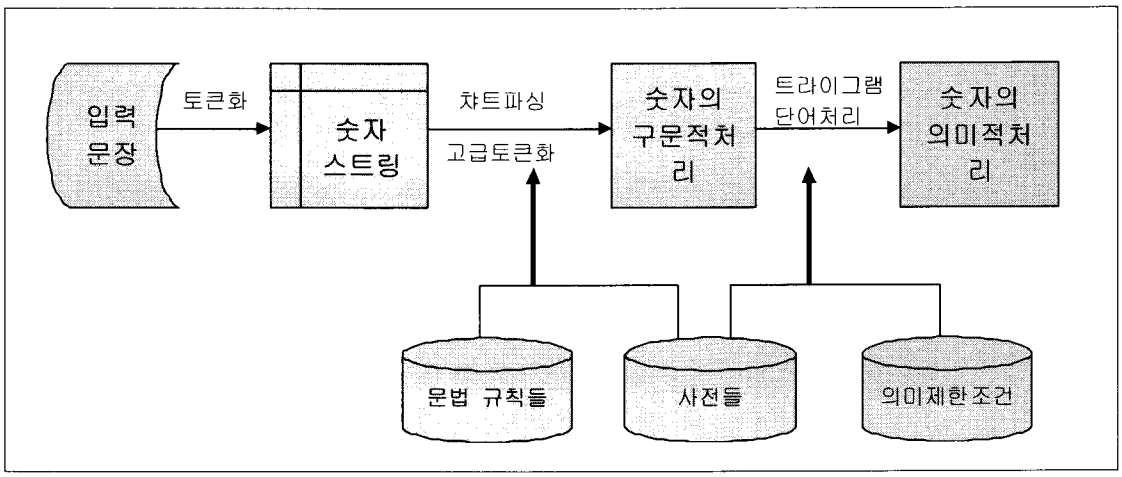

그림 1. 숫자 스트링 정보시스템 개요
Fig. 1. Outline of the Information System for Numeral Strings

이 연구에서는 영어로 쓰여진 웹문서 기사의 숫자 스트링을 대상으로 하여 〈그림 1〉에서 도시한 숫자 스트링 정보시스템을 적용한다. 다음 연구에서는 한글로 쓰여진 웹문서 기사를 대상으로 하여 숫자 스트링 정보시스템을 적용하고자 한다.

III. 숫자 스트링의 해석처리 기법

일반 웹문서에서 나오는 숫자 스트링의 분포는 실험 데이터의 조사 결과 약 1.9%에 달하였다. 인터넷의 신문기사를 주대상으로 분포를 조사하였다.

이 장에서 제안하는 숫자 스트링의 해석 처리 기법은 인간이 숫자 스트링을 인식하는 의미 분석 과정을 시뮬레이션한 것이다. 숫자 스트링 의미의 정확한 판별 처리를 수행하기 위하여, 구문적 판별 처리 기법과 의미적 판별 처리 기법을 차례로 적용하는 과정을 이 장에서는 기술한다.

숫자 스트링의 구문적 판별 처리를 하기 위하여 챠트파싱 기법을 사용하여 접사와 아울러 기본적 구문 파싱을 수행하게 된다. 또한 II 장에서 기술한 사항과 같이 숫자 스트링의 고급 토큰화 처리를 하기 위해서 64 개의 문맥자유 문법을 제한조건과 함께 사용한다. 발췌한 샘플 데이터에서 의미 범주에 관한 빈도수를 비교해보면, QUANTITY 〉 MONEY 〉 DATE 〉 YEAR 〉 LENGTH 등의 순서이다.

숫자 스트링의 의미적 판별 처리를 하기 위하여 트라이그램 단어와 제한조건을 구축하여 사용한다. 숫자 스트링의 의미적 해석을 처리하기 위한 트라이그램 단어의 제한조건은 다섯 가지 형태의 정보를 활용하여 구성한다. 다섯 가지 형태의 정보는 구문적/하위범주화 정보, 구두점 정보, 의미적 정보, 휴리스틱 정보 그리고 패턴매칭 정보이다.

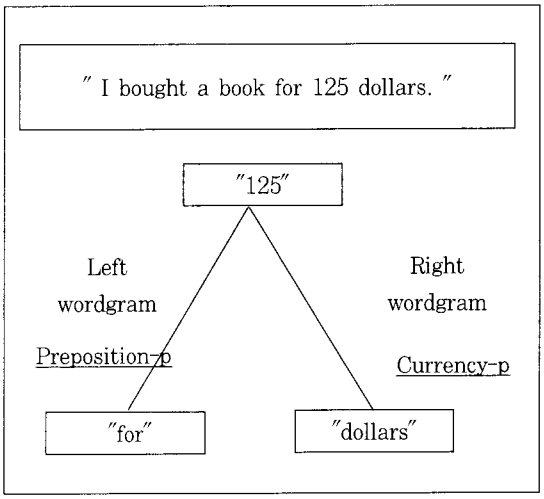

그림 2. 트라이그램 단어의 예
Fig. 2. An Example of Trigram Words

트라이그램 단어의 한가지 예를 〈그림 2〉에서 도시하고 있다. "I bought a book for 125 dollars." 란 문장에서 "for 125 dollars"에 대한 트라이그램이다. 〈그림 2〉에서 숫자 '125'를 중심으로 Left wordgram에 단어 'for'가 위치해 있으며 이는 preposition-p의 서술논리에 긍정적인 답을 준다.

그리고 숫자 '125'를 중심으로 Right wordgram에 단어 'dollars'가 위치해 있으며 이는 currency-p의 서술논리에 긍정적인 답을 준다. 〈그림 2〉 형식의 숫자 스트링을 중심으로 한 트라이그램 단어는 POS 태거를 사용하여 자동인식 되도록 시스템을 구성한다.

〈그림 3〉은 트라이그램 단어에서 다섯가지 제한조건의 각각 예를 도시하고 있다. 〈그림 2〉에서 제시한 트라이그램 단어의 의미 범주인 MONEY에 대하여, 다섯가지 형태 중 의미적 정보에 기반한 제한조건을 〈그림 3〉의 (1)과 같이 설계한다. 〈그림 3〉의 (1)은 '숫자'와 '통화표시'가 입력되었는지를 검사하는 의미적 판별처리를 위한 제한조건을 도시한다. 하나의 문장 "The thermometer rises to over 100 Fahrenheit in the hot months."에서 "over 100 Fahrenheit" 의 의미 처리를 위해서는 다섯 가지 형태의 트라이그램 단어 제한조건 중에서 휴리스틱 정보에 기반한 제한조건을 〈그림 3〉의 (2)에서 예시하고 있다. 〈그림 3〉의 (2)에서 '온도' 단위가 입력되었는지 그리고 '온도'의 숫자가 적정한 범위 내의 숫자인지 검사하는 휴리스틱 판별처리를 위한 제한조건을 표현한다.

```
(1) MONEY 제한조건의 예 (의미적 제한조건)
      "for 125 dollars"
  (and  (preposition-p  left-wordgram("for"))
        (currency-p  right-wordgram("dollars")))

(2) TEMPERATURE 제한조건의 예 (휴리스틱 제한조건)
      "over 100 Fahrenheit"
 (and (temperature-p right-wordgram("Fahrenheit"))
        ( 〉  numeral-string("100") -3000)
        ( 〈=  numeral-string("100")  3000))

(3) LENGTH 제한조건의 예 (구문적/하위범주화 제한조건)
      "about 180 meters"
  (and  (preposition-p left-wordgram("about"))
   (plural-noun-p  right-wordgram("meters")))

(4) SCORE 제한조건의 예 (패턴매칭 제한조건)
      "with 3-1 victory"
  (and
    (preposition-p   left-wordgram("with"))
    (score-right-pattern right-wordgram("victory")))

(5) AGE 제한조건의 예 (구두점 제한조건)
      "Smith, 28, has"
  (and  (comma-p  left-wordgram("Smith,"))
     ( capital-letter-p left-wordgram("Smith,"))
     ( comma-p numeral-string("28,")))
```

그림 3. 트라이그램 단어 제한조건의 예
Fig. 3. Examples of Constraints for Trigram Words

IV. 실험결과

이 논문에서 제안하는 숫자 스트링 해석 시스템의 샘플데이터를 온라인 신문에서 수집하였다. 약 15일에 걸쳐서 사회, 정치, 경제, 문화, 스포츠 등 각 분야의 기사를 대상으로 하였다. 이 샘플데이터를 입력 문장으로 하여, 〈그림 1〉에서 제시한 숫자 스트링 정보시스템을 구현하였다. 이 시스템은 입력 문장의 숫자 스트링 의미 판별을 하기 위하여, 토큰화 및 구문적 처리 그리고 의미적 처리를 수행하였다.

트라이그램 단어 구축 방법에서 표현된 제한 조건 즉 문맥정보는 숫자 스트링 해석 처리를 위한 마지막 단계에서 의미범주를 결정하는데 적용되어진다. 마지막 단계에서 제한조건을 적용하여 두 개 이상의 의미 범주가 결과로서 산출되었을 경우에 한 개의 의미 범주 결과를 선택하여야 한다. 이를 위하여 이 연구에서는 두 가지 휴리스틱 방법을 제안한다. 즉, 빈도수 비례 방법과 조건수 비례 방법이다.

첫 번째로, 빈도수 비례 방법은 샘플데이터를 활용하여 구문 규칙의 범주에 숫자 스트링의 해석을 적용한 결과, 산출된 의미 범주의 빈도수를 기준으로 의미를 채택하는 것이다. 일치되는 구문규칙이 발견되지 않을 경우에는 최다 빈도수를 가진 의미 범주를 채택한다. 발췌한 샘플 데이터에서 의미 범주에 관한 빈도수를 비교해보면, QUANTITY 〉 MONEY 〉 DATE 〉 YEAR 〉 LENGTH 등의 순서이다.

두 번째로, 조건수 비례 방법은 의미 범주에 대한 적용 빈도수가 존재하지 않을 경우에 의미 범주에 대한 제한조건수를 기준으로 의미를 채택하는 것이다. 일치되는 구문 규칙이 발견되지 않는 경우에는 최다 제한조건수를 가진 의미 범주를 채택한다. 이 방법에서 사용하기 위하여 발췌한 샘플 데이터에서 의미 범주에 대한 제한조건수를 비교해보면, QUANTITY(22개 제한조건) 〉 LENGTH(8개 제한조건) 〉 YEAR(7개 제한조건) 〉 AGE(5개 제한조건) 순이었다.

예를 들어, "He has 2m desks." 란 문장이 입력되었을 때, 이 논문에서 제안한 시스템은 구문적 판별과 트라이그램 단어 제한조건의 의미적 판별을 적용한 결과로 '2m' 의 의미 범주를 'LENGTH' 와 'NUMBER' 로 산출하였다. 이 때, '2m' 에 대한 복수 개의 의미 범주가 산출되었으므로 휴리스틱 선택 방법을 적용하여야 한다.

빈도수 비례 방법을 적용하면, LENGTH의 의미 범주의 빈도수가 NUMBER의 의미 범주의 빈도수보다 많으므로 LENGTH의 의미 범주가 채택된다. 어떤 의미 범주의 빈도수가 더 많다는 것은 그 의미 범주가 더 자주 사용된다는 의

미로 해석 가능하다.

조건수 비례 방법을 적용하면, LENGTH의 의미 범주의 제한조건수가 NUMBER의 의미 범주의 제한조건수보다 많으므로 LENGTH의 의미 범주가 채택된다. 어떤 의미 범주의 제한조건수가 더 많다는 것은 그 의미 범주가 더 다양하게 사용된다는 의미로 해석 가능하다.

표 2. 빈도수 비례방법과 조건수 비례방법의 성능비교
Table 2. Performance Comparison between Frequency Proportional Method and Condition Proportional Method

	정확률 (precision ratio)	재현률 (recall ratio)
빈도수 비례방법	86.3 %	77.5 %
조건수 비례방법	82.8 %	74.2 %

실험을 위하여, 167개의 숫자 스트링 샘플데이터는 숫자 스트링 해석을 위한 지식베이스 및 구문 규칙과 트라이그램 단어의 제한조건 규칙을 만드는 데 사용되었다. 그리고 테스트데이터인 숫자 스트링 473개는 이 시스템을 테스트하는 데 사용되었다. 샘플데이터 총 7952개의 스트링 중에서 숫자 스트링은 167개 (2.1%)이었고, 테스트데이터 총 24895개의 스트링 중에서 숫자 스트링은 473개 (1.9%)이었다.

이 논문에서 제안한 숫자 스트링 해석 시스템을 테스트데이터를 사용하여 실험한 결과, 빈도수 비례 방법은 86.3%의 정확률(precision ratio)을 나타냈고 조건수 비례 방법은 82.8%의 정확률을 나타냈다는 것을 〈표 2〉에서 보여주고 있다. 테스트데이터를 사용하여 숫자 스트링의 판별 방법에 대하여 실험한 결과, 빈도수 비례 방법은 77.5%의 재현률(recall ratio)을 나타냈고 조건수 비례 방법은 74.2%의 재현률을 나타냈다는 것을 〈표 2〉는 보여주고 있다. 실험 결과에서 제시한 정확률과 재현률에 의하면 이 정보시스템은 가치있는 연구 결과라고 할 수 있으며 실제 정보검색시스템에 접목하여 활용시 도움이 된다고 말할 수 있다.

V. 결론 및 향후 연구

지식 정보화 사회에서 산출되어지는 수많은 정보를 효과적으로 검색하기 위해서 자연어 정보처리가 필수적이다[1][17][18]. 이 자연어 정보처리 중에서 숫자 스트링의 의미 파악을 위한 연구는 기존에 많이 진행되지 못한 분야이다. 이 논문은 영어로 쓰여진 웹문서 기사를 대상으로 하여 숫자 스트링의 의미 파악을 위한 판별기법을 제안하였다. 이 논문에서 제안하는 숫자 스트링 해석 시스템의 샘플데이터를 온라인 신문에서 수집하였다. 사회, 정치, 경제, 문화, 스포츠 등 각 분야의 기사를 대상으로 하였다.

숫자 스트링 해석 판별기법은 챠트 파싱 기법과 함께 문맥자유 문법을 활용하여 숫자 스트링과 연관된 접사를 해석하였으며 N-그램 기반의 단어에 의거하여 조직적인 판별 분석 시스템을 설계하였다. 그리고 트라이그램 단어 제한조건이 POS 태거를 사용하여 자동인식 되도록 시스템을 구성하여, 점진적으로 효율적인 숫자 스트링의 의미 파악을 하도록 하였다. 실험을 위한 샘플데이터 총 7952개의 스트링 중에서 숫자 스트링은 167개 (2.1%)이었고, 테스트데이터 총 24895개의 스트링 중에서 숫자 스트링은 473개 (1.9%)이었다. 이 논문에서 제안한 숫자 스트링 해석 시스템을 실험한 결과, 빈도수 비례 방법은 86.3%의 정확률을 나타냈고 조건수 비례 방법은 82.8%의 정확률을 나타냈다.

기존의 연구들은 제한된 형태의 숫자 스트링을 해석하는데 중점을 두었으나, 본 논문은 광범위하고 일반적인 형태의 숫자 스트링을 대상으로 하여 기존의 연구보다 더 많은 문법규칙과 의미범주를 설정함으로써 광범위한 숫자 스트링에 대하여 구문적 판별을 효과적으로 수행하도록 하였다. 그리고 기존의 연구에서는 구문적 판별에 중점을 두었으나 본 논문에서는 구문적 판별과 의미적 판별을 차례로 수행하여 정확성과 정교함을 향상시켰다.

향후 연구 방향은 다음과 같다. 첫째, 숫자 스트링의 모든 패턴이 규칙으로 지정되어 있지 않아서 규칙의 지속적인 업데이트가 필요하다. 규칙적이고 지속적인 업데이트를 반자동화 또는 자동화시키기 위한 방법을 연구할 필요가 있다. 둘째, 이 연구에서는 영어로 쓰여진 웹문서 기사의 숫자 스트링을 대상으로 하였으나 향후 연구에서는 한글로 쓰여진 웹문서 기사의 숫자 스트링을 대상으로 하여 숫자 스트링 정보시스템을 적용하고자 한다. 셋째, 자연어 정보처리 시스템 및 정보 검색 시스템과 연계하여 실제적으로 활용함으로써 실생활에 접목시킬 필요성이 있다.

참고문헌

[1] Fensel, D., Hendler, J., Lieberman, H. & Wahlstet, W., "Spinning the Semantic Web," MIT Press, 2003.

[2] Jhingran, A. D. & Pirahesh, M. N., "Information Integration: A Research Agenda," IBM System Journal, Vol.41, No.4, pp.555-562, 2002.

[3] Nelson, G., Wallis, S. & Arts, B., "Exploring Natural Language - Working with the British Component of the International Corpus of English," John Benjamins, The Netherlands, 2002.
[4] Maynard, D., Tablan, V., Ursu, C., Cunningham, H. & Wilks, Y., "Named Entity Recognition from Diverse Text Types," Proceedings of Recent Advances in NLP, 2001.
[5] Wang, H. and Yu, S., "The Semantic Knowledge Base of Contemporary Chinese and its Application in WSD," Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp.112-118, 2003.
[6] Asahara M. & Matsumoto, Y., "Japanese Named Entity Extraction with Redundant Morphological Analysis," Proceedings of HLT-NAACL 2003, pp.8-15, 2003.
[7] Siegel, M. & Bender, E. M., "Efficient Deep Processing of Japanese," Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization, 2002.
[8] Polanyi, L. & van den Berg, M., "Logical Structure and Discourse Anaphora Resolution," Proceedings of ACL99 Workshop on The Relation of Discourse/Dialogue Structure and Reference, pp.10-117, 1999.
[9] Zhou, G. & Su, J., "Named Entity Recognition using an HMM-based Chunk Tagger," Proceedings of ACL 2002, pp.473-480, 2002.
[10] Black, W., Rinaldi, F. & Mowatt, D., "FACILE: Description of the NE System used for MUC-7," Proceedings of MUC-7, 1998.
[11] Chieu, L. & Ng, T., "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," Proceedings of the 19th COLING, pp.190-196, 2002.
[12] CoNLL-2003 Language-Independent Named Entity Recognition,*http://www.cnts.uia.ac.be/conll2003/ner/2, 2003.*
[13] Reiter, E. & Sripada, S., "Learning the Meaning and Usage of Time Phrases from a Parallel Text-Data Corpus," Proceedings of HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data, pp.78-85, 2003.
[14] Piprani, B., "Towards a Common Platform to Support Business Processes, Services and Semantics," 12th Annual Open Forum for Metadata Registries, 2009.
[15] Davis, J., "Semantic Frameworks: Meanings in the Architecture," 12th Annual Open Forum for Metadata Registries, 2009.
[16] Dale, R., "A Framework for Complex Tokenization and its Application to Newspaper Text," Proceedings of Australian document Computing Symposium, 1997.
[17] 이경호, 양룡, 이상범, "색상 정보를 이용한 자동 독화 특징 추출," 한국컴퓨터정보학회 논문지, 제13권, 6호, 107-116쪽, 2008년 11월.
[18] 김선옥, 이경호, " 얼굴 특징점을 이용한 한국어 8모음 독화시스템 구축," 한국컴퓨터정보학회 논문지, 제16권, 2호, 135-140쪽, 2008년 12월.

저 자 소 개

문 유 진

1987: 펜실베니아주립대학교 전산학 석사
1996: 서울대학교 컴퓨터공학박사
1997: 펜실베니아대학교 Post-Doc.
2002 - 현재
한국외국어대학교 경영정보학과 교수
관심분야: 정보검색, 지능정보시스템, 자연어처리