Identification of Numerical Expressions using Finite-State Automata
Identification of Numerical Expressions using Finite-State Automata

Ock-Jong Baek

Department of Computer Engineering, Korea Maritime University, Pusan, Korea

Abstract

There have been many trials to parse sentences in text to search complete and exact parses, but it is very hard because of unavoidable incompleteness of lexicon and grammar. Recently, to alleviate these difficulties, partial parsing appears as an alternative in the field. Partial parsing aims to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis.

As a part of partial parsing, the identification of Korean numerical expressions in text is described in this paper. Numerical expressions are required in several systems such as information extraction systems and question-answering systems. One of desired characteristics of these systems is the fastness. To achieve this goal, we use a finite-state automaton, for which we could use a tool like lex. So that we could rapidly implement the system. We observed that the system is fast and correct through several experiments. To evaluate our system, we used newspaper as test collection. We achieved the recall of 90.8%, and the precision of 86.9%. Experiments show that our system is comparatively correct.
1ex[] 1ex[] 1ex[] 1ex[] 1ex[] [9- 11].

ÀÌ¿ëÇѼö½ÄÇ¥ÇöÀνĽýºÅÛÀºÁÖ°¡Á¤º¸

ȯÀ²Á¤º¸

Àϱ⿹º¸Á¤º¸µîÀ»ÃßÃâÇÏ´Â

Á¤º¸ÃßÃâ½Ã½ºÅÛ

À̳ªÁúÀÇÀÀ´ä½Ã½ºÅÛ

[5, 7]

À̳ªÁúÀÇÀÀ´ä½Ã½ºÅÛ

[12]

¿¡³Î¸®»ç¿ëµÉ¼öÀÖ´Ù

±×¹Û¿¡

µµÁÖ¾îÁø¹®Àå¿¡¼ö½ÄÇ¥ÇöÀ»¸ÕÀúÀνÄÇÔÀ¸·Î½áÀÚ¿¬¾ð¾î󸮽ýºÅÛÀÇó¸®

ºÎ´ãÀ»´ú¾îÁٻӾƴ϶ó¼ÓµµµµÅ©°Ô°³¼±ÇÒ¼öÀÖÀ»°ÍÀÌ´Ù

º»³í¹®ÀDZ¸¼ºÀº´ÙÀ½°ú°°´Ù

Á¦

2

Àå¿¡¼­´Â°ü·Ã¿¬±¸·Î¼­Á¤±ÔÇ¥Çö°úÀ¯ÇÑ

»óÅ¿ÀÅ丶ŸÀǰü°è

ÀÚ¿¬¾ð¾î󸮿¡¼­À¯ÇÑ»óÅ¿ÀÅ丶Ÿ¸¦ÀÌ¿ëÇÑ»ç·Ê

±×

¸®°í¼ö½ÄÇ¥ÇöÀνĽýºÅÛÀÇÀÀ¿ë¿¡´ëÇØ¼­»ìÆìº»´Ù

Á¦

3

Àå¿¡¼­´ÂÇѱ¹¾î¼ö

½ÄÇ¥ÇöÀÇÇü½Ä¿¡´ëÇØ¼­»ìÆìº»´Ù

Á¦

4

Àå¿¡¼­´ÂÀ¯ÇÑ»óÅ¿ÀÅ丶Ÿ¸¦ÀÌ¿ëÇÑ

¼ö½ÄÇ¥ÇöÀνĽýºÅÛ¿¡´ëÇØ¼­±¸Ã¼ÀûÀ¸·Î±â¼úÇÑ´Ù

Á¦

5

Àå¿¡¼­´ÂÁ¦¾ÈµÈ½Ã

½ºÅÛÀǼº´ÉÀ»Æò°¡Çϰí,

¸¶Áö¸·À¸·ÎÁ¦

6

Àå¿¡¼­°á·Ð°úÇâÈÄ¿¬±¸¹æÇâ¿¡´ëÇØ

¼­±â¼úÇÑ´Ù

Á¦


2.1 Regular Expression (regular expression) vs. Regular Grammar (regular grammar)

Regular expression (regular expression)


Lex (lex)

- 3 -
Fig. 1 Regular expression, regular grammar, and finite-state automata

```c
< % % >
< % % >
< % % >
```
Table 1 Meta characters for patterns and their meanings

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>newline</td>
<td>\n</td>
</tr>
<tr>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>[ ]</td>
<td>Bracket</td>
</tr>
<tr>
<td>^</td>
<td>negation</td>
</tr>
<tr>
<td>$</td>
<td>0</td>
</tr>
<tr>
<td>{ }</td>
<td>disable</td>
</tr>
<tr>
<td>\</td>
<td>slash &quot;/&quot;</td>
</tr>
<tr>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>&quot;...&quot;</td>
<td>quotation mark</td>
</tr>
<tr>
<td>/</td>
<td>slash &quot;/&quot;</td>
</tr>
<tr>
<td>( )</td>
<td>special sequence</td>
</tr>
</tbody>
</table>

2.2 [1-3] Cascaded finite-state automata
2.2.1 Cascaded Analysis of Syntactic Structure (CA S S)

1 : NP -> D? A* N+ | Pron
   VP -> Md Vb | Vz | Hz Vbn | Bz Vbn | Bz | Vbg
2 : PP -> P NP
3 : SV -> NP VP
4 : S -> (Adv | PP)? SV NP? (Adv | PP)*

NP :¸í»ç±¸, VP :µ¿»ç±¸, PP :ÀüÄ¡»ç±¸,

±ÔÄ¢¾Õ¿¡ºÙ¾îÀִ°¢¼ýÀڴ±ױÔÄ¢ÀÇÀû¿ë·¹º§À»ÀǹÌÇÑ´Ù:

(longest match) [1].

1 : NP -> D? A* N+ | Pron
   VP -> Md Vb | Vz | Hz Vbn | Bz Vbn | Bz | Vbg
2 : PP -> P NP
3 : SV -> NP VP
4 : S -> (Adv | PP)? SV NP? (Adv | PP)*

NP :¸í»ç±¸, VP :µ¿»ç±¸, PP :ÀüÄ¡»ç±¸,

±ÔÄ¢¾Õ¿¡ºÙ¾îÀִ°¢¼ýÀڴ±ױÔÄ¢ÀÇÀû¿ë·¹º§À»ÀǹÌÇÑ´Ù:

(longest match) [1].

1 : NP -> D? A* N+ | Pron
   VP -> Md Vb | Vz | Hz Vbn | Bz Vbn | Bz | Vbg
2 : PP -> P NP
3 : SV -> NP VP
4 : S -> (Adv | PP)? SV NP? (Adv | PP)*

NP :¸í»ç±¸, VP :µ¿»ç±¸, PP :ÀüÄ¡»ç±¸,

±ÔÄ¢¾Õ¿¡ºÙ¾îÀִ°¢¼ýÀڴ±ױÔÄ¢ÀÇÀû¿ë·¹º§À»ÀǹÌÇÑ´Ù:

(longest match) [1].

1 : NP -> D? A* N+ | Pron
   VP -> Md Vb | Vz | Hz Vbn | Bz Vbn | Bz | Vbg
2 : PP -> P NP
3 : SV -> NP VP
4 : S -> (Adv | PP)? SV NP? (Adv | PP)*

NP :¸í»ç±¸, VP :µ¿»ç±¸, PP :ÀüÄ¡»ç±¸,

±ÔÄ¢¾Õ¿¡ºÙ¾îÀִ°¢¼ýÀڴ±ױÔÄ¢ÀÇÀû¿ë·¹º§À»ÀǹÌÇÑ´Ù:

(longest match) [1].
2.2.2 FASTUS (Finite-State Automation Text Understanding System)

Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.

1. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.
2. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.
3. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.
4. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.
5. Hobbs et al. [6] proposed a system for information extraction (FASTUS) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.

FASTUS was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.

2.2.3 Fidditch

Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.

Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks. Fidditch [13] proposed a system for unrestricted text (unrestricted text) that was the first to achieve a 100% match on the MUC (message understanding conference) tasks.
2.2.4 ¼ö½ÄÇ¥ÇöÀνÄÀ»ÀÌ¿ëÇÒ¼öÀִ½ýºÅÛÀºÅ©°ÔÁ¤º¸ÃßÃâ½Ã½ºÅÛ, ½Ã°£È¤ÀºÀº³¯Â¥¿¡´ëÇÑÁúÀÇÀÀ´ä½Ã½ºÅÛ,
±×¸®°í°¡°ÝÁ¤º¸¿¡¹Î°¨ÇÑÁõ±Ç°ü°èºñ±³¼îÇÎÀÌÀÖ´Ù.

2.3 ¼ö½ÄÇ¥ÇöÀνĽýºÅÛÀÇÀÀ¿ë

2.3.1 ¼ö½ÄÇ¥ÇöÀνÄÀ»ÀÌ¿ëÇÒ¼öÀÖ´ÂÀÚ¿¬¾îÅØ½ºÆ®ÀǾçÀÌÁõ°¡µÊ¿¡
µû¶óÇÊ¿äÇÑÁ¤º¸¸¸À»ÃßÃâÇÏ´ÂÁ¤º¸ÃßÃâ½Ã½ºÅÛÀÇÇʿ伺ÀÌÁõ´ëµÇ°íÀÖ´Ù.
Á¤º¸ÃßÃâÀ̶õƯÁ¤ºÐ¾ßÀdz»¿ëÀ»´ãÀºÀÚ¿¬¾îÅØ½ºÆ®·ÎºÎÅÍ¿øÇÏ´ÂÁ¤º¸¸¸À»
¾Æ³»¾îµ¥ÀÌÅͺ£À̽ºÈ­ÇÏ´ÂÀÛ¾÷À»¸»ÇÑ´Ù (template).
(overwrite) DB, [5, 7]. ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù [5, 7]. ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù [5, 7].
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
ÀåÇÏ´ÂÀÛ¾÷ÀÌ´Ù
2.3.2 服务取向

随着信息技术的发展，服务取向成为了现代企业的重要战略。服务取向不仅强调提供高质量的产品，更注重满足客户的需求和期望。这需要企业从产品设计、生产到销售的整个过程中，始终以客户为中心，提供个性化、差异化的服务。在服务取向的指导下，企业不仅能够提高客户满意度，还能够增强企业的竞争力。服务取向的成功实施，为企业带来了长期的发展和盈利。
3.1 

Table 2 Classification of numerals in Korean

<table>
<thead>
<tr>
<th>0, 1, 2, 3, ..., 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1, 2.3, ..., 9</td>
</tr>
</tbody>
</table>
3.2 3.2.1

(3-1) 2.7%
(3-2) 2,000,000
(3-3) 2001-3-20
(3-4) 12:30
### Table 3 Subcategory of unit bound nouns

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 13 -
3.3 ³­Â¥ (Date) ²­

3.3.1 ³­Â¥ (Date) ²­
(3-5) 1999 12 25
(3-6) 76 30
(3-7) 76
(3-8) 76
(3-9) 5
(3-10) 30
(3-11) 9
(3-12) 11 17
(3-13) 7
(3-14) 30
(3-15) 9 11
(3-16) 3

3.3.2 时间 (Time)

(3-14) 11 11 11
(3-15) 10 30
(3-16) 7 30
(3-17) 9
(3-18) 12
(3-19) 4

3.3.3 持续时间 (Duration)
(3-20) 12
(3-21) 10
(3-22) 3
(3-23) 4
(3-24) : 1994～2000
(3-25) : 14～16

3.3.4 Money (Money) (Money)

(3-26) 2
(3-27) 2,000
(3-28) 17
(3-29) 1$ 1200.36

3.3.5 Percent (Percent) (Percent) (Percent) (Percent)

(3-26) 2
(3-27) 2,000
(3-28) 17
(3-29) 1$ 1200.36
3.3.6 Measure (Measure) 

The measurements were conducted using various equipment and techniques, including ... 

3.3.7 Cardinal (Cardinal) 

The cardinality of the dataset was evaluated using various techniques and tools. ...
4.1 4.1 4.1

4.1 4.1 4.1
입력 문서

닝크(DTE), 시간(TME), 기간(DUR) 정규 표현
여정 입력 언어, 최장 구절 언어, 수사와 이존법사의 연질 가능성

금전(MNY), 비율(PCT), 측도(MSR) 정규 표현
여정 입력 언어, 최장 구절 언어, 수사와 이존법사의 연질 가능성

계수(CRD) 정규 표현
여정 입력 언어, 최장 구절 언어

출력 문서

Fig. 2 Overview of a numerical expression identifier

4.2

- 19 -
Fig. 3 A finite automata for recognizing numerals

4.3 :  

- 20-
NNN \((0-9)]+\)
NKC \(\"\"\"\"\"\"\"\"\"
NKK \(\"\"\"\"\"\"\"\"\"
NDy \(\{NNN\}\{NNN\}\{NNC\}\{SP\}?\)\"\"\"\"\"
NDm \(\{NNN\}\{SP\}?\)\"\"\"\"\"
NDd \(\{NNN\}+\)\"\"\"\"\"
NTh \(\{NNN\}\{NNK\}\)\"\"\"\"\"
NTm \(\{NNN\}\{SP\}?\)\"\"\"\"\"
NTs \(\{NNN\}\{SP\}?\)\"\"\"\"\"
NDymd \(\{TENT\ WELVE\ }\{NDy\}\{SP\}\{NDm\}\{SP\}\{NDd\}\)
NDym \{NDy\}\{SP\}\{NDm\}\{SP\}\{NDd\}\}
NDmd \{NDm\}\{SP\}\{NDd\}\}
NThms \{NTh\}\{SP\}\{NTm\}\{SP\}\{NTs\}\}
NThm \{NTh\}\{SP\}\{NTm\}\}
NTms \{NTm\}\{SP\}\{NTs\}\}

4.4 未来
4.5 (4-1) <MSR>4</MSR>
(4-2) <MSR>1</MSR>
(4-3) <MSR>1</MSR>
(4-4) <MSR>1</MSR>
(4-5) <CRD>1</CRD>
(4-6) <MSR>1</MSR>
(4-7) <MSR>1</MSR>
(4-8) <MSR>1</MSR>

4.6 (4-9) <MSR>4</MSR>
Lexical search, Rule (longest match), Rule (first given rule) 

4.7 "10" 10" "10" "10" "10" "10" "10" [18, 19].
- \( \text{(UNIT E)} \) : \text{UNIT E}

- \( \text{(UNIT K)} \) : \text{UNIT K}

- \( \text{(UNIT C)} \) : \text{UNIT C}

- \( \text{(SP)} | \text{(SYMBOL)} | \{\text{NNN} | \text{NNK}\} | \text{?} \) + \( \text{(UNIT} | \text{UNIT E} | \text{UNIT C}) \)

- \( \text{(SP)} | \text{(SYMBOL)} | \{\text{NNN} | \text{NNK}\} | \text{?} \) + \( \text{UNIT K} \)
5.1 5.1 5.1

Recall $R$ and Precision $P$. The $F$ score is defined as $F_{score} = \frac{(\beta \times P + 1)PR}{\beta \times P (P + R)}$ [22].

\begin{align*}
R &= \frac{N_R}{N_C} \quad (1) \\
\quad P &= \frac{N_R}{N_S} \quad (2) \\
F_{score} &= \frac{(\beta \times P + 1)PR}{\beta \times P (P + R)} \quad (3)
\end{align*}

SUN Sparc 10 Workstation.
5.3.1 Recall, Precision, Fscore

Table 4 System performances

<table>
<thead>
<tr>
<th>Recall</th>
<th>90.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>86.9%</td>
</tr>
<tr>
<td>Fscore</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

5.3.2

- 27-
表5 各タグのパフォーマンス

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Fscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>TME</td>
<td>98.6</td>
<td>75.0</td>
<td>85.2</td>
</tr>
<tr>
<td>DTE</td>
<td>94.4</td>
<td>98.8</td>
<td>96.5</td>
</tr>
<tr>
<td>DUR</td>
<td>74.8</td>
<td>91.7</td>
<td>82.4</td>
</tr>
<tr>
<td>MNY</td>
<td>92.7</td>
<td>90.6</td>
<td>91.7</td>
</tr>
<tr>
<td>PCT</td>
<td>97.2</td>
<td>99.5</td>
<td>98.3</td>
</tr>
<tr>
<td>MSR</td>
<td>94.1</td>
<td>89.1</td>
<td>91.5</td>
</tr>
<tr>
<td>CRD</td>
<td>73.9</td>
<td>52.6</td>
<td>61.5</td>
</tr>
</tbody>
</table>
5.4.1 오류 정의와 분석

오류 정의와 분석은 오류의 종류별로 구분하여 분석된다. 이는 오류의 원인을 파악하고 개선을 위한 근거로 사용된다.

\( (5-1) \) \(<MSR>\quad 44</MSR>\)
\( (5-2) \) \(<MSR>44</MSR>\)
\( (5-3) \) \(<MSR>44</MSR>\)
\( (5-4) \) \(<CRD>44</CRD>\)

Fig. 4 Error Analysis
5.4.2 5.4.3 5.4.4

...
5.4.5 おわりに

この文書は、情報の提供や理解を目的としています。お読みになった方には、使用条件や権利に関する注意事項を理解・遵守することをおすすめします。
(5-28) 10月 9日 7時 お楽しみ会を開催しております。
(5-29) 10月 9日 7時 お楽しみ会を開催しております。

食事後、(5-28)の後でお楽しみ会を開催いたします。お楽しみ会は(5-28)の後で開催いたします。
(5-29)の後で開催いたします。お楽しみ会は(5-29)の後で開催いたします。

...


- 36-
[24] 李小文，"多変量分析初步”，《统计研究》，10卷1期，91-100, 1996。

[25] 李小文，"关于bigram的语料库研究"，《统计与决策》，12卷1期，85-88, 2000。