한국어 문서 작성 지원 틀의 설계 및 구현

이 홍 수, 홍 순 재, 윤 지 희
한림대학교 자연과학대학 전자계산학과

Development of a System of Writing Tools for Korean Documents

Hongsu Lee, Soonjae Hong, and Jeehee Yoon
Department of Computer Science, Hallym University

요 약

한국어 문서 작성 지원 틀 「한림」은 기계 가독형의 한국어 문서 화일을 해석하여, 문서의 오류 검출을 포함한 문서 작성상의 수정 지침이 될 수 있는 자료를 사용자에게 제공하는 것을 목적으로 하는 시스템이다. 본 시스템에서는 기본적으로 문법 해석이나 사전을 이용하지 않고 알고리즘을 이용한 문자 분석만에 의하여 한국어 문서를 해석한다. 여기에서는 현재 개발중인 「한림」의 개발 목적, 방침, 구현 방법 등에 대하여 논하고, 부분적으로 완성된 몇 개의 틀을 소개한다.

Ⅰ. 서론

최근 컴퓨터 기술의 발달과 더불어 개인용 컴퓨터의 보급이 대중화되고 컴퓨터의 사용 방법 또한 다양해지고 있다. 특히 한글 워드프로세서의 개발, 보급으로 인하여 한국어 텍스트의 입력, 보존, 편집, 출력이 간단히 이루어질 수 있게 됨에 따라 컴퓨터에 의해 작성된 기계 가독형 한글 문서가 나날이 증가되고 있는 추세이다. 이와 함께 최근에는 단순한 워드프로세서의 기능을 넘어서 보다 적극적으로 문서 작성을 지원하기 위한 고기능 텍스트 처리 툴에 대한 요구가 높아지고 있다.

본 논문에서는 기계 가독형의 한국어 문서 화일을 해석하여, 문서의 오류 검출 등 사용자가 문서를 작성하는데 있어 수정 지침이 될 수 있는 정보를 사용자에게 제공할 수 있는 문서 작성 지원 틀 「한림」의 개발에 관하여 논한다.

이와 같은 문서의 작성을 적극 지원할 수 있는 문서 작성 지원 틀의 개발에 있어 영어 문서용으로 개발된 기존의 틀[1,2,3]을 참고할 수 있다. 그러나 한국어 문서는 한국어 특유의 성질을 갖고 있으므로 영어 문서를 대상으로 하는 이들 틀의 사양 및 알고리즘을 그대로 한국어용으로 변환, 사용할 수 없다. 따라서 「한림」의 개발에 있어서는 틀의 사양을 사전에 결정하기 어려운 관계로 최근 주목 받고 있는 프로토타입 기법[4]을 사용하여 틀의 개발을 진행하고 있다. 또한 알고리즘의 선정에 있어, 효율적인 틀의 개발을 위하여 한국어 텍스트의 고유의 성질을 고려한 문자열 검색 알고리즘 등 한국어 텍스트 처리를 위한 기초 연구 결과

[5,6]를 적극적으로 이용하고 있다.

여기에서는 개발중인 한국어 문서 작성 지원 툴 「한림」의 개발 목적, 방침, 구현 방법에 대하여 논하고, 현재 부분적으로 완성된 몇 개의 툴을 소개한다.

II. 개발 방침

본 문서 작성 지원 툴 「한림」은 사용자가 컴퓨터를 사용하여 문서를 작성하는 과정에 있어, 보다 좋은 문서를 작성하기 위한 수정 지침이 되는 사항을 사용자에게 제시하는 것을 목적으로 한다. 즉, 문서를 작성, 수정하는 것은 사용자로서 컴퓨터는 단지 문제가 될 수 있는 사항을 지적하여 사용자로 하여금 선택하게 한다.

해석 대상은 과학 기술 문서로 제한한다. 또한 실용성을 생각하여 일반적으로 작성하는 실용 규모의 문서(약 10,000자 정도)를 짧은 시간내에 처리하고자 한다.

문서 작성을 적극 지원하기 위해서는 단어, 문법 해석 등 자연 언어 처리 기법을 이용하는 것이 유용하다. 영어 문서용의 문서 작성 지원 툴로서는 IBM의 EPISTLE-CRITIQUE[1,2], UNIX의 Writer's Workbench[3]등을 들 수 있으며, 이들 툴에서는 자연 언어 처리 기법을 이용하여 단어, 문법, 문체 등에 관한 처리를 행하고 있다. 한국어 문서를 해석하기 위해서도 사전을 이용한 문법 해석을 행하는 것이 일반적이다. 단, 문법 해석을 행한다고 하여도 해석 결과에 모호한 점이 그대로 남아 있거나, 해석 시간이 상당히 길어질 우려가 있다. 「한림」에서는 문서 중의 문제가 될 가능성이 있는 부분을 지적하는 것만을 목적으로 하므로 문서 해석상의 정확도는 이 요구를 만족하는 정도로 충분하다. 한편 반드시 사전을 사용하거나 문법 해석 등을 행하지 않아도 문서 작성에 도움이 될 수 있는 정보를 추출할 가능성이 있다. 이와 같은 이유로 「한림」에서는 기본적으로 문법 해석이나 사전을 이용하지 않고 알고리즘을 이용한 문자 분석만에 의하여 한국어 문서를 해석한다.

III. 「한림」의 개발

3.1 프로토타입의 작성

사용자의 문서 작성을 보다 적극적으로 지원하기 위하여 툴이 제공하여야 할 정보는 무엇인가, 컴퓨터에 의해 어떠한 정보를 추출할 수 있겠는가, 컴퓨터를 사용함으로써 종래와는 다른 문서 작성 환경의 구축이 가능한가 등 이들 사항이 명확히 밝혀지지 않은 상황이다. 즉 문서 작성을 적극적으로 지원하기 위한 소프트웨어를 작성하려고 하나 그 사양이 명확치 않다. 또한, 툴의 품질을 좌우할 수 있는 사용자 인터페이스의 정의 역시 명확치 않다. 이와 같은 이유로 「한림」의 설계, 개발에는 프로토타입의 기법[4]을 사용하기로 하였다. 현재 사양이 결정 되었거나 부분적으로 완성된 프로그램은 다음과 같다.

(1) 문장의 추출

해석 대상의 문서로부터 문장을 가려내어 문장의 선두에 오는 12문자, 문장의 끝에 오는

12문자, 문장의 길이를 나열하여 문장의 출현 순서로 열거한다. 여기서 문장이란 구두점 (., ?, ! 등)으로 끝나는 것과 개행 기호에 의해 끝나는 것(표제 등)을 포함한다.

그림 1에 이 프로그램의 실행 예의 일부를 보인다. 상단의 윈도우는 해석 대상의 전체 문서를 표시하며, 하단의 윈도우는 본 프로그램에 의한 문장 추출 결과를 표시한다. 문장의 추출 결과에서는 문장의 서두 부분과 말미 부분을 맞추어 표시하고 있으므로 이 프로그램의 실행 결과를 이용하여 사용자는 문장을 시작하는 말, 끝내는 말의 단조성을 확인할 수 있다. 또한 각 문장의 끝에 각 문장의 문자수를 표시하여 긴 문장의 사용등에 대한 주의를 환기시킨다. 문장의 길이는 숫자로 나타내는 대신 20문자당 1개의 '*'의 기호를 사용하여 표시하였다. 하단의 윈도우와 상단의 윈도우는 연동하여 스크롤되며, 이를 이용하여 사용자는 현재 주목하고 있는 문장이 전체 문서 중 어디에 위치하고 있는 가를 확인, 교정할 수 있다.

또한 본 프로그램에서는 문서내의 총문자수, 문장의 수, 평균 문장의 길이 등 문서에 관한 통계적 정보를 제공한다.

그림 1. 문장의 추출 결과

(2) 문장 부호의 적절한 사용에 관한 조사

괄호의 대응 관계 및 문장 부호의 상호 관계를 조사한다. 문장 중의 괄호 등 ({ }, ( ), " ", ' ', [ ], < >)의 균형을 조사하여 오류가 있는 경우, 이를 사용자에게 제시한다. 또한 '쉼표와 마침표 등의 종지 부호는 따옴표 안에 둔다.', '줄표 앞이 완전한 문장이면 문장 부호를 붙인다.', '사물을 다 열거하지 않고 생략할 경우 줄임표 앞에는 쉼표를 찍는다.' 등의 문장 부호 규칙에 따라 이들이 적절하게 사용되었는가를 확인하여 그 결과를 제시한다. 그림 2는 본 프로그램에 의한 문서 중의 괄호 사용 검사를 행한 결과이다. 하단의 윈도우에서는 문장 단위로 괄호의 균형 검사를 행하여 오류가 있는 경우 이를 나타낸다. 상단의 윈도

우는 하단의 윈도우와 연동하여 스크롤되며, 문서 중 잘못 사용된 괄호가 있는 부분을 반전 표시한다.

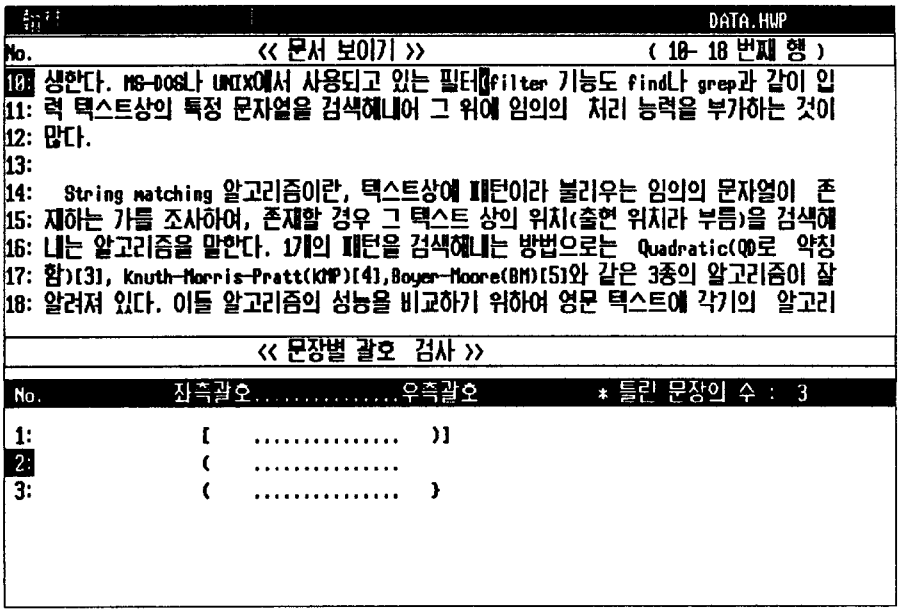

그림 2. 문장별 괄호 검사의 결과

(3) 지시대명사, 지시관형사의 검색

입력 문서 중에 출현하는 '이것', '그것', '저것', '이', '그', '저' 등의 지시대명사, 지시관형사를 검색, 제시한다. 이들은 문서 중의 비교적 가까운 범위의 지시 관계를 나타내는데 사용되지만 그것이 무엇을 가리키는지 명확하지 않은 경우가 있다. 본 프로그램에서는 이들을 검색하여 반전 표시함으로써 적절한 사용 방법에 대한 주의를 환기시키는 역할을 한다.

(4) 격조사 검사

워드프로세서를 사용하여 문서를 작성하는 경우, 처음에는 바른 조사를 사용하였으나 문서를 수정, 편집하는 과정에서 조사를 잘못 사용하는 경우가 빈번히 발생한다. 본 프로그램에서는 이와 같은 문서 작성 중 발생하기 쉬운 조사의 오용을 검색, 사용자에게 제시하여 준다. 그림 3은 격조사 중 목적격 조사의 사용에 대한 검사를 행한 결과이다. 여기에서는 문법 처리를 행하지 않고, 단순히 조사의 사용에 관한 특징 중 받침이 있는 경우와 받침이 없는 경우의 차이를 이용한 오용 검사를 행하고 있다. 하단의 윈도우는 문서 중 잘못 사용된 목적격 조사를 검색하여 반전 표시하며, 상단의 윈도우는 하단의 윈도우와 연동 스크롤되어 잘못 사용된 목적격 조사가 있는 행이 포함된 문장을 표시한다.

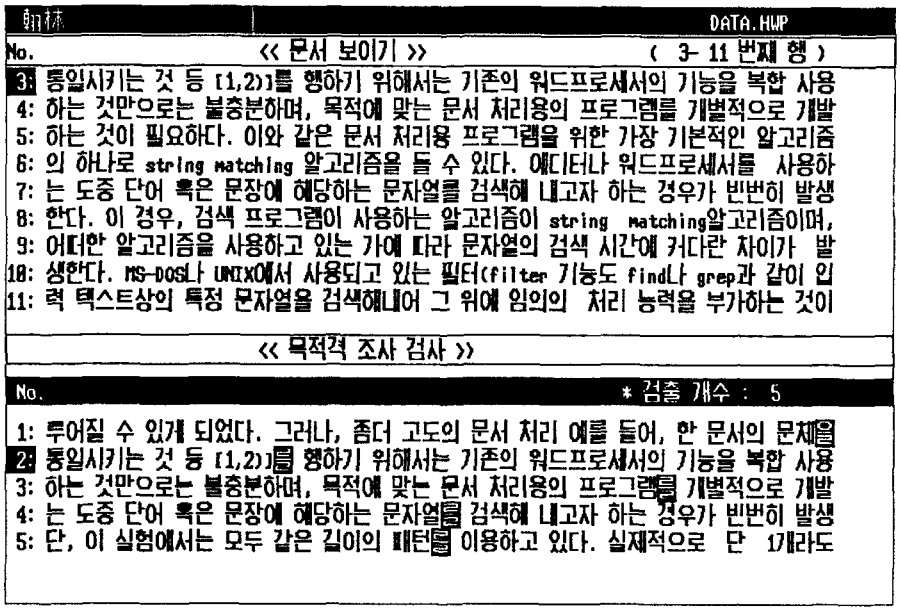

그림 3. 목적격 조사 검사의 결과

(5) 복수의 문자열 검색

사용자가 지정하거나 시스템에서 등록한 복수개의 문자열을 검색하여 그 출현 위치를 제시하여 준다. 과학 문서에 사용하기에는 부적합한 과장된 어휘나 구어체의 단어 등을 등록하여 입력 문서 중 이들의 출현을 검색하여 주는 프로그램으로 사용자의 습관 등에 의하여 잘못 사용되기 쉬운 어휘의 사용에 대한 주의를 환기시켜 준다.

(6) 외국어 검사

한글 문서 중에 포함된 외국어의 올바른 사용을 위한 정보를 제시한다. 본 프로그램에서는 문서 중의 외국어를 검색하여 이들을 알파벳순으로 정렬 표현하며 이를 이용하여 사용자는 외국어의 올바른 철자 사용에 관한 검토를 행할 수 있다. 또한 외국어가 괄호 안에 표기된 경우에는 그 앞의 단어를 함께 출력하여 외국어의 한국어 표기에 관한 주의를 환기시킨다.

(7) 시제 검사

문서 중의 시제를 검사하여 한 문장 중에 현재, 과거, 미래 등을 나타내는 어미가 섞여 존재하는 경우, 이를 사용자에게 지적하여 준다. 이를 이용하여 사용자는 각각에 대하여 그 사용이 적합한가를 검토할 수 있다.

(8) 수동형의 문장 검사

문서 중의 수동태의 형식으로 판단되는 문장을 검색하여 이를 사용자에게 제시한다. 수동태의 문장은 동작의 주체가 명확하지 않은 점 등의 특징을 갖는다. 본 프로그램에서는 문서 중 수동형의 문장을 검색, 제시하여 사용자로 하여금 이들의 사용에 관한 주의를 환기 시킨다.

3.2 시스템 구현

3.1에서 보인 프로그램의 처리 시간은 대부분 한국어 텍스트 상에서의 문자열 검색에 사용된다. 그러나 한국어 텍스트에 관해서는 한국어 특유의 성질을 살린 효율 좋은 문자열 검색 알고리즘이 개발되어 있지 않다. 이와 같은 이유로 우선 영문 텍스트상에서 유효성이 확인된 고속 문자열 검색 알고리즘[7,8]을 한국어 텍스트에 적용하기로 하였다. 그러나 한국어 텍스트는 문자의 종류가 많은 점, 1바이트의 문자와 2바이트의 문자가 섞여 존재하는 점 등의 고유의 특성을 가지고 있어 이들 알고리즘을 그대로 한국어 텍스트에 적용할 수 없다. 따라서 우리는 한국어 텍스트를 2바이트로 정규화하여 1바이트 단위의 검색을 행하는 등의 방법을 적용, 알고리즘을 일부 수정하여 이들 문제점을 해결하였다[5,6].

이들 프로그램은 IBM-PC의 MS-DOS 상에서 프로그래밍 언어 C를 사용하여 작성하였다.

Ⅳ. 결 론

본 논문에서는 한국어의 문서 작성을 지원하기 위한 문서 작성 지원 툴 「한림」의 개발에 관하여 논하였다. 여기에서는 현재 부분적으로 완성된 프로그램을 소개하였으나 앞으로 계속하여 이들 프로그램의 사양 및 프로그램을 재정비하여 발전시켜 나갈 예정이다. 특히, 격조사 검색, 시제 검사, 수동형 검사 등을 문자 분석만으로 행하기 위해서는 좀 더 많은 알고리즘의 보완이 필요하며, 필요에 따라 큰 부담이 가지 않는 한도 내에서 사전 등을 이용하는 것을 고려하고 있다. 또한 문자 분석만에 의하여 문서의 수정 작업을 지원할 수 있는 새로운 프로그램의 개발 및 사용자 인터페이스의 개선 작업을 수행하고 있다.

한편 본 논문 중의 그림 1, 2, 3은 한글 워드프로세서 훈글 1.51을 사용하여 작성한 문서를 「한림」으로 처리한 결과이다.

참고 문헌

[1] Heidorn, G. E., Jensen, L. T., Miller, L. A., Byrd, R. J. and Chodorow, M.S., "The EPISTLE Text-Critiquing System," IBM Syst. J., Vol.21, No.3, pp. 305-326, 1982.

[2] Richardson, S. D., "Enhanced Text Critiquing Using a Natural Language Parser," IBM Research Report, #51041, 1985.
[3] Gingrich, P. S., "The UNIX Writer's Workbench Software-Result of a Field Study," Bell Syst. Tech. J., Vol.62, No.6, pp. 1909-1921, 1983.
[4] "Special Issue On Rapid Prototyping ," ACM SIGSOFT Software Engineering Notes, Vol. 7, No. 5, 1982.
[5] 이영진, 윤지희, "한국어 텍스트를 위한 pattern matching 알고리즘의 개발", 한국정보과학회 '90 봄 학술 발표 논문집 Vol.17, No. 1, pp. 477-480, 1990.
[6] 윤지희, "한국어 텍스트 처리를 위한 문자열 검색 알고리즘의 개발 및 응용", 한국과학재단 최종 연구 보고서, 1990.
[7] Aho, A. V. and Corasick, M. J., "Efficient String Matching : An Aid to Bibliographic Search," Comm. ACM, Vol.18, No.6, pp. 333-340, 1975.
[8] Kowalski, G. and Meltzer, A., "New Multi-Term High Speed Text Search Algorithms," Proc. 1st Int. Conf. on Computers and Applications, pp. 514-522, 1984.