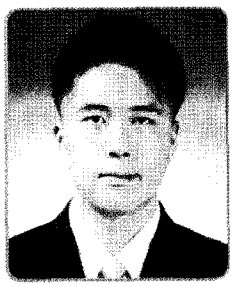
Text Mining을 이용한 영문 특허텍스트 DB의 텍스트 경제성 및 피검색성을 평가하는 기법에 관한 연구



김 현 태
화학생명공학사업팀

본 연구보고서는 Text Mining 기법을 기반으로 영문 특허텍스트 DB를 구성하는 텍스트(Text) 부분의 경제성 및 피검색성을 정량적으로 평가하는 모델을 제시하고, 이를 바탕으로 2차 가공된 영문 특허텍스트 DB의 성능을 일정범위 내에서 관리하는 품질관리모델의 개발 가능성을 탐색하는데 그 목적이 있다.

Ⅰ. 들어가며

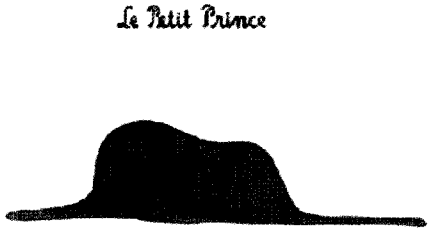
20세기 후반에 들어 산업분야별 기술의 라이프사이클이 급격히 줄어들고 있으며, 또한 인터넷, 정보통신 및 생명공학의 발달과 이종 기술 간의 융합이 활발해짐에 따라, 불과 10년 전만 해도 일반인이 상상하지 못했던 기술분야가 기업 · 연구소 · 대학 등의 연구개발(R&D) 주체를 중심으로 활발히 개척되고 있으며, 그 성과물로서 특허출원이 급증하여 2000년대 중반에는 연간 전세계 출원건수가 200만 건이 넘기에 이르렀다. 이에 따라 특허 데이터베이스의 몸집은 급격히 커지게 되었고, 대량의 특허를 한정된 시간 내에 검색해야 하는 전문적인 기술조사원(Professional Searcher)들은 특허 데이터베이스를 구성하고 있는 텍스트의 문장이 얼마나 짜임새 있고 간결하게 구성되어 있는가, 즉 『텍스트의 경제성(Text compactness)』을 중요하게 생각하기 시작했다. 왜냐하면 텍스트의 내용(Content)이 본질적으로 동일한 경우, 문법적으로 간결하게 작성된 텍스트 집합체, 즉 특허 데이터베이스를 사용하는 것이 기술조사원(Professional Searcher)이 검색된 텍스트를 읽고 분석하는데 소요되는 시간[1]을 절약해 주기 때문이다.

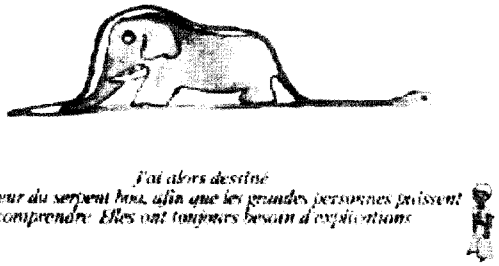
한편 『텍스트의 경제성』과 더불어 기술조사원(Professional Searcher)에게 중요한 이슈로 떠오른 다른 한 가지는, 특정 특허에 있어서 해당 출원의 기술적 요체를 나타내는 주제적 단어가 텍스트 내에서 얼마나 유효하게 재수용[2] 되고 있는가에 대한 여부이다. 즉, 주제적 단어에 대한 등가어, 유사어 혹은 유의어가 얼마나 효과적으로 텍스트에 존재하는가 하는 것이다. 연구에 따르면 인간은 비록 동일한 사물을 감각기관을 통해 수용하는 경우라 하더라도, 대뇌의 인지적 작용을 거쳐 언어로 발화할 때 그 인지적 판단의 결과물인 발화어는 개개인의 지식수준 혹은 사회문화적 환경에 따라 일치하지 않을 수 있다고 한다. 즉 같은 사물이라도 다른 표현으로 나타낼 수 있다는

1) 기술조사원(Professional Searcher)이 8시간 동안 검토(Review)할 수 있는 특허문서(abstract)가 일반적으로 200건 내외이며, 특허문서 하나의 텍스트가 250개의 단어로 이루어져 있을 경우, 200건을 검토(Review)하는데 있어서 기술조사원(Professional Searcher)이 읽어야 하는 단어는 5만개에 달한다. 그러나 만일 하나의 특허문서(Abstract)를 이루는 텍스트가 250개의 단어로 쓰인 경우와 내용(Content)적 측면에서 본질적으로 동일하지만 보다 짧은 200개의 단어로 구성되어 있는 경우, 200건을 검토하는데 있어서 기술조사원(Professional Searcher)이 읽어내려야 하는 단어의 수는 4만개로 줄어든다. 5만과 4만, 즉 20%의 시간이 절약되는 셈이다.

2) 다음 페이지의 모래시계형 발명품의 예시에 나타난 형태적 정의를 살펴보면, "원추형을 갖는 두 개의 물체가 있다. 이 원뿔체는 서로의 밑면이 평행인 상태로 원형의 접면을 이루며 맞닿아 있다."라고 기술되어 있다. 이 문장에서 "원추형의 물체"는 바로 다음 문장에서 "원뿔체"란 단어로 재수용 되고 있다.

것이다. 따라서 언어적 표현에 있어 형태의 불일치에 따른 검색실패를 회피하기 위해 『텍스트의 피검색성』 또한 중요한 이슈로 대두되었다. 아래의 예는 생텍쥐페리의 어린왕자에 나오는 보아뱀에 관한 일화로서, 개개인의 사회문화적 경험에 따라 같은 사물이라도 바라보는 시각이 얼마나 다를 수 있는지 보여주는 대표적인 경우이다.

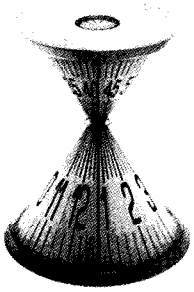


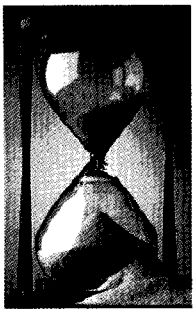


'보아뱀은 먹이를 씹지 않고 통째로 삼킨다. 그런 다음 몸을 움직일 수가 없게 되어 먹이가 소화될 때까지 여섯 달 동안 잠을 잔다. 나는 그 그림을 보고 나서 밀림의 여러 가지 모험들을 곰곰이 생각해 보았으며, 드디어는 나도 색연필을 들고 나의 첫 그림을 용케 그려 내었다. 나의 그림 제 1호, 그건 다음과 같았다. 나는 내 걸작을 어른들에게 보여주며 내 그림이 무섭지 않느냐고 물어 보았다. 어른들은 대답했다. "아니, 모자가 다 무서워?" 내 그림은 모자를 그린 것이 아니라 코끼리를 소화시키고 있는 보아뱀을 그린 것이었다. 그래서 나는 어른들이 알아볼 수 있도록 보아뱀의 속을 그렸다. 어른들에겐 항상 설명을 해 주어야 한다. 내 그림 제 2호는 아래와 같았다.'

그렇다면, 텍스트에서 주제적 단어의 유목적적 재수용, 즉 키워드의 확장이 중요한 까닭은 무엇일까? 개인적 경험차에 의한 발화어의 불일치가 과연 특허 텍스트의 생산과 소비라는 순환 사이클에서 어떤 상황을 초래할 수 있을까? 아래의 가정적 실제상황을 통해 알아보도록 하자.

A씨는 모래시계 제작자이다. 그는 모래시계의 형태적 특징인 중간이 오목한 형태는 유지하되, 모래입자의 중력낙하를 이용하지 않고, 태엽과 나사로 구동되는 시계를 만들기를 고심하던 중, 도면 1과 같은 모래시계 형태의 태엽시계를 고안하게 되었다. 일반적으로 기하학적인 도형의 이미지를 투사(透寫)하여 발명품 A의 형태적 특징을 언어로 표현한다면 "두개의 원뿔체가 서로의 밑면이 평행하도록 마주보도록 위치하며, 그 원추의 꼭지점 일부가 원형의 접면을 이루면 맞닿아 있는 형태"라고 표현할 수 있다. 그러나 A씨는 발명품 A를 원추(원뿔)형 태엽시계라는 명칭 대신에 『모래시계형 태엽시계』라 명명하여 특허출원 하였다. 왜냐하면 A씨의 전생애적인 시각적 각인(刻印)에 있어 모래시계가 매우 우월한 위치에 있었기 때문이다.





도면 1 [발명품 A] 도면 2 [모래시계] 도면 3 [장구]

하지만, 발명품 A는 상품화되지 못하고 특허권만 유지된 채로 사장되어 있었다. 그러던 중, 우연한 기회에 발명가 B는 발명품 A의 존재를 알게 되었고, 이를 그대로 모방하여 발명품 B를 출원하면서 특허 명세서 작성 시 『모래시계형 태엽시계』라는 용어 대신, 두 개의 『원추체』가 마주보며 접촉하는 형태라고 표현하였다. 과연 특허 조사원은 발명품 B에 대한 특허심사를 위해 선행기술조사를 실시할 때, 발명품 B의 출원서에 표현된 『원추』란 키워드를 사용하여 『모래시계형 태엽시계』이라고 표현된 발명품 A를 쉽게 찾아낼 수 있었을까?

위에서 『텍스트의 경제성』과 『텍스트의 피검색성』이 전문적인 기술조사원(Professional Searcher)이 대량의 데이터를 검색하는데 있어서, 시간적 효율성 그리고 검색결과의 신뢰성(Reliability)이란 측면에서 어떤 방식으로 긍정적 혹은 부정적인 영향을 줄 수 있는지를 간단히 살펴보았다.

그러면 이제 본론에서는 『텍스트의 경제성』과 『텍스트의 피검색성』을 Text Mining 기법을 이용하여 평가하는 방법에 관해 탐색해볼 것이며, 도출된 텍스트 평가기법은 영문초록 특허데이터베이스의 텍스트 품질을 일정수준으로 관리하는 하나의 지침이 될 수 있으며, 나아가 DB의 실제 사용자인 기술조사원(Professional Searcher)의 니즈(Needs)를 만족시킬 수 있는 특허 영문초록 텍스트의 작성에 대한 방향을 제시해 줄 수 있을 것이다.

한편 기술조사원(Professional Searcher)은 『텍스트의 경제성』과 『텍스트의 피검색성』이 낮은 특허 데이터베이스를 사용할 때 어떤 감정을 가질까? 아래 스포츠 기사의 마지막 구절은 위의 질문에 대한 은유적인 대답으로서 충분히 음미해볼만한 가치가 있지 않을까?

2009프로축구 K리그가 팀당 16~17경기를 치르며 반환점을 돌았다. 지난달 26일까지 한 경기 평균 관중은 1만156명. 지난해 평균 관중(1만1642명)에 비해선 1500여명 정도 줄었으나, 피부로 느끼는 K리그의 위기감은 숫자 이상이다. 무엇보다 심각한 것은 스포츠 팬 사이에서 '재미없는 K리그'라는 인식이 굳어지는 것이다. 그 해결책 중 하나가 'APT(Actual Playing Time · 실제경기시간)'를 늘리는 것이다. 작전 시간이 따로 존재하지 않는 축구의 최대 매력은 90분 내내 쉴 새 없이 몰아치는 긴박감에 있다. 자주 끊기는 축구만큼 지겨운 것도 없다. 파울이나 프리킥 · 코너킥 · 스로인 · 선수교체 등을 위해 허비하는 시간을 뺀, 실제로 플레이가 벌어지는 시간을 따지는 APT가 '재미있는 축구'의 척도가 될 수 있다. 분석 자료에 따르면, 잉글랜드 프리미어리그의 평균 APT는 63분10초, 일본 J리그는 62분48초에 달한다. 이에 비해 포항은 56분01초에 그쳤다. 즉 똑같은 90분 동안 K리그 팬들은 6~7분이나 더 멍하게 정지된 축구공을 지켜봐야 한다는 얘기다. (중략) 한준희 KBS 축구해설위원은 "데드타임이 긴 축구를 보는 기분은 액션 영화를 보러 간 관객이 영화 대부분이 멜로로 채워질 때 느끼는 배신감과 비슷한 것"이라고 말했다.

자, 이제 박진감 있는 액션영화를 만드는 방법을 알아보자.

II. 본론

서론에서는 텍스트의 품질평가에 대한 필요성에 대해 몇 가지 사례를 들어 알아보았다. 본론에서는 텍스트 품질의 객관적 평가를 구현하기 위한 방법론을 소개할 것이며, 이 방법론은 텍스트로부터 추출할 수 있는 평가요소(Estimative factor)에는 어떤 것이 있으며, 이 평가요소를 어떻게 결합할 때 품질평가산식으로서 유의미한 결과를 나타내는지 탐색해 볼 것이다. 본론의 목차는 아래와 같다.

본론에서는 먼저 텍스트는 어떻게 정의되며, 텍스트는 무엇으로 구성되어 있는지 살펴보도록 하겠다. 이를 통해, 텍스트에서 추출할 수 있는 텍스트 품질 평가요소에는 어떤 것이 있는지 알아볼 것이다.

A. Text의 정의 및 구성요소

텍스트는 언어적인 단위이자 의사소통의 단위로서, 단순한 문장의 나열이 아니라 하나의 주제에 대해 응집력을 가지고 결속된 언어의 조직망이라고 정의된다. 문법적인 관점에서 볼 때, 이러한 『텍스트』의 핵심적인 구조 단위는 『문장』이다. 『문장』은 일반적으로 마침표, 물음표 또는 느낌표 등의 완결형 구두점에 의해 명시적으로 분할될 수 있다. 이러한 『문장』은 다시 『단어』라는 세부단위로 구성되며, 영어(英語)의 경우 『단어』는 그 기능에 따라 명사/대명사/동사/형용사/부사/접속사/전치사/감탄사 『8대 품사[3]』로 분류된다. 따라서 텍스트를 구성하는 가장 작은 단위는 단어이며, 이 단어가 텍스트의 품질평가 산식이라는 일품요리의 식재료가 된다.

B. 유의미어 vs 무의미어

만약 누군가 당신에게 바람직한 검색용 텍스트란 과연 무엇인가라는 질문을 던진다면, 당신은 어떻게 대답할 수 있을까? 여러 가지 경우가 있겠지만, 그 목적이 정보의 검색을 위해 만들어진 텍스트라면, 간결한 문장으로 최대한의 정보를 다양한 키워드로 사용자에게 제공하는 텍스트라고 대답할 수 있다. 이를 다른 형식으로 표현한다면, 텍스트 내에서 정보전달의 핵심이며 독립적으로 의미를 가지고 키워드로서의 역할을 수행할 자격이 있는 단어(이하 유의미어)와, 유의미어 상호간의 문법적 연결에는 관여하지만 독립된 의미를 갖지 못하여 키워드로서 역할을 수행할 수 없는 단어(이하 무의미어)가 적절한 비율로 조합되어 있는 텍스트가 바람직한 텍스트라고 할 수 있다. 따라서 텍스트의 『바람직한』정도를 판단하기 위해 텍스트의 적절한 유의미어 대 무의미어 비율에 대한 기준 또는 범위를 정할 필요가 있다. 그리고 이를 위해서는 유의미어와 무의미어의 구분에 대한 합리적이고 일관성 있는 기준을 설정하는 과정이 반드시 선행되어야 한다.

앞 절에서 우리는 텍스트의 기본 구성단위는 단어라고 규정하였으며, 영문에서 모든 단어는 그 기능에 따라 8개의 품사로 분류된다는 사실을 확인하였다. 그렇다면 영문 텍스트의 단어를 유의미어와 무의미어로 구분하기 위한 기준을 설정함에 있어, 『품사』라는 잣대를 사용하는 것은 어떨까?

8개 품사의 특징을 살펴보면, 명사/동사는 단어 자체로서 독립적인 의미를 가지고 있으므로, 유의미어의 정의와 일치한다고 볼 수 있다. 그리고 대명사/부사/접속사/전치사/감탄사는 명사/동사의 의미를 부가하거나 문장의 연결에 관여하는 등의 역할을 하므로 무의미어의 성격에 부합한다. 따라서 본 보고서에서는 명사/동사[4]는 항상 유의미어로 취급할 것이며, 나머지 대명사/접속사/전치사/감탄사는 항상 무의미어로 분류할 것이다. 그런데 일부 형용사나 부사 중에는 텍스트의 구조와 내용을 고려하여 경우에 따라 유의미어로 혹은 무의미어로 양쪽으로 분류되는 경우가 발생할 수 있다. 이것은 문장 내에서 무의미어에 해당하는 문법적 기능, 즉 유의미어의 의미를 제한, 확장 혹은 보완하는 형용사 내지 부사의 역할을 수행하지만, 실제 유의미어에 해당하는 동사 또는 형용사에서 파생된 무의미어의 경우에 해당한다. 아래 예문을 보도록 하자.

A drawer with supporting rollers is slidingly inserted into a receiving part by guiding rails.

상기 문장에서 단어를 품사에 따라 유의미어와 무의미어로 구분하면 아래와 같다.

◈ 유의미어 : *drawer(명사) / supporting(형용사) / rollers(명사) / inserted(동사) / receiving(형용사) part(명사) / guiding(형용사) / rails(명사)*

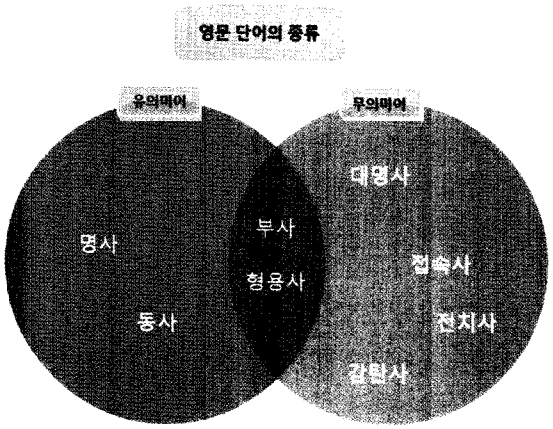
◈ 무의미어 : *with(전치사) / is(be동사) / slidingly(부사) / into(전치사) / a(관사) / by(전치사)*

위 예문에서 *"slidingly"*은 동사 slide에 ing가 결합해

3) 명사/대명사/동사/형용사/부사/접속사/전치사/감탄사/형용사 중에서 명사나 동사의 어미에 형용사형 어미가 붙어 품사가 전성된 단어는 항상 유의미어로 분류

4) 그러나 Be 동사, Have 동사, Take 동사, Get 동사 및 조동사는 항상 무의미어로 분류

형용사형으로 1차 전성된 후, 여기에 부사형 어미인 ly가 결합하여 문장에서 부사로 쓰인 경우이다. 이와 같이 전성어미의 결합으로 인해 텍스트에서 부사로 쓰이는 단어는 유의미어로 분류된다. 왜냐하면 비록 문장에서 품사는 부사일지라도 동사 "Slide"가 가진 의미를 그대로 가지고 있기 때문이다. 그러므로 본 연구에서는 문장에서 품사상 형용사 또는 부사로 분류되는 단어 중에서 형태소 분석결과 『동사』에서 파생된 『동사의 명사형』, 『동사의 형용사형』 또는 『동사의 부사형』은 비록 문법적으로는 무의미어에 해당하더라도, 실제적으로 유의미어로 취급될 것이다.



C. 텍스트의 경제성 측정

C-1. 유의미어 점유율

앞 절에서는 유의미어와 무의미어를 어떤 기준에 의해 분류할 것인가에 대해 논의하였다. 그렇다면 이러한 분류의 본원적 목적으로 다시 거슬러 올라가 보자. 과연 유의미부와 무의미부가 어느 정도의 비율로 조합되어 있는 텍스트가 바람직한 텍스트라고 할 수 있을까? 상기의 예문을 대상으로 텍스트내의 유의미어 점유율을 계산해보면, 57.1%라는 수치를 얻을 수 있다. 그러나 단문장으로 이루어진 상기 예문의 유의미어 점유율 값인 57.1%[5]가 일반적인 텍스트의 바람직한 유의미어 점유율이라고 판단하는 것은 통계적으로 무리가 있다. 따라서 보다 객관적으로 텍스트에서 유의미어가 차지하는 바람직한 비율을 알아보기 위해 다문장 텍스트를 대상으로 아래와 같은 분석을 실시하였다.

[분석예 A]

분석예 A는 하기의 조건을 모두 충족시키는 92건의 특허 중 무작위로 10건을 추출하여 실시하였다.

조건 1. IPC 주분류 또는 부분류가 C02F*인 한국특허
조건 2. 출원일자가 2008.1.1~2008.12.31인 한국특허

위의 조건으로 추출된 10건[6]의 한국특허에 대응하는 한국특허영문초록(KPA)을 키프리스(www.kipris.or.kr)에서 검색하였다. 검색된 텍스트의 유의미어 및 무의미어를 앞 절의 규칙에 따라 분류한 한 결과 총 1500 단어중 유의미어는 843단어, 무의미어는 657단어였다. 텍스트의 유의미어 점유율을 계산한 결과는 아래와 같다.

* 텍스트 유의미어 점유율
{유의미어수/(유의미어수+무의미어수)}
= 843/(657+843) = 56.2%

분석예 A에서 보듯이, 14개의 단어로 이루어진 단문장 텍스트의 유의미부 점유율은 57.1%, 1500개의 단어로 이루어진 다문장 텍스트의 유의미부 점유율은 56.2%로 거의 유사한 결과를 나타내었다. 그러므로 텍스트의 유의미부 점유율은 대략 55% 혹은 그 이상일 경우 평균적인 품질수준의 텍스트라고 가정할 수 있을 것이다. 물론 텍스트의 유의미부 점유율이 해당 텍스트의 품질(경제성 및 피검색성)을 측정하는 유효한 인자인지 나아가 텍스트의 유의미부 점유율이 높을수록 텍스트의 품질이 높은가에 대한 판단에는 보다 많은 텍스트를 대상으로 하는 분석이 이루어져야 할 것이다. 그러나 한 가지 명확한 사실은 일정 수준이상으로 단단하게 직조된 텍스트라면 일정 수준 이상의 유의미부 점유율을 반드시 나타낸다는 것이다. 따라서 텍스트 품질의 계량적 평가를 위한 평가산식을 도

5) 8 (유의미어수) / 14 (유의미어수+무의미어수)
6) KR1020090064554A KR1020080083351A KR1020080067340A KR1020080110873A KR1020080033260A KR1020080108099A KR1020080079213A KR1020080078550A KR1020090030232A KR1020087030803A

출함에 있어서, 텍스트 유의미부 점유율을 포함하는 것이 타당할 것이다.

C-2. 평균 단어반복율

앞 절에서는 텍스트의 품질을 결정하는 여러 인자 가운데 하나인 텍스트 유의미부 점유율을 검토해 보았다. 본 절에서는 예문을 통해 텍스트의 경제성 평가에 있어 텍스트 유의미부 점유율이 지닌 한계를 알아보고, 나아가 이와 같은 한계를 보완하는 인자로서평균 단어반복율에 대해 검토해 보도록 한다.

[분석예 B]

분석예 B에서는 내용적인 측면에서는 서로 같지만, 형식적인 측면에서는 서로 다르게 작성된 텍스트를 대상으로 유의미부 점유율을 분석하였으며, 밑줄이 그어진 단어가 유의미어에 해당한다.

version Ⅰ. 한국특허 1020070116273 A의 KPA 텍스트
version Ⅱ. 한국특허 1020070116273 A에 대응하는 일본 우선출원건의 PAJ 텍스트
version Ⅲ. 한국특허 1020070116273 A를 집필자가 작성한 영문초록 텍스트

Version I (PAJ)

The phenols-containing waste water treatment method comprises the following first step and second step: the first step where the alkali agent is added to the phenols-containing waste water; and the second step where the waste water with the alkali agent added thereto in the first step is distilled to thereby separate the waste liquid into which phenols are concentrated, in the case that the addition of the alkali agent is controlled in such a manner that the pH of the phenols-containing waste water reaches 10 to 14.

총 단어수 : 88개	유의미어 수 : 43개	무의미어 수 : 45개
텍스트의 유의미어 점유율 : 43 / 88 * 100 = 48.9%		

Version II (KPA)

Disclosed is a method for processing a waste water containing an oxygen-containing compound having a phenolic hydroxyl group or a carbonyl group, which comprises the following first step and second step. First step: An alkaline agent is added into a waste water containing an oxygen-containing compound having a phenolic hydroxyl group or a carbonyl group so that the waste water has a pH of not less than 10. Second step: The waste water to which the alkaline agent is added in the first step is distilled for separating a waste liquid in which the oxygen-containing compound having a phenolic hydroxyl group or a carbonyl group is concentrated.

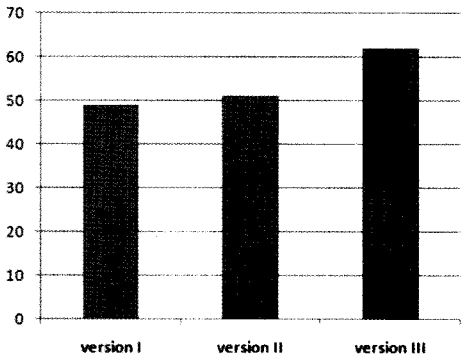
총 단어수 : 109개	유의미어 수 : 56개	무의미어 수 : 53개
텍스트의 유의미어 점유율 : 56 / 109 * 100 = 51.2%		

Version III (Customized KPA)

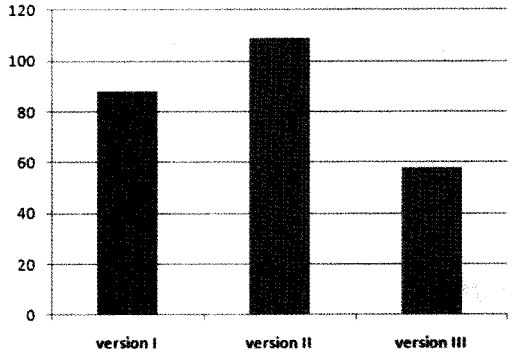
Alkali agent such as ammonia (NH4) is added to waste water that contains oxygen-based compounds having hydroxy phenol (phenolic hydroxide) or carbonyl to adjust hydrogen ion concentration of the waste liquid to higher than pH 10. After alkali addition, the waste water is vaporized to separate concentrate having hydroxy phenol or carbonyl from the waste water by distillation.

총 단어수 : 58개	유의미어 수 : 36개	무의미어 수 : 22개
텍스트의 유의미어 점유율 : 36 / 58 * 100 = 62.1%		

[그래프 1]은 위 예문의 유의미어 점유율이다. 앞 절에서 가정한 것처럼 만일 유의미어 점유율과 텍스트의 경제성이 정비례 관계라면, 텍스트 경제성 순위는 *version Ⅲ〉version Ⅱ〉version Ⅰ*일 것이다. 실제로 텍스트의 경제성을 가장 명시적으로 보여주는 텍스트 총단어수를 [표2]에서 검토해보면 유의미어 점유율이 가장 높은 *version Ⅲ*의 총 단어수가 56개로 가장 낮은 값을 보였다. 따라서 유의미어 점유율이 높을수록 텍스트의 경제성이 높아진다는 가정은 타당해 보일 수 있다.



[그래프 1] 유의미부 점유율



[그래프 2] 총 단어수

그러나 *version Ⅰ*과 *version Ⅱ*의 관계를 분석해 보면 좀 다른 해석이 가능하다. *version Ⅱ*의 총단어수는 109개로 *version Ⅰ*의 88개보다 21.6%나 높은 수치를 나타내었음에도 불구하고, *version Ⅱ*의 유의미어 점유율은 *version Ⅰ*과 거의 유사한 수치를 나타내고 있다. 이와 같은 현상은 *version Ⅰ* 및 *version Ⅱ*가 비슷한 수의 유의미어와 무의미어로 작성되어 있으나, 텍스트의 양에 있어서 *version Ⅱ*가 더 많은 단어로 구성되어 있음을 의미한다. *version Ⅰ*과 *version Ⅱ*의 내용이 실질적으로 동일한 내용을 서술하고 있고, 그렇다면 많은 단어로 작성된 것은 바람직하지 않다. 따라서 실제적인 텍스트 경제성 순위는 *version Ⅲ〉version Ⅰ〉version Ⅱ*로 볼 수 있다. 이와 같이 상기 그래프는 유의미어 점유율만으로 텍스트의 경제성을 평가하는 것이 항상 유효한 결과를 나타내는 것은 아니라는 사실을 보여준다.

그렇다면 왜 *version Ⅱ*의 텍스트 양이 다른 *version Ⅰ* 및 *version Ⅲ*보다 현저히 높을까? 이를 알아보기 위해 *version Ⅰ*, *Ⅱ*, *Ⅲ*를 구성하는 단어의 종류 및 그 종류에 따른 출현빈도에 관한 데이터를 추출해 보았으며, 그 결과는 아래의 표와 같다.

단어의 평균 반복율이 낮다는 것은 주어, 목적어 및 보어가 효율적인 유의미 조직체를 이루고 있어 문장의 불필요한 중복이나 부연설명이 적절하게 통제되어 있다는 것을 의미한다.따라서 검색용 텍스트에서 단어 평균 반복률은 일정수준까지는 낮을수록 바람직하다고 할 수 있으며, 반 복율의 관점에서 볼 때, 텍스트의 경제성은 *version Ⅲ〉version Ⅰ〉version Ⅱ*의 순서라고 할 수 있다. 물론 유의미어 점유율의 경우와 마찬가지로 단어평균 반복율이 낮을수록 텍스트의 경제성이 항상 정비례의 관계로 증가한다고는 볼 수 없지만, 일정 수준이상으로 단단하게 직조된 텍스트라면 일정 수치 이하의 단어 평균 반복율을 반드시 나타낸다. 따라서 텍스트의 경제성 판단을 위한 산식을 도출함에 있어서, 평균 단어 반복율을 포함하는 것이 타당할 것이다.

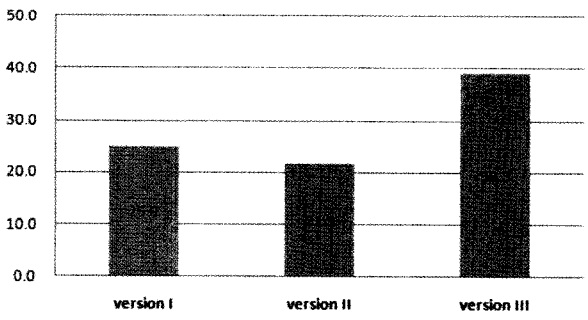
단어평균 반복율(단어종류/단어수)					
Version I		Version II		Version III	
0.43		0.39		0.64	
단어종류 : 38	단어수 : 88	단어종류 : 42	단어수 : 109	단어종류 : 37	단어수 : 58

단어평균 반복율(단어종류/단어수)											
Version I		Version II		Version III		Version I		Version II		Version III	
phenol	4	phenol	3	phenol	3	is(are)	4	is	5	is	2
contain	3	contain	5	contain	1	of	2	of	1	of	1
waste	5	waste	5	waste	4	that	2	that	1	that	1
water	4	water	4	water	4	the	15	the	6	the	3
method	1	method	1			to	3	to	1	to	4
comprise	1	comprise	1			in	3	in	2		
follow	1	follow	1			into	1	into	1		
step	5	step	5			first	3	first	3		
alkali	3	alkali	2	alkali	2	and	2	and	1		
agent	3	agent	2	agent	1	a(an)	1	a(an)	14		
add	3	add	2	add	2	second	2	second	2		
distill	1	distill	1	distill	1	which	1	which	3		
separate	1	separate	1	separate	1			have(has)	4	has(have)	2
pH	1	pH	1	pH	1			or	3	or	2
concentrate	1	concentrate	1	concentrate	2			than	1	than	1
liquid	1	liquid	1			thereby	1				
		disclose	1			thereto	1				
		process	1			where	2				
		oxygen	3	oxygen	1	with	1				
		compound	3	compound	1	such	1			such	1
		hydroxy	3	hydroxy	2					as	1
		group	6							by	1
		carbonyl	3	carbonyl	2					from	1
control	1									high	1
manner	1									after	1
case	1							for	2		
reach	1							so	1		
treat	1							less	1		
				ion	1			not	1		
				base	1						
				hydrogen	1						
				hydroxide	1						
				adjust	1						
				vaporize	1						
				ammonia	1						
				NH4	1						
유의미 단어 종류	유의미 단어수	유의미 단어 종류	유의미 단어수	유의미 단어 종류	유의미 단어수	무의미 단어 종류	무의미 단어수	무의미 단어 종류	무의미 단어수	무의미 단어 종류	무의미 단어수
21	43	23	56	23	36	17	45	19	53	14	22

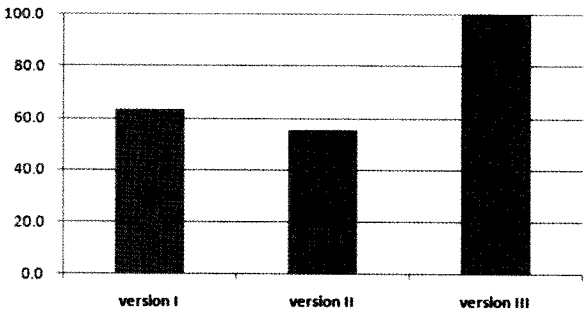
C-3. 텍스트의 경제성 평가산식

이상과 같이 앞 절에서는 영문 텍스트의 문장을 구성하는 단어를 품사에 따라 유의미어와 무의미어로 분류하였고, 여기에서 『유의미어 점유율』이라는 경제성 평가를 위한 하나의 인자(Factor)를 추출하였다. 그리고 주어, 동사 및 보어 등 문장 구성요소의 효율적인 문법적 배열을 통해 문장에서 유의미어의 단순 재등장을 얼마나 효과적으로 회피하였는지를 평가하기 위하여 『평균 단어반복율』이라는 또 하나의 인자(Factor)를 추출하였다. 그리하여 두 가지 인자(Factor)를 하나로 결합하기 위해 곱의 상관관계를 도입하였으며, 다음과 같이 『텍스트의 경제성 평가산식』을 최종 도출 하였다.

텍스트 경제성 평가산식
= 유의미어 점유율 * 평균 단어반복율



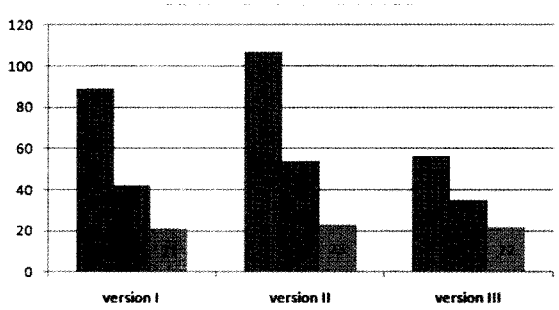
[그래프 3] 텍스트 경제성 절대값



[그래프 4] 상대적 텍스트 경제성 값

[그래프 3]은 위의 경제성 평가산식을 이용하여 도출한 각 version 별 텍스트 경제성 인덱스 수치로서, version Ⅲ가 39.3으로 가장 높았으며, version Ⅱ가 21.9로 가장 낮았다.

[그래프 4]는 각 version의 상대적 품질을 나타낸 것으로, version Ⅲ을 100점이라고 할 때, version I(PAJ version)은 63.6점, version Ⅱ(KPA version)은 불과 55.7점으로 매우 낮았다. 그리고 version I(PAJ version)을 100점이라고 할 때, version Ⅱ(KPA version)은 87.6점을 나타내었다.



[그래프 5] 총단어 수, 유의미어 수, 유의미어 종류

[그래프 5]는 version I, Ⅱ, Ⅲ의 단어수, 유의미어수 및 유의어 종류를 각각 나타내고 있다. version Ⅲ의 경우처럼 유의미어 점유율이 높고 단어평균 반복율이 낮으면 단어수, 유의미어수 및 유의어 종류의 수를 표시하는 막대그래프의 상대적 체감율이 0.6~0.65 사이에서 안정한 값을 가짐을 알 수 있다(마치 통일신라시대의 삼층석탑 같은 황금비율을 연상케 한다). 앞 절에서 검토한 바와 같이 유의미어 점유율은 텍스트내 유의미어의 중복이 다수 존재하는 경우 텍스트의 경제성 평가시 착시현상을 일으킬 가능성이 있으며, 단어평균 반복율은 텍스트내 무의미어의 비율에 따라 착시현상의 원인이 될 수 있다. 따라서 상기의 두가지 인자를 따로 분리하여 텍스트 경제성 평가에 사용하는 것은 타당하지 않다고 판단된다. 만일 텍스트의 유의미부 점유율을 높이기 위해 극단적으로 전치사/접속사/관사/부사 등의 무의미어 사용을 배제한다면 그 텍스트의 문장은 올바른 독해가 불가능한 수준이 될 수 있으며, 극단적으로는 낱말의 단순나열에 가까운 형태를 가질 수도 있다. 따라서 유의미어 점유율은 높을수록 바람직하겠지만 그 상한값이 존재한다고 보아야 할 것이다. 앞 절의 [분석예 A]에서 보듯이, 검색용 텍스트라면 최소 50% 이상, 일반적으로 55% 내외의 유의미어 점유율을 가지는 것이 바람직하며, version Ⅲ의 경우

처럼, 텍스트 구조학적인 측면에서 매우 정밀하게 구성되었다면 60% 이상의 유의미어 점유율을 나타내는 것도 가능하다. 또한 텍스트의 평균 단어반복율은 최대 250% 이내, 일반적으로 200% 내외가 바람직할 것으로 예상되나, version Ⅲ의 경우처럼 160% 내외도 가능하다. 하지만 유의미어 점유율에 상한값이 존재하는 것처럼, 평균 단어반복율에도 하한값이 존재할 것으로 보이며, 대략 150% 내외로 예상된다. 유의미어 점유율의 상한값과 평균 단어반복율의 하한값을 알아보기 위해서는 보다 많은 수의 텍스트를 분석할 필요가 있다.

D. 텍스트의 검색성 평가에 대한 고찰

앞 장에서는 『텍스트 유의미어 점유율』 및 『평균 단어 반복율』이라는 두가지 새로운 개념을 도입하였고, 이를 곱의 관계로 결합하여 텍스트 경제성 평가산식을 도출하였다. 경제성이라는 낱말이 보여주듯, 앞 장의 주된 내용은 "검색용 텍스트는 정보전달이라는 스스로의 본질에 충실하면서도 최대한 간결하고 콤팩트하게 작성돼야 한다."라는 측면에 초점이 맞추어져 있다. 즉 경제성은 텍스트 구조학적으로 문장의 경 · 박 · 단 · 소를 추구하는 축소지향적 이미지이다. 그렇다면 텍스트의 검색성은 어떤 이미지를 가지고 있을까? 아니 그 이전에 검색어란 과연 무엇일까?

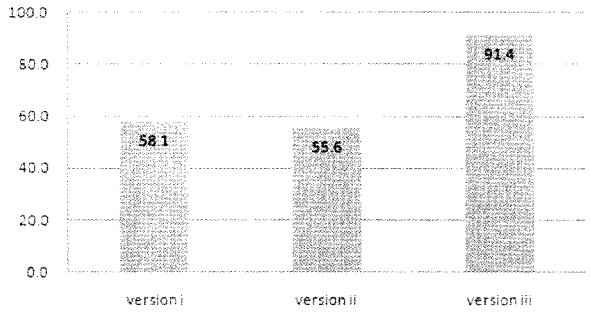
D-1. 검색어 점유율

검색(檢索語)이란 텍스트의 집합체인 데이터베이스에서 특정한 주제(Topic)를 포함하는 단위 텍스트를 찾아내는 유목적적인 행위로서, 이러한 행위의 목적을 달성하기 위해서 검색하는 자가 사용하는 단어를 검색어라고 하고, 키워드라는 용어와 동일한 의미를 가진다. 이러한 키워드의 특징은 검색하고자 하는 유형적 또는 무형적 대상의 특징 또는 본질을 표현 또는 묘사하고 있다는 점이다. 이러한 검색어는 앞 장에서 살펴본 유의미어에 대해 부분집합의 성격을 가진다.

그렇다면 유의미어 중에서 검색어로 사용할 수 있는 단어와 사용할 수 없는 단어는 어떻게 분류할 수 있을까? 원칙적으로 모든 유의미어는 검색어로 쓰일 수 있는 가능성을 가지고 있다. 하지만 특허문헌으로 그 범위를 한정할 경우 검색어로 쓰일 수 없는 유의미어의 종류는 보다 명확해 진다. version Ⅰ, Ⅱ, Ⅲ에서 사용된 유의미어 중에서 특허문헌에서 구성성분의 포함여부를 나타내는 contain, 방법 또는 프로세스 발명을 나타내는 method 또는 process, 청구항의 전제부(Transition)에서 쓰이는 comprise, 방법발명에서 단계를 나타내는 step, 불특정 제제를 뜻하는 agent, 일반적인 처리라는 뜻의 treat, 영문 특허명세서의 패턴화된 구문인 in the manner of / in case of / is characterized 등에서 쓰이는 manner, case, characterize, 화합물에서 기(基)를 나타내는 group, 불특정 화합물을 나타내는 compound, 발명의 개시를 나타내는 disclose 등은 검색어의 자격을 갖지 못한다. 따라서 이러한 검색어로서의 능력이 없는 유의미어를 『무능력 유의미어』, 검색어로서의 능력이 있는 유의미어를 『유능력 유의미어』로 명명하며, 텍스트 내의 유의미어가 점유하는 부분에서 『무능력 유의미어』의 비율이 낮을수록 해당 텍스트의 검색성은 높다고 할 수 있다.

검색어 점유율 = 유능력 유의미어수/
유의미어수(유능력 유의미어수+무능력 유의미어수)

앞장의 version Ⅰ, Ⅱ, Ⅲ의 예문을 대상으로 각각의 유능력 유의미어를 추출한 후 이의 개수를 살펴보았다. 페이지 9의 표[1]의 왼쪽인 측면인 유의미어수에서 붉은색으로 표시된 단어는 유능력 유의미어, 검은색으로 표시된 단어는 무능력 유의미어를 나타낸다.



[그래프 6] 검색어 점유율

	총 유의미어 수	유능력 유의미어 수
Version Ⅰ	43	25
Version Ⅱ	54	30
Version Ⅲ	35	32

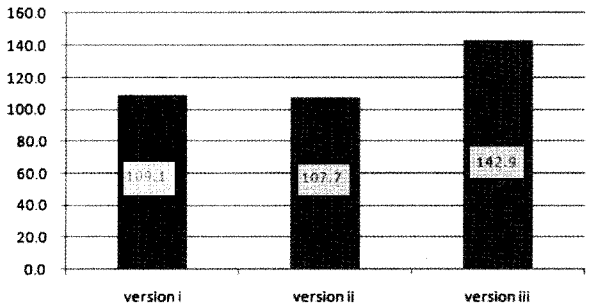
검색어 점유율 측정결과, version Ⅲ의 텍스트는 유능력 유의미어수가 90%가 넘는 수준으로 군더더기가 없이 작성된 검색용 텍스트이며, 나머지 예문의 유능력 유의미어 점유율을 55~50%사이로서 높지 않으며, version Ⅰ이 version Ⅱ보다 앞서있으나, 그 차이는 2.5%에 불과하여 그리 유의미한 값은 아닌 것으로 판단된다.

D-2. 검색어 변주율

변주(Variation)라는 단어는 음악에서 주로 사용되는 용어로서, 비교적 짧은 하나의 주제선율을 바탕으로 그 선율의 가락, 리듬, 조, 박자, 화성, 빠르기 등을 여러 모양으로 변화 및 반복시켜 모은 형식의 악곡으로 정의 할 수 있으며, 곡의 통일성을 유지하면서도 변화를 주는 효과를 준다. 이러한 변주(Variation)는 음악뿐만 아니라 검색용 텍스트를 구성함에 있어서도 적용이 가능한데, 서론에서 밝힌바 있는 단어의 재수용이란 개념과 같은 선상에 있다고 말할 수 있다. 그렇다면 음악이 아닌 텍스트에서의 단어의 변주란 어떻게 정의할 수 있으며, 그 효과는 무엇일까? 서론에서 예시한 모래시계형 발명품의 예를 다시 한번 상기해 보면(3page), 원추형의 물체를 묘사하는데 있어, 첫 문장에서는 "원추형", 그리고 두 번째 문장에서 "원뿔형"이라고 표현하였다. 텍스트 내에서 유능력 유의미어(검색어)를 재수용하는데 있어서 상기와 같이 하나 이상의 표현의 변주해 주는 것은, 검색성(Searchability)을 높여주는데 큰 영향을 준다. 왜냐하면 전문조사분석원(Professional Searche)r라 하더라도, 검색을 위해 특정 키워드를 사용함에 있어서, 모든 동의어 유사어 등가어를 알 수는 없다. 그러므로 텍스트 내에서 핵심적인 역할을 하는 유능력 유의미어(검색어)는 가능하다면 2가지 이상으로 변주해서 표현해 주는 것이 발견의 확률을 높여줄 수 있다. 따라서 검색어 변주율이 높을수록 텍스트의 피검색성은 높아진다고 할 수 있다. 검색어 변주율은 아래와 같이 정의된다.

검색어 변주율 = 검색어 종류 (형태) / 검색어 종류 (의미)

그러면 version Ⅰ, Ⅱ, Ⅲ의 예문을 대상으로 각각의 검색어 변주율(유능력 유의미어 변주율)을 살펴보자. 그 이전에 우선 예문의 텍스트에서 꼭 필요한 유능력 유의미어를 추출하는 작업을 해 보면 다음과 같다.



[그래프 7] 유능력 유의미어 변주율 (%)

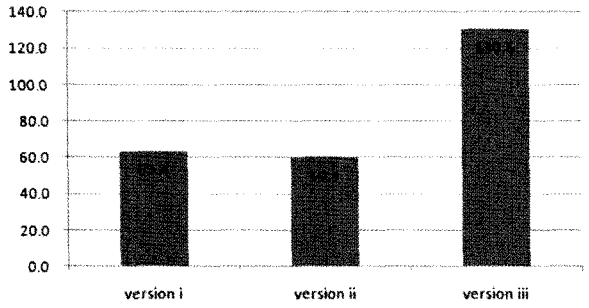
유능력 유의미어 변주율 측정결과, version Ⅲ의 텍스트가 142.9%로 매우 높은 값을 나타내었으며, version Ⅰ 및 version Ⅱ는 110% 미만으로서, 주요 키워드에 대한 변주가 거의 이루어지지 않고 있음을 알 수 있다.

D-3. 텍스트의 검색성 평가산식

앞선 절에서는 텍스트 검색성의 평가시 적용 할 수 있는 인자(Factor)로서 유능력 유의미어 점유율과 유능력 유의미어 변주율을 각각 검토해 보았다. 텍스트의 검색성을 측정하는데 있어서 상기의 두가지 인자는 서로 긴밀하게 관계를 맺고 있으나, 또한 판단의 측면이 다르므로 그 상관관계를 곱으로 연결하여 텍스트의 검색성 평가산식을 도출하는 것이 바람직하고 그 정의는 아래와 같다.

	Version I	Version II	Version III
페놀	phenol	phenol	phenol
폐(廢)	waste	waste	waste
수(水)	water / liquid	water / liquid	water / liquid
알칼리	alkali	alkali	alkali / ammonia / NH4
투입	add	add	add
수소이온농도	ph	ph	ph / hydrogen ion concentration
조절	control		adjust
증발	distill	distill	distill / vaporize
농축	concentrate	concentrate	concentrate
분리	separate	separate	separate
산소		oxygen	oxygen
히드록시기		hydroxyl	hydroxyl
카보닐기		carbonyl	carbonyl
증발	distill	distill	distill / vaporize
유능력 유의미어 종류 (형태)	12	14	20
유능력 유의미어 종류 (의미)	11	13	14
유능력 유의미어 변주율	12/11*100 = 109.1%	14/13*100 = 107.7%	20/14 = 142.9%

텍스트의 검색성 평가산식
= 검색어 점유율 * 검색어 변주율



[그래프 8] 텍스트의 피검색성 (절대값)

[그래프 8]는 version Ⅰ, Ⅱ, Ⅲ의 유능력 유의미부 점유율 및 유능력 유의미부 변주율을 각각 곱의 관계로 결합하여 계산한 텍스트의 검색성 평가결과를 나타낸다.

F. 텍스트 성능평가 산식의 도출

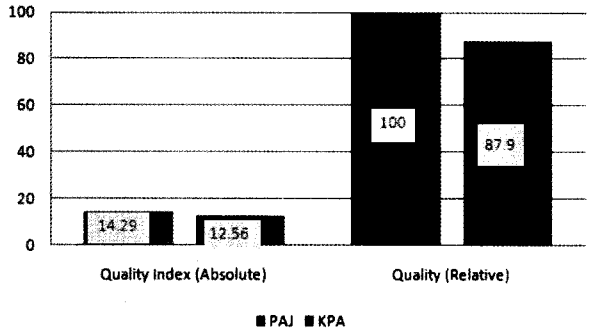
우리는 앞절에서 유의미어 및 무의미어를 기반으로 추출한 텍스트 경제성 평가산식과 텍스트 검색성 평가산식을 각각 도출하였고, 본 절에서는 각각의 산식을 곱의 관계로 연결하여 최종적인 텍스트 성능(품질)평가산식을 도출하였다.

텍스트의 성능(품질) 평가식

= 텍스트 경제성 * 텍스트 피검색성
or
= (유의미어 점유율 * 평균단어반복율)
* (검색어 점유율 * 검색어 변주율)

G. Case Study : 텍스트 성능평가 평가식을 이용한 특허 DB의 성능평가의 실례

본 Case Study는 분석예 A(6쪽)의 한국특허영문초록(KPA) 10개와, 이에 대응하는 일본특허영문초록(PAJ)을 각각 10건씩 추출하였다. 각각 10개의 단위체로 이루어진 KPA와 PAJ를 작은 단위의 KPA DB 및 PAJ DB로 간주하고, 앞선 장에서 도출한 텍스트 성능 평가산식을 이용하여 KPA DB와 PAJ DB의 텍스트 성능을 비교 분석하였다.



[그래프 9] PAJ DB & KPA DB Quality Index

[그래프 9]의 KPA와 PAJ의 성능평가 결과, Quality Index값은 KPA 12.56, PAJ 14.29이며, 이를 상대값으로 환산하면, PAJ가 100점 일때, KPA는 약 87.9점을 기록하였다. Quality Index 값을 구성하는 4개 요소를 각각 분석해 보면 [그래프 10~13]와 같다.

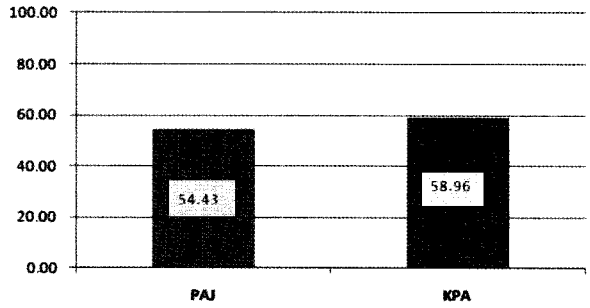
첫째, 유의미어 점유율은 KPA가 PAJ에 비해 약 4.5%로 근소한 우위를 나타냈고,

둘째, 단어평균 반복율은 PAJ가 KPA에 비해 0.02 차이로 근소한 우위를 보였으며,

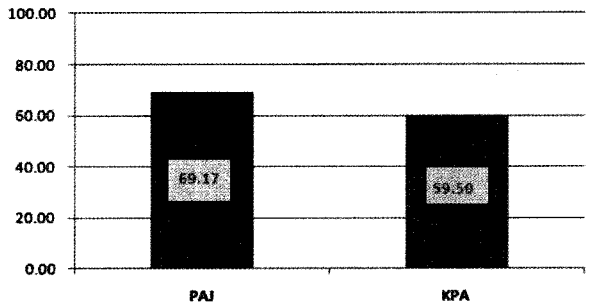
셋째, 유능력 유의미어 점유율은 PAJ가 KPA에 비해 약10% 차로 현저한 우위를 나타냈고,

넷째, 유능력 유의미어 변주율은 PAJ가 KPA에 비해 0.01 차이로 근소한 우위를 보였다.

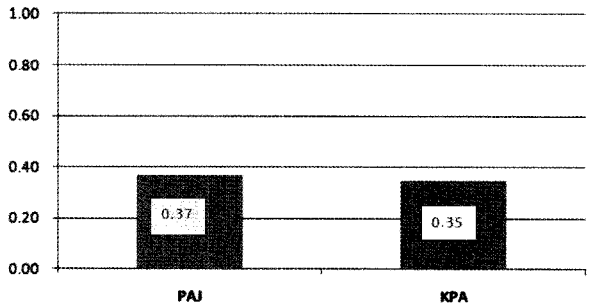
결론적으로, KPA의 Quality Index 절대값을 15점 수준까지 개선하기 위해서는, 현재처럼 높은 유의미어 점유율을 유지하되, 일정한 분량의 텍스트에서 검색어의 자격을 가질 수 있는 유능력 유의미어의 점유율을 보다 높은 수준으로 유지하도록 하는 것이 필요하다.



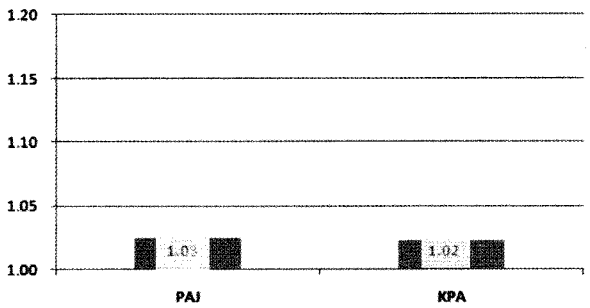
[그래프 10] 유의미어 점유율



[그래프 12] 유능력 유의미어 점유율



[그래프 11] 단어 평균 반복도



[그래프 13] 유능력 유의미어 변주율

Ⅲ. 마치며

이제까지 가공을 거친 영문 특허요약 텍스트의 품질평가는 주로 아래의 3가지 관점으로 이루어져 왔다.

첫째, 텍스트의 문법적 정확성
둘째, 텍스트와 대표도면과의 내용 합치성
셋째, 기술용어의 한영번역 적절성

이러한 종래의 품질평가 모델은 개별 특허 영문초록 텍스트의 문법적 그리고 내용적 관점에 국한된 것으로, 개별 영문초록텍스트의 집단체인 영문초록 DB에 대한 텍스트 경제성 및 검색성을 고려한 정량적인 측면에서의 품질평가 모델은 이제껏 전무하였다.

본 연구결과로 도출한 평가산식은 검색용 텍스트로 이루어진 데이터베이스의 검색성 및 간결성을 동시에 측정할 수 있기 때문에 일차적으로 텍스트로 이루어진 데이터베이스의 품질을 정량적으로 관리하는 하나의 Tool로 사용될 수 있다. 한편 각 국가에서 영문으로 구축하는 검색용 텍스트들은, 비영어권 국가인 경우 대개 자국어로 쓰여진 raw document를 영문으로 요약 가공하고, 이를 Boliven社 같은 정보업체에 일정 금액을 받고 제공하고 있는데, 이와 같은 2차 가공자료의 유통에 있어서 본 평가산식을 적용시킬수 있는 Tool을 개발하여 정보업체에게 텍스트의 성능평가에 대한 서비스를 유상으로 제공한다며, 특허정보의 유통에 있어 새로운 블루오션이 될 가능성이 있다.

효율적으로 작성된 검색용 데이터베이스의 텍스트는 생태적으로 조화로운 숲과 같다. 숲의 천이(遷移)단계에서 가장 안정된 극상림(極上林)에는 토양과 암석 등의 무기체와, 각종 동식물이 적당한 비율로 어우러져 있으며, 식물상의 분포에 있어서도 단일 수종의 인공림(人工林)과는 달리 지표의 이끼에서부터, 높이 자라나는 喬木(교목) 그리고 그 중간의 공간을 메워주는 關木(관목)에 이르기까지 다양한 수종(水種)으로 이루어져 있다. 이처럼 종(種)의 다양성이 확보된 숲은 병충해에도 쉽게 해를 입지 않으며, 각종 동물에게 비옥한 삶의 터전이 된다.

텍스트로 이루어진 검색용 데이터베이스도 마찬가지이다. 무의미어와 유의미어가 적절한 비율로 어우러져 있으며, 유의미어의 분포에 있어서도 텍스트에서 설명하고자 하는 바를 여러 가지 측면에서 설명하기 위한 다양한 유능력 유의미어 및 그의 유사어/유의어가 존재하고 있다면 이는 극상림의 숲과 같다고 할 수 있다. 효율적인 검색용 데이터베이스의 구축을 위해 본 연구가 조금이나마 도움이 되었으면 하는 바람이다.