정보검색에서의 언어모델 적용에 관한 분석*

An Analysis of the Applications of the Language Models for Information Retrieval

김 희 섭(Heesop Kim)**
정 영 미(Youngmi Jung)***

〈 목 차 〉

Ⅰ. 서론
Ⅱ. 언어모델과 정보검색
1. 언어모델의 개념
2. 언어모델과 정보검색
Ⅲ. 언어모델링 정보검색에 관한 선행연구 개관
Ⅳ. 언어모델링 정보검색의 연구동향 및 결과분석
1. 언어모델링 연구 동향분석
2. 언어모델링 정보검색 평가 실험들의 결과 분석
Ⅴ. 요약 및 결론

초 록

본 연구의 목적은 정보검색 분야에서의 언어모델의 적용에 관한 연구동향을 개관하고 이 분야의 선행연구 결과들을 분석해 보는 것이다. 선행연구들은 (1)전통적인 모델 기반 정보검색과 언어모델링 정보검색의 성능 비교 실험에 초점을 두고 있는 1세대 언어모델링 정보검색(LMIR)과 (2)기본적인 언어모델링 정보검색과 확장된 언어모델링 정보검색의 성능 비교를 통해 보다 우수한 언어모델링 확장기법을 찾아내는 것에 초점을 두고 있는 2세대 LMIR로 구분하여 분석하였다. 선행연구들의 실험결과를 분석해 본 결과 첫째, 언어모델링 정보검색은 확률모델, 벡터모델 정보검색보다 그 성능이 뛰어나고, 둘째 확장된 언어모델들은 기본적인 언어모델 정보검색보다 그 성능이 우수한 것으로 나타났다.

주제어: 언어모델, 정보검색, 통계적 언어모델

ABSTRACT

The purpose of this study is to examine the research trends and their experiment results on the applications of the language models for information retrieval. We reviewed the previous studies with the following categories: (1) the first generation of language modeling information retrieval (LMIR) experiments which are mainly focused on comparing the language modeling information retrieval with the traditional retrieval models in their retrieval performance, and (2) the second generation of LMIR experiments which are focused on comparing the expanded language modeling information retrieval with the basic language models in their retrieval performance. Through the analysis of the previous experiments results, we found that (1) language models are outperformed the probabilistic model or vector space model approaches, and (2) the expended language models demonstrated better results than the basic language models in their retrieval performance.

Key Words: Language Model(LM), Information Retrieval, Statistically Language Model(SLM)

---

* 본 논문은 2005년도 한국도서관·정보학회 하계학술발표대회(2005.6.3-6.4, 전북대학교)에서 발표된 내용을 수정한 것임.

** 경북대학교 문헌정보학과 조교수(heesop@knu.ac.kr) (제1저자)
*** 경북대학교 문헌정보학과 강사(yomjung74@hanmail.net) (공동저자)

• 접수일: 2005년 5월 20일 • 최초심사일: 2005년 5월 30일 • 최종심사일: 2005년 5월 30일

I. 서 론

정보검색의 근본적인 과제는 이용자 질의에 적합한 문헌을 찾아냄과 동시에 적합하지 않은 문헌을 배제시키는 것이다. 이 과제를 해결하기 위한 노력의 중심에 있는 것이 순위화(Ranking)이며, 순위화 결정 알고리즘에 의해 정보검색에 적용될 검색모델들이 결정된다. 전통적인 정보검색 모델인 불리안, 백터, 확률모델을 기반으로 오늘날의 퍼지, 확장불리안, 일반 벡터, 잠재의미색인, 신경망, 추론망, 신념망에 이르기까지 다양한 방법과 성능을 지닌 모델들이 지난 수년간 제안되었다. 이들 중 몇몇은 현재 여러 정보검색 분야에서 적용되어 실험되거나 사용되고 있지만 성능상의 한계점과 해결해야 할 여러 문제점들이 지적되어, 정보검색 연구자들은 새로운 모델 개발에 대한 연구에 많은 노력을 기울이고 있다.

최근 몇 년간 TREC(Text REtrieval Conference)[1]과 INEX(INitiative for the Evaluation of XML Retrieval)[2]와 같은 세계적인 정보검색 연구그룹에서 새로운 검색 모델로 화제의 중심에 있는 것이 바로 언어모델(LM: Language Model)이다. 언어모델은 과거 30여 년간 자연언어처리 분야, 특히 음성인식, 기계번역 그리고 자동교정 등의 응용분야에서 간편하고 빠른 데이터 처리를 위해 광범위하게 적용되어 왔다.

언어모델을 정보검색에 적용하기 시작한 것은 최근의 일이며, 1998년 Ponte와 Croft의 연구[3]가 그 시작이라고 할 수 있다. 이후 최근까지 언어모델과 관련된 다양한 기술들이 정보검색 분야에 꾸준하게 적용되고 다양한 측면에서 실험되어 왔다. 언어모델링 정보검색(LMIR: Language Modeling Information Retrieval)에 관한 연구는 일반적인 문헌검색을 위한 ad hoc 검색에 관한 것이 대부분이었지만, 최근에는 교차언어정보검색(CLIR: Cross Language Information Retrieval), 요약(Summmarization)과 필터링, 그리고 QA(Question Answering)와 같이 정보검색 기술 전반에 점차 확대되어 적용되고 있다. 대부분의 이들 분야의 실험들이 정보검색 성능 향상을 위한 언어모델 적용의 가능성과 시스템 성능향상의 잠재력을 확인하려고 하였다. 하지만 각 각의 연구들에서 수행된 언어모델링 방법이나 확장에 관한 기법들이 너무 다양하고 단편적으로 적용되어 실험되었기 때문에, 실제로 LMIR 시스템 구축에 어떤 언어모델링 방법과 확장기법이 적용 가능한 것인지 그리고 어떤 LMIR이 검색성능에서 효율적이었는지를 파악하기 어렵다.

따라서 본 연구에서는 최근 5-6년간 SIGIR(Special Interest Group on Information Retrieval), CIKM(Conference on Information and Knowledge Management), TREC 그리고 INEX Proceedings나 정보검색 분야 학회지들을 통해 발표된 LMIR에 관한 연구들의 망라적인 문헌조

1) TREC Home page, <http://trec.nist.gov/>
2) INEX Home page, <http://inex.is.informatik.uni-duisburg.de>
3) J. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval," *SIGIR'98, Melbourne* (1998), pp.275-281.

사를 통해 정보검색에 적용된 언어모델들과 언어모델의 확장을 위해 사용된 보조적인 기술들을 검토하고, 각 연구들에서 실험된 상대적인 성능 비교 결과를 일괄적으로 고찰하고자 한다. 또한 LMIR 연구에서 많이 적용된 언어모델들과 기술들의 동향을 파악함으로써 언어모델 기반 정보검색 시스템 구축을 위한 이론적 토대를 마련하고자 한다.

본 연구에서는 TREC이나 INEX에서 제공된 테스트 컬렉션(Test Collection)을 통한 실험적 연구들 중에서 텍스트 기반의 ad hoc 검색을 다룬 연구들만을 문헌조사 대상으로 제한하였다.

Ⅱ. 언어모델과 정보검색

1. 언어모델 개념

언어 모델링(Language Modeling)은 자연어 안에서 문법, 구문, 단어 등에 대한 어떤 규칙성을 찾아내고 그 규칙성을 이용하기 위한 노력이다. 이런 방법을 통해 얻어진 언어모델(LM: Language Model)은 오랫동안 음성 인식이나 기계 번역, 문자 인식, 철자 교정 등 다양한 응용분야에서 시스템의 정확도를 높이고 수행 시간을 줄이는 데 유용한 방법으로 각광을 받아왔다.

언어모델은 크게 지식 기반 모델(Knowledge-based Model)과 통계적 모델(Statistical Model)로 나눌 수 있다. 지식 기반 모델은 정규 문법(RG: Regular Grammar)나 문맥 자유 문법(CFG: Context-Free Grammar)을 만들고, 이러한 문법 구조에 어긋난 구조를 탐색 공간에서 제거함으로써 탐색 범위를 줄이고 인식률을 높이는 방식이다.[4] 그러나 지식 기반 모델은 문법 구조를 만들기가 까다롭고 대용량의 어휘를 수용하기 어려울 뿐만 아니라 언어의 비문법성에 의해 규칙정의가 어렵기 때문에 범용 언어 모델이나 새로운 영역에 대한 언어 모델을 구성할 때 많은 시간과 노력을 요구하게 된다. 따라서 이 모델링 방법은 주로 특정적이고 협소한 분야의 자연언어처리 분야에서 일부 사용되고 있을 뿐 대규모의 데이터를 처리해야 하는 분야에서는 적용되기 어려운 접근법이다.

이에 반해 통계적 모델은 대량의 말뭉치(Corpus)에서 언어 규칙을 확률로 나타내고 확률값을 통해서 탐색 영역을 제한하는 방법이다. 그래서 통계적 언어모델은 음성인식에서 정확성 뿐만 아니라 탐색 공간을 급격히 줄이는 효과를 보여준다.[5] 통계적 언어모델의 목적은 주어진 인식영역에 맞는 단어열 s의 확률을 예측하는 것으로 단어열 s는 w_1, w_2, ..., w_i으로 이루어진 단어열이라고

4) 이진석, 박재득, 이근배, "K-SLM Toolkit을 이용한 한국어의 통계적 언어 모델링 비교," 제11회 한글 및 한국어정보처리 학술대회 논문발표집(1999), <http://nlp.postech.ac.kr/lab_papers/9910_h%26h_wolfpack.doc> [cited 2005. 4. 12]

5) R. Lyer and M. Ostendorf, "Relevance Weighting for Combining Multi-domain Data for N-gram Language Modeling," *Computer Speech and Language*, Vol.13(1999), p.281.

가정하면 단어열 s의 확률 P(s)는 식 (1)과 같다.[6]

$$p(w_1, w_2, ..., w_i) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1w_2) \ldots p(w_i \mid w_{1\ldots}w_{i-1}) \qquad (1)$$

그러나 일반적인 통계적 언어모델링은 어떤 단어열에서 i번째 단어 w_i가 나타날 확률을 구하기 위해 w_1에서 w_{i-1}까지 i-1개의 단어열이 나타날 확률을 구해야 하는 번거로움이 있다. 그래서 단어의 확률은 이전 단어에 의존적이라는 Markov 가정에 의해 i번째 단어가 나타날 확률을 구하기 위해 이전의 N-1개의 단어의 확률만을 적용하는 N-Gram 모델이 일반적이다.[7] 특정한 문장 s에 대한 N-gram 모델은 식 (2)와 같다.[8]

$$P(s) \cong P(w_1, w_2, ..w_n) = \prod_{i=1}^{n} P(w_i | w_{i-m+1}, ..., w_{i-1}) \qquad (2)$$

w_i는 문장에서 i 번째 단어이고 m은 연관되어 사용된 단어 수를 정의한다. bigram 모델 $(m = 2)$은 앞의 단어에 이어 뒤의 단어가 나타날 확률을 사용해서 측정한다. 반면 unigram 모델$(m = 1)$은 오직 개별 단어의 확률로 측정한다. 음성인식과 기계번역과 같은 응용분야에서는 단어순서가 중요하기 때문에 일반적으로 대용량의 학습데이터가 가능하다면 trigram$(m = 3)$ 모델이, 소량의 학습데이터가 가능하다면 bigram이 사용되어 왔다. 정보검색에서 단어순서의 역할은 덜 명백하다고 생각되었고 이에 따라 가장 친숙하고 기본적인 언어모델인 unigram 모델이 광범위하게 사용되고 있다.[9]

통계적 언어모델이 모두 그렇듯이 N-Gram 통계언어 모델링은 모든 가능한 문장의 확률적 분포를 추정하기 위해 언어 모델을 사용하기 때문에, 우선적으로 전체 컬렉션을 대표할 만한 샘플 데이터를 선정하고 그 데이터를 학습함으로써 언어모델을 측정하는 것이 선행되어야 한다. 따라서 통계적 언어모델의 좋은 성능을 위해서는 대용량의 학습 데이터가 필요하고, 만약 빈약한 학습데이터에 의해 빈도수가 '0'인 단어가 질의어에 포함될 경우 전체 질의어의 확률분포 값이 '0'이 심각한 오류가 발생하게 된다. 즉 미등록어(OOV: Out-Of-Vocabulary) 발생시에는 그 대처가 불가능하다는 것이 통계적 언어모델링의 가장 큰 단점이다. 하지만 지난 30여 년 동안 언어모델링의 다양한

6) 박선희, 노용완, 홍광성, "문장음성인식을 위한 VCCV 기반의 언어모델과 Smoothing 기법 평가," 정보처리학회 논문지B, 제11-B권, 제2호(2004, 4), p.242.

7) 이진석, 박재득, 이근배, 전게논문, p.2.

8) R. Rosenfeld, "Two Decades of Statistical Language Modeling : Where Do We Go From Here?" *In Proceeding of the IEEE*, Vol.88, No.8(2000), p.1274.

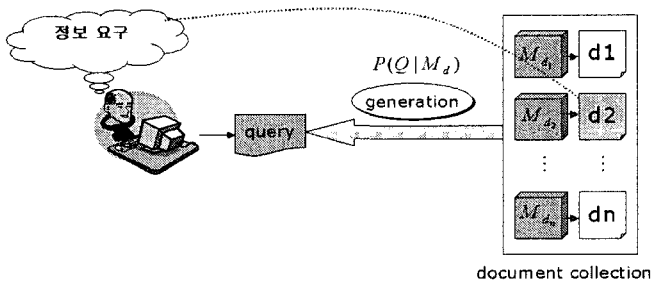
9) W. B. Croft, J. Callan, J. Lafferty, *Workshop on Language Modeling and Information Retrieval*, (Carnegie Mellon University, Pittsburgh, Pennsylvania, 2001), p.4.

확장 기법에 대한 연구들을 통해 미등록어 발생의 단점들을 보완함으로써 음성인식, 기계 번역, 자동 교정 등과 같은 대부분의 언어모델링 응용분야에서 성공적인 결과를 보이고 있다. 이중에 통계적 언어모델인 N-gram이 가장 성공적이고 간편한 언어모델링 방법으로 알려져 있다.[10] 다음 장에서는 언어모델이 정보검색 분야에서는 어떻게 적용되는지에 대하여 간략하게 살펴보고자 한다.

2. 언어모델과 정보검색

1998년의 Ponte와 Croft의 연구 이후, 언어모델링 연구 분야에서 정보검색은 음성인식, 기계번역과 같은 주요한 응용분야의 하나가 되었고 정보검색 연구 분야에서 언어모델은 주목해야할 검색모델의 하나로 자리 잡고 있다. 언어모델 적용은 여러 가지 이유에서 이점을 지닌다. 예를 들면, 언어모델을 사용하여 정보검색시스템을 구축하는 것은 시스템 설계와 실험 성과에 대한 이유를 확률적인 방법을 사용하여 명확한 수치 데이터로 제시해 줄 수 있다는 것이다. 또 다른 큰 이점은 지난 30여 년 동안 자연언어처리 분야에서 실험되어 축적되어온 다양한 언어모델과 완화기법의 결합에 대한 연구결과물들을 응용할 수 있다는 것이다.[11]

정보검색에서 언어모델 적용의 기본적인 가정은 이용자가 생각하고 있는 하나의 '이상적인' 문헌이 존재하고, 이용자는 그 문헌을 요구하기 위해 질의어 텍스트로 표현한다는 것이다. 이와 같은 이용자의 질의생성과정을 모델링 하는 것이 전통적인 LMIR의 기본 개념이 된다. 즉, 문헌으로부터 질의를 생성한다는 것이다.[12] 다시 말해 LMIR은 컬렉션 내의 각각의 문헌 d를 위한 언어모델 M_d를 구축한다는 것이고, 그리고 어떻게 이런 각각의 문헌 모델로부터 질의 Q가 생성될 수 있는가에 따라 문헌들을 순위화하는 것이다. 이것은 $P(Q|M_d)$로 표현된다.



〈그림 1〉 언어모델 기반 정보검색 개념[13]

10) W. B. Croft, "Language Models for Information Retrieval," *In Proceedings of the 19th International Conference on Data Engineering*(2003), p.4.

11) *Ibid.*, p.4.

12) K. Sparck Jones et al., "Language Modeling and Relevance," In W. B. Croft and J. Lafferty., eds, *Language Modeling for Information Retrieval*(London : Kluwer Academic Publishers, 2003), pp.57-71.

정보검색을 위한 언어모델들은 $P(Q|M_d)$을 어떻게 정의하고 계산하느냐에 따라 그 접근방법이 달라진다.

Ponte와 Croft[14]는 질의 Q을 단일어로 취급하고 $P(Q|M_d)$의 근사치를 구하기 위해 질의어를 생성해낼 확률과 질의어에 포함되지 않은 단어를 생성하지 않을 확률의 곱으로 계산했다. 질의에서 같은 단어의 중복 출현은 고려되지 않았고 이를 위해 제시된 식(3)은 다음과 같다.

$$P(Q|M_d) = \prod_{w \in Q} P(w|M_d) \prod_{w \notin Q} (1.0 - P(w|M_d) \qquad (3)$$

여기에서 $P(w|M_d)$는 비모수적 방법에 의해 계산된 것으로, w와 위험요인을 포함한 문헌에서 w의 평균 확률을 사용하였고, 전혀 출현하지 않은 단어를 위해서는 컬렉션에서 w의 전체 확률을 대신 사용하였다. Ponte와 Croft의 모델은 정보검색분야에 적용된 초기 언어모델로 최근의 많은 연구들에서 기본적인 또는 전통적인 언어모델로 간주되고 있다.

전통적인 LMIR은 확률적인 계산방법에 의한 전형적인 수학모델을 따르기 때문에 비교적 단순하고 이해하기 쉬운 체제이다. 또한 LMIR은 각 문헌의 모델링 과정이 곧 색인작성 과정이 되며 이렇게 측정된 문헌모델에서 질의 생성확률을 추정하는 것으로 색인작성과 검색모델이 통합된 접근법이라는 장점을 가진다. 하지만 LMIR이 대량의 훈련 데이터(Training Data)가 필요하다는 점, 적합성 피드백(Relevance Feedback)이나 질의 확장(Query Expansion)을 직접적으로 적용하기 어렵다는 점, 또한 불리언 연산자나 구, 구조적인 질의 처리가 어렵다는 단점을 지니고 있다. 이런 단점들을 해결하고 보다 나은 성능을 위해 다양한 언어모델링 방법들이 계속 시도되었고, 제안된 언어모델을 확장하기 위해 자연언어처리 분야에서 개발되어온 여러 확장기법들이 정보검색 분야에 적용되어 실험되고 있다.

다음 장에서는 정보검색에 적용된 언어모델링 방법들과 언어모델 확장기법들을 고찰하기 위하여 LMIR 분야에서 수행되어온 선행연구를 개관하였다.

Ⅲ. 언어모델링 정보검색에 관한 선행연구 개관

본 장에서는 언어모델링 정보검색에 관한 선행연구를 연대기별 순으로 나누어 단순히 각각의

13) V. Lavrenko, C. Zhai, "Text Retrieval and Mining," 〈http://www.stanford.edu/class/cs276a/handouts/lecture12.ppt〉 [cited 2005. 5. 22]

14) J. Ponte and W. B. Croft, *ibid.*, pp.277-278.

특징에 대해서만 정리하고자 한다.

1998년의 Ponte와 Croft[15]의 연구를 시작으로 ad hoc 정보검색에 언어모델이 적용되어왔다. 이들은 정보검색 분야에 적절한 색인작성 모델이 부족한 많은 이유들에 대해 논의하였고, 적절한 색인작성 모델로 확률적 언어모델링에 기반을 둔 검색 접근을 제안하였다. 여기에서 언어모델 접근은 문헌의 색인작성과 문헌의 검색을 하나의 모델로 통합하여 제시하였고, 실험을 통해 표준적인 tf · idf(단어빈도수 · 역문헌빈도수) 가중치를 사용한 시스템보다 현저하게 성능이 우수하다는 결론을 내렸다.

이와 비슷한 시기에 연구를 수행한 Hiemstra[16]는 샘플 데이터에서 측정된 문헌모델에 말뭉치(Corpus) 모델을 사용하여 보간하는 방법을 제안했고 말뭉치에서 단어의 확률을 측정하기 위해서 tf · idf를 사용하였다. Hiemstra 모델의 실험 결과는 코사인 함수와 tf · idf를 사용한 전통적인 벡터모델 기반 검색시스템보다 좋은 성능을 보인 것으로 나타났다.

Song과 Croft[17]의 연구에서 제시된 언어모델은 Ponte와 Croft 모델 그리고 Hiemstra 모델과 기본적인 개념은 같지만 언어모델 적용의 가장 큰 장애물인 빈약한 데이터 문제를 해결하기 위해 Good-Turing과 Curve-Fitting을 통해 문헌모델을 완화하였다. 또한 말뭉치 모델을 통해 문헌모델을 확장하고 선형 보간법(Linear Interpolation)을 통해 unigram과 bigram 모델을 결합한 언어모델을 제안하였다. 월 스트리트 저널(Wall Street Journal)과 TREC4의 데이터를 테스트 컬렉션(Test Collection)으로 사용하여 시스템의 성능을 평가한 결과, INQUERY 시스템의 성능에 필적하고 Ponte와 Croft의 모델보다 성능이 우수한 것으로 나타났다.

2001년 ARDA(Advanced Research and Development Activity in Information Technology)의 후원으로 메사추세츠 대학(University of Massachusetts)과 카네기 멜론 대학(Carnegie-Mellon University)을 중심으로 조직된 LMIR을 위한 워크샵[18]이 카네기 멜론 대학에서 개최되었다. 이 워크샵에는 5개의 나라들로부터 32명의 연구자들이 참여하였고 20개의 연구에 대한 발표가 있었는데 LMIR에서의 적합성 이론, 다른 확률적 검색 모델들과의 관계, 완화기법의 상대적인 성능, 문헌모델을 사용한 완화의 이점과 말뭉치 기반 질의 모델의 비교, unigram 기반 확률에 더 많은 정보와 결합하기 위한 기술 등에 관한 ad hoc 검색과 자동 요약과 카테고리 작성과 같은 정보시스템 전반에 걸쳐 이 분야의 광범위한 주제가 논의되었다. 이 워크샵을 기폭제로 LMIR 분야의 많은 연구들이 활발하게 진행되었다.

---

15) J. Ponte and W. B. Croft, *ibid.*, pp.277-278.

16) D. Hiemstra, "A Linguistically Motivated Probabilistic Model of Information Retrieval," *Second European Conference on Digital Libraries*(1998), pp.569-584.

17) F. Song and W. B. Croft, "A General Language Model for Information Retrieval," *CIKM'99*, Kansas City, Mo(1999), pp.316-321.

18) W. B. Croft, J. Callan, J. Lafferty, *ibid.*, pp.4-6.

Zhai와 Lafferty[19]는 질의 언어의 확장을 위한 언어모델기반 문헌 피드백과 일반적인 확률기반 문헌 피드백의 두 가지 접근에 대해 제안하고 평가했는데, 실험결과 두 가지의 접근 모두 Rocchio 피드백 접근보다 성능이 우수하고 효과적인 것으로 밝혀졌다.

Zhai와 Lafferty[20]는 또 다른 논문에서 베이시안 결정 이론(Bayesian Decision Theory)을 바탕으로 확률적 순위화를 사용하여 문헌모델과 질의모델을 결합한 프레임웍을 제시했는데, 이 프레임웍에 사용된 검색모델로서 언어모델을 사용하였다. 이 새로운 접근법은 Rocchio를 사용한 질의확장을 가지는 벡터 모델과 기본적인 언어모델 보다 그 성능이 우수한 것으로 나타났다.

Lavrenko와 Croft[21]는 전통적인 확률모델과 언어모델 정보검색 접근의 차이를 설명했고, 문헌모델을 기반으로 한 전통적인 언어모델들과는 달리 오직 질의만을 사용하여 생성한 적합성 모델(Relevance Model)을 제시하였다. 이 모델을 사용한 시스템은 TREC을 통한 실험에서 전통적인 언어모델 기반 시스템보다 그 성능이 우수한 것으로 나타났다.

Zin과 Hauptmann[22]은 새로운 언어모델로 'Title Language Model'을 제안했는데, 이것은 정보검색에 사용된 전통적인 언어모델과 달리 문헌 D의 제목으로써 질의 Q를 사용하는 확률에 의해 조건부확률 P(Q|D)를 정의한 것으로 두 가지의 새로운 완화(Smoothing)기법이 사용되었다. 네 가지의 다른 TREC 문헌집합을 사용하여 실험한 결과, 이 모델은 전통적인 언어모델과 벡터모델보다 그 성능이 우수한 것으로 나타났다.

또한 이 연구그룹의 Jin 외 3인[23]은 문헌들의 카테고리 정보를 이용한 확장된 LMIR을 제안했고, TREC4를 통한 실험 결과 전통적인 언어모델보다 그 성능이 우수하다는 결론을 내렸다.

Zhai와 Lafferty[24]는 두 단계의 완화기법 사용을 제안했는데, 먼저 문헌모델이 Dirichlet 매개변수를 사용하여 완화되고 두 번째 단계에서는 Dirichlet 매개변수를 통해 완화된 문헌모델이 질의모델을 기반으로 다시 보간되는 방법을 사용하였다. 5종의 다른 데이터베이스와 네 개의 질의 유형을 사용하여 평가한 결과, 두 단계의 완화기법을 사용한 정보검색은 평균 정확률에서 단일 완화기법을 사용한 방법에서 얻은 최고의 결과에 근접하게 나타난 것으로 밝혀졌다.

---

19) C. Zhai and J. Lafferty, "Model-based Feedback in the Language Modeling Approach to Information Retrieval," *CIKM'01*, Atlanta, Georgia(2001), pp.403-410.

20) C. Zhai and J. Lafferty, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval," *SIGIR'01*, New Orleans, Louisiana(2001), pp.111-119.

21) V. Lavrenko and W. B. Croft, "Relevance-based Language Models," In W. B. Croft, D. Harper, D. H. Kraft, J. Zobel eds., *SIGIR'01*, New Orleans, Louisiana(2001), pp.123-125.

22) R. Jin and A. G. Hauptmann, "Title Language Model for Information Retrieval," *SIGIR'02*, Tempere (2002), pp.42-48.

23) R. Jin et al., "Language Model for IR Using Collection Information," *SIGIR'02*, Tampere(2002), pp.419-420.

24) C. Zhai and J. Lafferty, "Two-Stage Language Models for Information Retrieval," *SIGIR'02*, Tampere (2002), pp.49-56.

Srikanth와 Srikanth[25]은 정보검색을 위해 두 단어들 간의 의존성을 표현하면서 연속하는 두 단어가 순서에 상관없이 나타날 확률을 기반으로 bi-term 언어모델을 제시하였다. 실험결과 근소한 차이지만 unigram 언어모델보다 bi-term 언어모델이 그 성능에서 우수한 것으로 나타났다.

Croft[26]는 언어모델 접근의 주요 변수에 대한 간단한 리뷰와 언어모델이 검색관련 언어기술의 개발에 어떻게 사용되는지를 논하고, 텍스트로부터 추출된 구조화된 데이터에 이런 접근이 사용될 수 있는지에 대하여 언급하였다.

Zaragoza, Hiemstra 그리고 Tipping[27]은 ad hoc 정보검색을 위한 언어모델에서 베이지안 확장을 제안했다. ad hoc 언어모델은 질의모델의 다항식에 확장된 추정량이 베이지안 예측치를 위해 사용되었다. 실험결과 제안된 모델이 Bayes-smoothing보다 성능이 우수했고, 베이지안 확장과 선형 보간 완화기법을 결합한 것이 다른 모든 추정량보다 성능이 우수한 것으로 결론을 내렸다.

Zhai와 Lafferty[28]는 각 문헌에 대한 언어모델을 측정하고, 측정된 언어모델에 따라 질의의 확률에 의해 문헌을 순위화하는 것을 전제로, 언어모델 측정시 완화기법의 사용여부가 검색시스템의 성능 향상에 영향을 미친다는 결론과 함께 다양한 완화기법을 적용한 시스템의 성능을 비교하였다. 또한 이 논문에서는 완화 매개변수의 자동생성 방법을 제안하였다.

Metzler와 Croft[29]는 언어모델링과 추론망 접근을 단일 프레임워크로 결합한 새로운 모델을 제시했다. 이 모델은 언어모델링 추정을 사용하여 평가를 위한 구조화된 질의어처리를 수용했다. 실험을 통해 제시된 새로운 모델, 추론망 검색 모델, 그리고 질의우도 기반 언어모델의 성능을 평가한 결과, 단일 프레임워크로 결합한 새로운 모델이 추론망 검색 모델 기반의 INQUERY 검색엔진보다 더 높은 정확률을 보인 것으로 나타났다.

Gao et al.[30]은 unigram에 기반한 기본적인 언어모델링 접근을 확장할 목적으로 질의속에 나타나는 단어들간의 의존성을 표현하는 새로운 의존 언어모델(Dependence LM)을 제시했다. 기존의 bigram 이나 bi-term 기반 언어모델들이 인접하는 단어들간의 의존성만 나타낸 것이라면 여기에

25) M. Srinkanth and R. Srinkanth, "Bi-term Language Models for Document Retrieval," *SIGIR'02*, Tempere (2002), pp.425-426.

26) W. B. Croft, "Language Models for Information Retrieval," *Proceedings of the 19th International Conference on Data Engineering*(2003).

27) H. Zaragiza, D. Hiemstra, M. Tipping, "Bayesian Extension to the Language Model for Ad Hoc Information Retrieval," *SIGIR'03*, Toronto(2003), pp.4-9.

28) C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Transactions on Information Systems*, Vol.22, No.2 (April 2004), pp.179-214.

29) D. Metzler and W. B. Corft, "Combining the Language Model and Inference Network Approaches to Retrieval," *Information Processing & Management*, Vol.40(2004), pp.735-750.

30) J. Gao et al. "Dependence Language Model for Information Retrieval," *SIGIR'04*, Sheffield, South Yorkshire (2004), pp.170-177.

서 제시된 모델은 질의속의 단어들이 인접하지 않더라도 두 단어간에 존재하는 의존성을 표현한 언어모델이다. 전통적인 확률모델을 기반으로 한 Okapi 시스템과 unigram 언어모델, bigram 언어모델, bi-term언어모델, 그리고 의존 언어모델 기반 시스템의 각각에 대한 성능을 TREC을 통해 비교·실험한 결과 의존 언어모델의 성능이 확률모델이나 unigram 언어모델의 그것보다 상당히 우수하게 나타났다. 실제의 정보검색에서 bigram 언어모델 적용의 한계를 극복하기 위한 대안으로, bi-term 언어모델이 제기되나 그 성능이 탁월한 것은 아니었기 때문에 의존 언어모델을 제안한 것으로 밝혔다.

지금까지 개관한 언어모델링 정보검색에 대한 연구들이 SIGIR이나 TREC을 중심으로 한 비구조화된 문헌의 ad hoc 검색에 관한 것이라면, 아래에서 언급될 연구들은 최근에 INEX를 통해 구조화된 XML 문헌 검색을 위한 언어모델 적용에 관한 개관이다.

List et al.[31]은 Tijah라 불리는 XML 검색시스템을 INEX 2003을 통해 그 성능을 평가하였다. 이 검색모델은 단순한 전통적인 언어모델로 각 계층 요소들간의 관계에 대한 값으로 언어모델을 측정하지 않고, 각 요소들이 속한 텍스트와 각 요소의 자식 노드(Child Node)를 사용하여 각 구성요소의 언어모델을 측정하였다. 이 모델은 INEX 2003을 통해 수행된 다른 검색모델 기반 XML 검색시스템과의 성능비교에서 상위권을 차지함으로써 그 우수성을 입증하였다.

Ogilvie와 Callan[32]의 연구는 INEX 2003을 통해 단순한 언어모델 기반 XML 검색 시스템을 평가한 결과 그 성능은 상위에 속한 것으로 나타났다. 또한 INEX 2004에서의 계속적인 연구[33]를 통해 XML의 각 요소를 검색 대상으로 하기 위해 각 요소의 계층마다 언어모델을 적용하였다. 실험결과 다른 전통적인 검색모델 기반 모델들 보다 그 성능이 우수한 것으로 나타났다.

앞에서 살펴본바와 같이 언어모델은 많은 ad hoc 검색에서 적용되었다. 또한 최근에는 자원 선택(Resource Selection)과 다른 텍스트 데이터베이스로부터의 결과 통합(Result Merging)[34], 단락검색(Passage Retrieval)[35], 문헌 브라우징(Document Browsing)[36] 등의 광범위한 정보검색 분야에서도 언어모델에 대한 다양한 연구가 이루어지고 있다. 넓게는 이런 연구들 또한 정보검색의

---

31) J. List et al., "The Tijah XML-IR System at INEX 2003," *In INEX 2003 Workshop Proceedings*(2003), pp.102-109.

32) P. Ogilvie, and J. Callan, "Language Models and Structured Document Retrieval," *In Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval*(2003), pp.12-18.

33) P. Ogilvie and J. Callan, "Hierarchical Language Models for XML Component Retrieval," *In Pre-Proceedings of the Workshop of the INitiative for the Evaluation of XML Retrieval*(2004), pp.119-125.

34) L. Si et al. "A Language Modeling Framework for Resource Selection and Results Merging," *CIKM'02*, Mclean, Virginia(2002), pp.391-397.

35) X. Liu and W. B. Croft, "Passage Retrieval Based On Language Models," *CIKM'02*, Mclean, Virginia (2002), pp.375-382.

36) D. J. Harper, S. Coulthard and Y. Sun, "A Language Modeling Approach to Relevance Profiling for Document Browsing," *JCDL'02*, Portland, Oregon(2002), pp.76-83.

영역에 포함되지만, 앞서 말한 대로 본 연구에서는 ad hoc 검색에서 언어모델 적용에 관한 연구만을 다루기로 한다.

국내의 경우 LMIR 연구는 미비하며, 좁은 개념에서 LMIR은 아니지만 음성인식, 자연언어처리, 자동교정과 같은 다른 응용분야에서의 언어모델에 관한 연구는 심철민과 권혁철[37], 강승식[38], 강미경과 권혁철[39], 이도길 외[40] 그리고 최학윤[41] 등 다수가 있다.

다음 장에서는 앞에서 살펴본 LMIR 문헌조사를 통해 정보검색을 위한 다양한 언어모델들의 접근과 확장을 위한 기술들의 적용에 관한 연구동향 및 결과 그리고 주요한 기술들에 대하여 분석하고자 한다.

Ⅳ. 언어모델링 정보검색의 연구동향 및 결과 분석

1. 언어모델링 연구 동향 분석

본 연구에서는 1998년에서 2004년까지 수행된 LMIR의 실험적 연구들을 그 성격에 따라 크게 1세대 LMIR과 2세대 LMIR의 두 가지로 대별하였다. 1세대 LMIR 연구의 범주에는 주로 1998년 직후의 이 분야 초기의 연구들로 언어모델링 정보검색의 타당성과 우수성을 검증하고자 확률모델, 벡터모델과 같은 전통적인 모델 기반 정보검색과의 성능비교에 대한 연구들을 포함시켰다. 2세대 LMIR 연구 범주는 보다 나은 성능의 언어모델링 정보검색 개발을 위해 언어모델의 다른 응용분야들에서 연구 개발된 언어모델링 도구들의 적용과 그것이 검색성능에 미치는 영향들에 관한 실험적 연구들을 포함시켰다. 물론 Song과 Croft[42]의 연구와 같이 1세대 LMIR과 2세대 LMIR의 성격 모두를 포함한 연구들도 다수 존재한다.

가. 1세대 언어모델링 정보검색(LMIR)

Ponte와 Croft의 모델과 같이 unigram 기반의 전통적인 LMIR 연구를 포함한 비교적 단순한

37) 심철민, 권혁철, "연어 정보에 기반한 한국어 철자 검사와 교정기의 구현," 정보과학회논문지, 제23권, 제8호(1996), pp.776-785.
38) 강승식, "음절 bigram을 이용한 띄어쓰기 오류의 자동 교정," 음성과학회논문지, 제8권, 제2호(2001), pp.83-90.
39) 강미경 · 권혁철, "효율적인 문서처리를 위한 띄어쓰기 교정 기법 개선," 한국정보과학회 2003 봄 학술발표논문집(B)(2003), pp.486-488.
40) 이도길 등, "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델," 정보과학회논문지 : 소프트웨어 및 응용, 제30권, 제4호(2003), pp.358-370.
41) 최학윤, "Back-off bigram을 이용한 대용량 연속어의 화자적응에 관한 연구," 한국통신학회논문지, Vol.28, No.9C (2003, 9), pp.884-890.
42) F. Song and W. B. Croft, *ibid.*, pp.316-321.

언어모델을 적용한 것을 1세대라고 볼 수 있다. 이 시기의 연구는 주로 LMIR과 전통적인 확률모델이나 벡터 모델과의 성능 비교를 통해 정보검색에서 언어모델 적용의 타당성과 우수성을 증명하고자 하였다. 대부분 LMIR의 초기 연구들이 이 범주에 포함되지만, 최근의 INEX를 중심으로 한 구조화된 XML 문헌 검색에서 구조적인 질의 처리를 위해 기본적인 언어모델을 사용하는 연구들도 이 범주에 포함된다고 볼 수 있다. 이 범주에 속하는 대표적인 연구들로는 Ponte와 Croft[43], Hiemstra[44], Song과 Croft[45], Lavrenko와 Croft[46] 등이 있고, 또한 비교적 최근의 구조화된 질의처리를 위한 List 외[47]과 Ogilvie와 Callan[48]연구들이 이 범주에 속한다. 이들 연구들 중 일부는 단순한 완화기법을 통한 확장된 언어모델을 실험한 2세대적 LMIR 요소를 포함하고 있는 것도 있다.

나. 2세대 언어모델링 정보검색(LMIR)

2세대 LMIR의 연구 범주에는 자연언어처리 분야에서 축적되어온 다양한 언어모델링 기술들을 통해 보다 발전적이고 확장된 LMIR이 포함된다. LMIR의 성능 향상과 언어모델 적용의 최대 단점인 빈약한 훈련데이터를 해결하기 위해 문헌모델을 해당 분야의 말뭉치 모델이나 적합성 모델, 전체 컬렉션 모델 등과의 결합을 통해 확장하거나 이런 모델들과의 결합을 지원하기 위해 다양한 완화기법을 사용하여 완화하거나 일정한 상수를 통해 전체 확률분포의 이동시키는 등의 다양한 노력 등이 있었다. 앞서 개관한 연구들 중 Song과 Croft[49], Zhai와 Lafferty[50], Zhai와 Lafferty[51], Zin, Hauptmann[52]의 연구 등이 이 범주에 속한다고 볼 수 있다.

2세대 LMIR의 연구는 세부적으로 세 개의 그룹으로 범주화될 수 있다. 그것은 첫째, 언어모델 추정 대상을 확장하는데 초점을 맞춘 연구, 둘째, 완화기법을 통한 추정된 언어모델 확장에 관한 연구, 그리고 셋째, 통계적 언어모델링에 의한 개념적 확장에 관한 연구이다.

이들 각각에 대하여 보다 자세히 살펴보면 다음과 같다.

---

43) J. Ponte and W. B. Croft, *ibid.*, pp.275-281.
44) D. Hiemstra, *ibid.*, pp.569-584.
45) F. Song and W. B. Croft, *ibid.*, pp.316-321.
46) V. Lavrenko and W. B. Croft, *ibid.*, pp.123-125.
47) J. List et al. *ibid.*, pp.102-109.
48) P. Ogilvie, and J. Callan, *ibid.*, pp.12-18.
49) F. Song and W. B. Croft, *ibid.*, pp.316-321.
50) C. Zhai and J. Lafferty, "Model-based Feedback in the Language Modeling Approach to Information Retrieval," *CIKM'01*, Atlanta, Georgia(2001), pp.403-410.
51) C. Zhai and J. Lafferty, "Document Language Models, Query Models, and Risk Minimization for Information Retrieval," *SIGIR'01*, New Orleans, Louisiana(2001), pp.111-119.
52) R. Jin and A. G. Hauptmann, *ibid.*, pp.42-48.

(1) 언어모델 추정 대상에 관한 연구

Song과 Croft는 문헌 모델을 확장시키기 위해 말뭉치 모델을 결합시켰다. 문헌 모델은 대규모의 미등록어들이 있을 수 있다는 관점 그리고 어떤 알려진 단어들의 변칙적인 분포가 있을 수 있다는 관점에서 안정적이지 않다. 하지만 문헌모델에 말뭉치 모델의 결합함으로써 대규모의 문헌들로부터 언어모델이 얻어지는 것이기 때문에 그런 언어 모델은 안정적이게 된다. 정보검색에 관한 말뭉치에서, 예를 들면 "keyword"라는 단어는 "crocodile" 단어보다 아마 자주 출현하게 될 것이고 결과적으로 그것은 말뭉치 모델로 문헌 모델을 확장하는 것을 돕는다. Ponte와 Croft의 모델 $P(w|M_d)$는 선형 보간(Linear Interpolation)에 의해 문헌 언어모델과 말뭉치 모델을 결합하여 식(4)로 계산된다.[53] 이 식에서 λ은 0에서 1까지의 매개변수 가중치이고 $tf(w,d)$은 문헌 d에 나타난 단어 w의 출현빈도이고 dl_d은 문헌d의 문헌 길이, cf_w는 전체 컬렉션에서의 w 출현빈도, cs : 컬렉션에서 전체 토큰 수를 나타낸다.

$$P(w|M_d) = \lambda \frac{tf(w,d)}{dl_d} + (1-\lambda)\frac{cf_w}{cs} \qquad (4)$$

Zhai와 Lafferty(2001b)[54]는 문헌의 언어모델 뿐만 아니라 질의를 위한 언어모델을 측정하기 위해 기존의 언어모델링 접근을 확장하는 새로운 틀을 제시했다. 문헌과 질의의 유사성은 문헌 모델과 질의 모델사이의 Kullback-Leibler(KL)의 차이에 의해 측정된다. 여기에서 제시된 모델은 전통적인 확률 모델의 개념을 따르고 기존의 언어모델들을 수행하고 또한 질의와 문헌의 언어모델을 확장하기 위해 Markov chain 방법을 따른다. Markov chain 방법은 질의생성을 위한 번역모델 $t_\alpha(q_i \mid w)$과 최초로 이용자 U에 의해 선택된 단어 w의 사전 분포 $p(w \mid U)$에 따라 단어의 사후 확률이 계산되는 방법이다. 질의어가 독립적으로 생성된다고 가정하고 질의 q을 위한 확장된 언어모델은 식(5)에 의해 계산된다.

$$P(w|\hat{\theta}_q) \propto \sum_{i=1}^{m} t_\alpha(q_i|w)p(w|U) \qquad (5)$$

유사한 방식으로 문헌 d도 Markov chain 방법을 사용하여 확장될 수 있다.

$$P(w|\hat{\theta}_d) \propto t_\alpha(d|w)p(w|U) \qquad (6)$$

53) F. Song and W. B. Croft, *ibid.*, p.318.
54) C. Zhai and J. Lafferty, *ibid.*, pp.111-119.

Lavrenko와 Croft[55]는 질의 생성과정에서 모델을 시도하지 않고, 질의의 적합성 모델을 측정하는 새로운 모델을 제시했다. 이것은 적합한 문헌에서의 단어 출현에 대한 확률 $P(w|R)$을 할당한 적합성 모델 R을 제시한 것이다. 적합한 문헌은 $P(w|R)$ 분산으로부터 랜덤 표본을 추출하고 질의와 문헌들은 R로부터의 표본이 된다. 이것의 계산 식(7)은 다음과 같다.

$$P(w|R) \approx P(w|Q) = \frac{P(w, q_1, ...q_m)}{P(q_1, ...q_m)} = \frac{P(w, q_1, ...q_m)}{\sum_{\nu \in \text{vocabulary}} P(\nu, q_1, ...q_m)} \qquad (7)$$

(2) 완화기법을 통한 언어모델의 확장에 관한 연구

완화기법을 사용한 언어모델의 확장에 대해서는 언어모델의 보조적인 수단으로 가장 일반화되어 있는 확장 방법이다. 물론 이런 완화 기법들은 음성인식 분야의 오랜 연구의 결과물들 중 하나이며 몇몇 완화기법들이 정보검색분야에서 자주 적용되고 성능 향상에 많은 역할을 수행하고 있다. 많은 연구들에서 언어모델 확장을 위해 다양한 완화기법들이 적용되고 있지만 주로 보조적인 수단으로 사용되고 있기 때문에 연구의 초점이 그것에 있지는 않다. Zahi와 Lafferty는 2004년의 논문에서 단순한 언어모델과 몇몇 완화기법을 적용한 확장된 언어모델과의 성능 비교를 통해 언어모델 성능에서 완화기법의 중요성을 재확인시켜 주었다. 또한 정보검색에서 적용하기 쉽고, 여러 연구들에서 자주 사용된 완화기법 Jelinek-Mercer와 Dirichlet Priors, 그리고 Absolute Discounting을 소개하고 이들 각 각을 적용한 정보검색시스템의 성능을 실험하였다.

Jelinek-Mercer 방법은 주로 둘 이상의 언어 모델을 결합할 때 간단하게 사용할 수 있는 완화기법으로 계수 λ를 사용하여 각 언어모델의 영향을 통제한다. Dirichlet Priors를 사용한 베이지언 완화기법은 둘 이상의 언어모델을 결합할 때 각 언어모델의 이질성이 인정되며 각 단어의 확률 분포는 각 언어모델의 산출 과정을 통해 산출되어 보존된다. 즉 Drichlet 다항분포 방법은 다차원적인 언어 모델을 허락하는 것이다. Absolute Discounting 방법은 산출된 확률을 낮추기 위해 일정한 상수를 빼주는 방법이다. 이것은 단어의 확률을 낮추기 위한 것으로 Jelinek-Mercer 방법과 개념적으로는 유사하지만, Jelinek-Mercer 방법의 경우는 (1-λ)의 값을 곱해주는 반면, 이것은 상수를 빼주는 것으로 단어의 확률들을 낮춘다. 〈표 1〉은 이들 세 완화기법에 대한 요약이다.[56]

55) V. Lavrenko and W. B. Croft, "Relevance-based Language Models," In W. B. Croft, D. Harper, D. H. Kraft, J. Zobel eds., *SIGIR'01*(2001), pp.123-125.

56) C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Transactions on Information Systems*, Vol.22, No.2(April 2004), p.185.

〈표 1〉 세 가지 주요 완화기법의 요약

방법	$P_s(w \mid d)$	α_d	매개 변수
Jelinek-Mercer	$(1-\lambda)P_{ml}(w \mid d) + \lambda P(w \mid C)$	λ	λ
Dirichlet	$\frac{C(w;d) + \mu P(w \mid C)}{\mid d \mid + \mu}$	$\frac{\mu}{\mid d \mid + \mu}$	μ
Absolute Discount	$\frac{max(C(w;d) - \delta, 0)}{\mid d \mid} + \frac{\delta \mid d \mid_u}{\mid d \mid} P(w \mid c)$	$\frac{\delta \mid d \mid_u}{\mid d \mid}$	δ

(3) 통계적 언어모델링에 의한 개념적 확장에 관한 연구

Ponte와 Croft 모델과 같이 대부분의 LMIR은 unigram 기반 언어모델을 적용하고 있다. unigram 모델은 각 단어들이 독립적으로 발생한다는 강한 가정을 통해 만들어진다. 그래서 unigram 모델은 단어들의 연속 확률이 개개의 단어들의 확률들의 곱으로 나타나게 된다. 하지만 bigram과 trigram 모델은 공간의 문맥을 고려한 것으로 bigram에서 새로운 단어들의 확률은 이전 단어의 확률에 의존하고 trigram에서 새로운 단어들의 확률은 이전 두 단어의 확률에 의존한다.[57] 정보검색 분야의 특성상 trigram 이상의 통계적 언어모델이 적용된 경우는 없으나, 최근에 들어 다수의 논문들에서 bigram과 같이 두 단어들 간의 의존성을 표현하고자 하는 예들이 있다. Song과 Croft의 모델에서는 단어쌍과 구로써 더 의미를 갖는 단어들의 동시 확률을 위해 unigram 모델과 bigram 모델이 선형보간법에 의해 결합되었고 Srikanth와 Srikanth의 모델은 연속하는 두 단어가 순서에 상관없이 나타날 확률에 기반한 bi-term 언어모델을 적용하였고, Gao 외의 모델은 두 단어가 인접하지 않아도 의존성을 가지고 있다는 가정아래 두 단어의 의존성을 표현하는 의존 언어모델을 제시하였다. bigram기반 언어모델에서 질의 생성확률은 『$P(Q \mid D) = P(q_1 \mid D)\prod_{i=2}^{m} P(q_i \mid q_{i-1})$』 이고 bi-term 기반 언어모델에서 질의 생성확률 『$P_{BT}(q_i \mid q_{i-1}, D) = \frac{1}{2}(P_{BG}(q_i \mid q_{i-1}, D) + P_{BG}(q_1 \mid q_{i-1}, D))$』이다.[58] 의존 언어모델에서 질의 생성확률은 식(8)에 의해 추정된다.[59] 여기에서 L(Linkage)은 모든 가능한 결합정보이다.

$$\log P(Q \mid D) = \log P(L \mid D) + \sum_{i=1}^{m} \log P(q_i \mid D) + \sum_{(i,j) \in L} MI(q_i, q_j \mid L, D) \qquad (8)$$

2. 언어모델링 정보검색 평가 실험들의 결과 분석

LMIR의 실험적 연구들은 다양한 목적을 위해 수행되었다. 그 목적의 첫째는 새롭게 개발된

57) F. Song and W. B. Croft, *ibid.*, p.316.
58) Gao et al. *ibid.*, pp.175-176.
59) *Ibid.*, p.172.

LMIR의 성능 평가와 다른 모델들과의 성능 비교이고, 둘째는 성능 비교를 통해 정보검색분야에 적합한 언어모델링 방법들과 기법들을 고찰하는 것이다. 1세대 LMIR의 대부분은 확률모델과 벡터모델과 같은 전통적인 검색모델들과 기본적인 언어모델과의 성능 비교에 관한 것이었고, 2세대 LMIR은 기본적인 언어모델과 확장된 다양한 언어모델들과의 성능을 비교하기 위한 것이었다. 각각의 실험적 연구들은 정보검색의 성능 평가를 위해 TREC과 INEX에서 제공하는 대규모의 테스트 컬렉션을 통해 실험하였는데, TREC을 통한 실험적 연구의 경우 TREC2~TREC8, WSJ(Wall Street Journal), ZIFF(Information from Computer Select disks), LT(Los Angeles Times), 그리고 FT(Financial Times)와 같은 다양한 문헌집합을 사용하거나 다양한 토픽(Topic)을 사용하고 있다. 또한 실험에서 적용된 변수 정의가 다양하기 때문에 실험에서 평가된 검색시스템의 평균정확률(Avg:P)이나 재현율-정확률(RPr) 값의 절대적인 비교는 불가능하다. 하지만 다수의 실험결과들에서 언어모델 성능평가에 대한 또는 확장된 언어모델 성능 평가에 대한 상대적인 결과값들이 주어져 있다. 몇 몇 연구들에서 도출된 주목할 만한 실험 결과들을 정리해보면 〈표 2〉와 같다.

〈표 2〉 LMIR 실험들의 검색 성능 비교

연구자	테스트 컬렉션	검색모델	실험 결과	
			Avg:P*	RPr**
Ponte와 Croft(1998)	TREC 2와 3	tf·idf	0.1868	0.2473
		LM	0.2233	0.2876
Song과 Croft(1999)	TREC4	INQUERY	0.1917	·
		LM	0.1890	·
		GLM(40)	0.1905	·
		GLM(40+90)	0.1923	·
Lavrenko와 Croft(2001)	AP(title)	LM	0.2021	0.2546
		Relevance LM	0.2617	0.2935
Zhai와 Lafferty(2001a)	TREC8	Simple LM	0.2560	·
		Mixture Feedback	0.2820	·
Zhai와 Lafferty(2001b)	TREC8	Simple LM	0.2410	·
		Query Model	0.2660	·
		Markov Chain QM	0.2940	·
		tf·idf + Rocchio	0.2560	·
Jin, Hauptmann, Zhai(2002)	WSL	Okapi	0.1719	·
		LM	0.1844	·
		Title LM	0.1950	·
Jin et al.(2002)	TREC4	LM	0.1825	·
		New LM	0.1911	·
Zhai와 Lafferty(2004)	TREC7(title)	Jelinek-Mercer	0.1670	·
		Dirichlet prior	0.1860	·
		Absoulte Discounting	0.1720	·
Gao et al.(2004)	ZIFF	Okapi	0.1536	·
		LM-unigram	0.1647	·
		LM-bigram	0.1717	·
		LM-bi-term	0.1766	·
		Dependence LM	0.1818	·

각 실험결과에서 상대적인 결과값들이 유사한 경향을 보임을 알 수 있다. Ponte와 Croft, Jin, Hauptmann, Zhai, Gao 외 등의 실험 결과는 확장되지 않은 기본적인 언어모델링 정보검색이라 할지라도 전통적인 정보검색모델보다 성능이 우수한 것으로 나타났다. 그리고 Lavrenko와 Croft, Zhai와 Lafferty, Zhai와 Lafferty, Jin, Hauptmann, Zhai, Jin 외, Zhai와 Lafferty 그리고 Gao 외(2004) 등의 연구들에서의 실험결과는 일반적으로 기본적인 언어모델보다 다양한 언어모델링 기술들을 사용하여 확장된 언어모델들의 성능이 더 뛰어남을 보여주고 있다.

또한 Song과 Croft의 실험결과와 같이 전통적인 모델과 언어모델이 성능에서 많은 차이가 나타나지 않는 실험들도 다수 존재한다. 하지만 이런 실험들에서 조차 60년대부터 계속해서 실험되어 정제되어온 전통적인 모델과 대등한 수준의 성능을 보임으로 새로운 검색모델로써 언어모델이 상당히 매력적이라는 사실을 확인할 수 있다.

V. 요약 및 결론

본 연구의 목적은 지난 5-6년간 연구되어온 LMIR 관한 망라적인 문헌조사를 통해 정보검색 분야에 적용된 언어모델들과 언어모델의 확장을 위해 사용된 보조적인 기술들을 검토하고, 각 연구들에서 실험된 전통적인 모델들과 상대적인 성능 비교 결과를 고찰하고자 한 것이었다. 이를 통하여 LMIR 연구에서 많이 적용된 언어모델들과 기술들의 동향을 파악하고, 언어모델 기반 정보검색 시스템 개발을 위한 이론적 토대를 마련하고자 하였다.

LMIR은 1998년 Ponte와 Croft의 연구 이후 다수의 연구들에서 개발되고 실험되었는데, 1세대 LMIR로 구분되는 초창기의 연구들은 정보검색에서의 전통적인 검색모델과 언어모델 적용의 성능 비교를 통해 언어모델 적용의 우수성과 타당성을 입증하는데 초점을 두었다. 이후 2세대 LMIR 연구에서는 이를 바탕으로 보다 나은 성능을 위해 기본적인 언어모델과 다양한 언어모델링 기술들을 사용하여 확장된 언어모델의 성능을 비교·실험하고자 한 것으로 나타났다.

다수의 선행연구들에서 도출된 결과로는 첫째, LMIR의 성능이 전통적인 검색모델의 그것과 필적하거나 더 우수하다는 것이었고, 둘째, 다양한 언어모델링 기법을 사용한 확장된 언어모델들이 기본적인 언어모델보다 그 성능이 우수하다는 것이었다.

본 연구를 통하여 LMIR은 음성인식 분야에서 30여 년간 축적되어온 언어모델링의 다양한 기술들을 응용함으로써 성능 향상의 많은 잠재력을 가지고 있음을 또한 인지할 수 있었다. LMIR 연구들은 아직 시작단계에 있음을 감안하면 그 적용 가능성과 잠재력은 매우 주목할 만한 것으로 간주할 수 있다. 따라서 LMIR에 대한 지속적인 연구를 통하여 지금까지 나타난 단점들을 보완하고 정보검색 분야에 보다 적합한 언어모델링 기법의 개발에 많은 노력을 기울여 나가야 할 것이다.

참 고 문 헌

강미경, 권혁철. "효율적인 문서처리를 위한 띄어쓰기 교정 기법 개선," 한국정보과학회, 2003 봄 학술발표논문집(B)(2003), pp.486-488.

강승식. "음절 bigram을 이용한 띄어쓰기 오류의 자동 교정." 음성과학회논문지, 제8권, 제2호(2001) pp.83-90.

박선희, 노용완, 홍광성. "문장음성인식을 위한 VCCV 기반의 언어모델과 Smoothing 기법 평가." 정보처리학회논문지B, 제11-B권, 제2호(2004, 4), pp.241-246.

심철민, 권혁철. "언어 정보에 기반한 한국어 철자 검사와 교정기의 구현." 정보과학회논문지, 제23권, 제8호(1996), pp.776-785.

이도길 외. "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델." 정보과학회논문지 : 소프트웨어 및 응용, 제30권, 제4호(2003), pp.358-370.

이진석, 박재득, 이근배, "K-SLM Toolkit을 이용한 한국어의 통계적 언어 모델링 비교." 제11회 한글 및 한국어정보처리 학술대회 논문발표집(1999).
〈http://nlp.postech.ac.kr/lab__papers/9910__h%26h__wolfpack.doc〉 [cited 2005. 4. 12]

최학윤. "Back-off bigram을 이용한 대용량 연속어의 화자적응에 관한 연구." 한국통신학회논문지, Vol.28, No.9C(2003, 9), pp.884-890.

Croft, W. Bruce, Jamie Callan, John Lafferty. *Workshop on Language Modeling and Information Retrieval.* Carnegie Mellon University, Pittsburgh, Pennsylvania, 2001.

Croft, W. Bruce. "Language Models for Information Retrieval." *Proceedings of the 19th International Conference on Data Engineering*(2003), pp.3-7.

Gao, Jianfeng et al. "Dependence Language Model for Information Retrieval." *SIGIR'04*, Sheffield, South Yorkshire(2004), pp.170-177.

Harper, David J., Sara Coulthard and Sun Yixing. "A Language Modeling Approach to Relevance Profiling for Document Browsing." *JCDL'02*, Portland, Oregon(2002), pp.76-83.

Jin, Rong, Alex G. Hauptmann. "Title Language Model for Information Retrieval." *SIGIR'02*, Tempere(2002), pp.42-28.

Jin, Rong et al. "Language Model for IR Using Collection Information." *SIGIR'02*, Tampere(2002), pp.419-420.

Lavrenko, Viktor, Chengxiang Zhai. "Text Retrieval and Mining."
〈http://www.stanford.edu/class/cs276a/handouts/lecture12.ppt〉 [cited 2005. 5. 22]

Lavrenko, V., W. B. Croft. "Relevance-based Language Models." In W. B. Croft, D. J. Harper, D. H. Kraft, J. Zobel, eds., *SIGIR'01*(2001), pp.123-125.

List, J., V. Mihajlovic, G. Ramirez, and D. Hiermstra. "The Tijah XML-IR System at INEX 2003." *In INEX 2003 Workshop Proceedings*(2003), pp.102-109.

Liu, Xiaoyon, W. Bruce Croft. "Passage Retrieval Based On Language Models." *CIKM'02*, Mclean, Virginia(2002), pp.375-382.

Luk, R. et al. "A Survey in Indexing and Searching XML Document." *JASIS&T*, Vol.53, No.6(Feb. 2002), pp.415-437.

Lyer, R. and M. Ostendorf. "Relevance Weighting for Combining Multi-domain Data for N-gram Language Modeling." *Computer Speech and Language*, Vol.13(1999), pp.280-284.

Metzler, Donald and W. Bruce Croft. "Combining the Language Model and Inference Network Approaches to Retrieval." *Information Processing & Management*, Vol.40(2004), pp.735-750.

Miller, D., T. Leek, R. Schwartz. "A Hidden Markov Model Information Retrieval System." *In Proceedings of the 22nd Annual International ACM SIGIR Conference*(1999), pp.214-221.

Ogilvie, P. and J. Callan. "Language Models and Structured Document Retrieval." *In Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval(INEX)*(2003), pp.12-18.

Ogilvie, P. and J. Callan. "Hierarchical Language Models for XML Component Retrieval." *In Pre-Proceedings of the Workshop of the INitiative for the Evaluation of XML Retrieval(INEX)*(2004), pp.119-125.

Ponte, Jay M., W. Bruce Croft. "A Language Modeling Approach for Information Retrieval." *SIGIR'98, Melbourne*(1998), pp.275-281.

Rosenfeld, R. "Two Decades of Statistical Language Modeling: Where Do We Go From Here?" In *Proceeding of the IEEE*, Vol.88, No.8(2000), pp.1274-1278.

Si, Luo et al. "A Language Modeling Framework for Resource Selection and Results Merging." *CIKM'02*, Mclean, Virginia(2002), pp.391-397.

Song, Fei, W. Bruce Croft. "A General Language Model for Information Retrieval." *CIKM'99*, Kansas City, Mo(1999), pp.316-321.

Sparck Jones, Karen et al., "Language Modeling and Relevance." In W. B. Croft and J. Lafferty.,

editors, Language Modeling for Information Retrieval. London: Kluwer Academic Publishers, 2003.

Srinkanth M. and R. Srinkanth. "Bi-term Language Models for Document Retrieval." SIGIR'02, Tempere(2002), pp.425-426.

Zaragiza, Hugo, Djoerd Hiemstra, Michael Tipping. "Bayesian Extension to the Language Model for Ad Hoc Information Retrieval." *SIGIR'03*, Toronto(2003), pp.4-9.

Zhai, Chengxiang, John Lafferty. "Document Language Models, Query Models, and Risk Minimization for Information Retrieval." *SIGIR'01*, New Orleans, Louisiana(2001), pp.111-119.

Zhai, Chengxiang, John Lafferty. "Model-based Feedback in the Language Modeling Approach to Information Retrieval." *CIKM'01*, Atlanta, Georgia(2001), pp.403-410.

Zhai, Chengxiang, John Lafferty. "Two-Stage Language Models for Information Retrieval." *SIGIR'02*, Tampere(2002), pp.49-56.

Zhai, Chengxiang, John Lafferty. "A Study of Smoothing Methods for Language Models Applied to Information Retrieval." *ACM Transactions on Information Systems*, Vol.22, No.2(April 2004), pp.179-214.

INEX Home page. <http://inex.is.informatik.uni-duisburg.de>

TREC Home page. <http://trec.nist.gov/>