기계번역을 이용한 교차언어 문서 범주화의 분류 성능 분석

Classification Performance Analysis of Cross-Language Text Categorization using Machine Translation

이 용 구(Yong-Gu Lee)*

목 차

초 록

교차언어 문서 범주화(CLTC)는 다른 언어로 된 학습집단을 이용하여 문헌을 자동 분류할 수 있다. 이 연구는 KTSET으로부터 CLTC에 적합한 실험문헌집단을 추출하고, 기계 번역기를 이용하여 가능한 여러 CLTC 방법의 분류 성능을 비교하였다. 분류기는 SVM 분류기를 이용하였다. 실험 결과, CLTC 중에 다국어 학습방법이 가장 좋은 분류 성능을 보였으며, 학습집단 번역방법, 검증집단 번역방법 순으로 분류 성능이 낮아졌다. 하지만 학습집단 번역방법이 기계번역 측면에서 효율적이며, 일반적인 환경에 쉽게 적용할 수 있고, 비교적 분류 성능이 좋아 CLTC 방법 중에서 가장 높은 이용 가능성을 보였다. 한편 CLTC에서 기계번역을 이용하였을 때 번역과정에서 발생하는 자질축소나 주제적 특성이 없는 자질로의 번역으로 인해 성능 저하를 가져왔다.

ABSTRACT

Cross-language text categorization(CLTC) can classify documents automatically using training set from other language. In this study, collections appropriated for CLTC were extracted from KTSET. Classification performance of various CLTC methods were compared by SVM classifier using machine translation. Results showed that the classification performance in the order of poly-lingual training method, training-set translation and test-set translation. However, training-set translation could be regarded as the most useful method among CLTC, because it was efficient for machine translation and easily adapted to general environment. On the other hand, low performance was shown to be due to the feature reduction or features with no subject characteristics, which occurred in the process of machine translation of CLTC.

키워드: 교차언어 문서 범주화, 문헌자동분류, 다국어 분류, 다국어 학습, 교차언어 학습
Cross-Language Text Categorization, CLTC, Document Classification, Multilingual Classification, Poly-Lingual Training, Cross-Language Training

* Visiting Scholar, School of Information Sciences, University of Pittsburgh(yglee@mail.sis.pitt.edu)
논문접수일자: 2009년 2월 27일 최초심사일자: 2009년 3월 2일 게재확정일자: 2009년 3월 9일

1. 서 론

최근 웹의 인기는 영어를 포함한 다양한 외국어로 된 문헌의 증가를 초래하였으며, 이용자도 이들 문헌을 처리해야 할 기회가 증가하였다. 또한 이러한 증가는 국제화 내지 세계화 추세와 맞물러 더욱 가속화되고 있다.

이와 같은 환경의 변화는 다양한 분야에서 다국어 문헌을 이용한 정보처리를 요구하고 있다. 즉 여러 언어로 된 문헌을 이용하기 위해 자동색인이나 편리한 검색 등에 대한 조직화를 필요로 한다. 대표적인 예로서 한 언어로 된 질의를 입력하면 다른 언어 또는 여러 언어로 된 문헌을 검색해 주는 교차언어 또는 다국어 정보 검색(cross-language information retrieval: CLIR or multi-lingual information retrieval)을 들 수 있다. 이는 이용자가 자신의 주 언어로 된 질의어를 검색시스템에 입력하면 다른 언어로 된 적합문헌을 검색 결과로 제공하여 주는 검색기능이다.

일반적인 문서 범주화 또는 자동분류(text categorization or text classification)는 컴퓨터가 특정 주제범주(label)로 분류된 문헌집합인 학습집단(training set)을 사용하여 학습한 후, 새로운 문헌에 대해 주제범주를 할당하는 작업(자동분류)을 수행하게 된다. 따라서 이러한 문서 범주화 기법을 사용하기 위해서는, 특정 문헌이 어느 범주에 속하는지를 태깅한 정보를 포함하는 학습문헌이 먼저 갖추어져 있어야만 한다.

일반적으로 각 문헌에 주제범주를 부여하는 작업은 수작업에 의존하게 되며, 이는 높은 처리 비용을 초래한다. 만약 한 언어에서 얻을 수 있는 범주 정보를 이용하여 분류기를 만든 후, 범주 정보가 부여되지 않은 다른 언어로 된 문헌에 대해 별도의 수작업 분류 없이 자동 분류를 한다면, 비용 절감 효과와 함께 매우 효율적인 작업 처리가 가능할 것이다. 이를 가능하게 해 주는 기법이 바로 교차언어 문서 범주화(cross-language text categorization: CLTC)이다.

교차언어 문서 범주화 또는 자동분류는 문서 범주화의 새로운 연구영역이다. 교차언어 문서 범주화는 기존의 문서 범주화와 유사하지만, 한 단계 더 나아가 서로 다른 언어들로 이루어진 학습집단과 검증집단(test set)을 그 대상으로 자동 분류를 수행한다. 즉 한 언어로 된 학습집단을 이용하여 분류기를 구축한 후 이를 이용하여 다른 언어로 된 문헌을 자동분류하거나, 여러 언어로 된 학습집단을 동시에 이용하여 여러 언어로 구성된 문헌집단을 자동분류하는 것이다. 예를 들어 정치, 경제, 사회, 문화 등의 범주로 이루어진 영어 뉴스 기사 학습집단을 이용하여 한글로 된 동일 범주의 뉴스기사를 분류할 수 있다.

일반적으로 한글로 된 문서 범주화 실험 문헌집단은 매우 한정적이거나 그 규모가 작다. 이로 인하여 문서 범주화 기법의 성능이 우수함에도 불구하고, 한글로 된 학습집단의 부재로 인해 실제 업무에 적용하는데 어려움이 따른다. 따라서 기존의 다른 언어로 된 범주정보를 이용할 수 있는 교차언어 문서 범주화 기법은 더더욱 한글 환경에 있어서 유용할 수 있다.

이 연구에서는 여러 기계 번역기를 이용하여 다양한 학습집단과 검증집단을 포함하는 실험 문헌집단을 만든 후, 이를 CLTC 기법에 적용하여 자동 번역에 따른 범주화 성능 차이를 비

교실험하고, 여러 번역 데이터를 결합하여 자동 분류 실험을 하였을 때 어떤 결과를 가져오는지 제시하고자 하였다. 즉, 이 연구의 목적은 CLTC 기법을 적용하기 위해 언어가 다른 문헌 사이에 필요한 번역 방법 중 기계번역을 이용하였을 때 가장 좋은 성능을 보이는 CLTC 기법을 확인하고, 더 나아가 기계번역을 이용할 때 발생할 수 있는 문제점과 해결책에 대한 방향을 제시하고자 하였다. 이를 위한 선행연구를 통해 다양한 CLTC의 기법을 파악하고 각각의 기법에 따라 실험문헌집단을 구축한 후, SVM 분류기를 이용하여 CLTC 기법 간에 자동 분류 성능을 평가하였다.

2. 교차언어 문서 범주화

CLTC는 새롭게 대두되는 연구영역으로서 CLIR과 밀접히 관련되어 있다. 두 분야 모두 다국어로 된 문헌을 대상으로 정보처리 작업이 이루어지는데, 특히 CLTC를 이해하기 위해 많은 선행 연구가 이루어진 CLIR 분야의 주요 기법들을 살펴보면 다음과 같다.

CLIR에서 문헌집단과 질의를 매칭하기 위한 기법은 네 가지로 나누어진다(Oard and Diekema 1998; Peters and Sheridan 2001; Kishida 2005). 먼저 언어 사이에 번역이 필요 없는 기법과 반드시 필요한 기법으로 나눌 수 있다. 전자는 실험대상 언어의 기원이 같아서 번역과정이 필요 없다. 후자는 다시 어떤 대상을 번역하느냐에 따라 질의번역 방법, 문헌번역 방법, 그리고 중간언어 방법(interlingual techniques)으로 나눌 수 있다. 마지막 방법인 중간언어 방법은 자동번역의 한 기법에서 그 기원을 찾을 수 있는데, 질의와 문헌 모두를 언어 독립적인 표현(language-independent representation)으로 변환한 후에, 그 결과를 비교하는 방법이다. 관련 연구로서 LSI(latent semantic indexing), 다국어 시소러스 또는 워드넷(WordNet) 등과 같은 통제어휘 자원을 이용하는 연구들이 있다.

CLTC의 기법들을 살펴보면, CLIR의 주요 기법들과 유사함을 발견할 수 있다. CLTC의 방법론을 정리하면 다음과 같다(Bel, Koster and Villegas 2003; Rigutini, Maggini and Liu 2005; Amine and Mimoun 2007). 먼저 CLTC를 이용하여 자동분류를 하고자 하는 문헌집단이 각각의 언어에 대해 고유한 학습집단을 갖고 있다면, 다국어 학습(poly-lingual training) 방법을 이용한다. 만약 다국어로 된 시소러스나 온토로지와 같은 외부 자원을 이용할 수 있다면, 중간언어 방법을 이용한다. 또한 서로 다른 언어 간의 번역 과정을 거친 후, 그 결과를 자동 분류에 이용하는 교차언어 학습(cross-lingual training) 방법을 이용한다. 각각의 방법을 구체적으로 살펴보면 다음과 같다.

2.1 다국어 학습 방법

다국어 학습방법은 각기 다른 언어로 된 학습집단이 존재해야 이용할 수 있다. 먼저 서로 다른 언어로 된 학습집단을 이용하여 분류기를 구축하고 해당 언어의 문헌을 분류한다. 분류기가 사용한 자질 정보는 해당 언어로 된 모든 어휘들의 합집합에 대응한다. 따라서 이 방법에서는 한 언어를 다른 언어로 번역하는 과정

을 필요로 하지 않는다.

다국어 학습 방법은 세부적으로 여러 언어로 된 학습문헌을 이용하여 하나의 분류기를 구축하는 방법과 언어별로 별개의 분류기를 구축하는 방법으로 나눌 수 있다. Adeva, Calvo와 Ipina (2005)는 뉴스기사 75,000건을 대상으로 나이브 베이즈, 로치오 그리고 kNN 분류기를 이용하여 실험한 결과, 나이브 베이즈 분류기를 언어별로 구축한 방법이 다른 방법보다 분류 정확도가 약간 높은 것으로 나타났다. Gliozzo와 Strapparava(2005)는 영어와 이탈리아어로 된 뉴스 데이터로 학습집단을 만든 후, 용어-문헌 행렬에 SVD(Singular Value Decomposition) 방법을 적용하여 얻은 MDM(Multilingual Domain Models)을 CLTC에 이용하였으나 좋은 성능을 보이지는 않았다.

또한 다국어 학습 방법은 경우에 따라서는 분류기를 구축하기 전에 언어 식별 단계가 필요할 수 있다. Adeva, Calvo와 Ipina(2005)는 N-gram 빈도를 이용하여 언어별로 프로파일을 구축한 뒤, 이를 대상 문헌과 비교하여 문헌에 쓰인 언어를 식별하였다. 즉 언어식별 방법으로 한 언어에서 추출한 400개 정도의 최빈도 N-gram은 항상 그 언어와 매우 밀접히 관련되어 있다는 속성을 이용하였다. 이러한 속성을 통해 언어별 프로파일을 생성하고 이를 새로운 문헌과 비교하여 문헌에 쓰인 언어를 식별하였다.

2.2 교차언어 학습 방법

교차언어 학습방법에서는 오직 한 언어로 된 학습집단을 이용하여 분류기를 구축하고, 그 분류기를 이용하여 다른 언어로 된 문헌을 분류한다. CLTC를 수행하고자 할 경우에는 두 언어 또는 다국어 중에서 최소한 하나의 언어에는 주제범주가 부여된 학습문헌 집단이 존재하여야 한다. 따라서 이 방법을 적용하는 경우가 가장 보편적이라고 볼 수 있다. 이 방법에서는 두 가지 서로 다른 하위 방법을 이용할 수 있는데, 이는 다음과 같다.

- 학습집단 번역: 범주 정보가 부여된 학습집단을 대상언어(target language)로 번역한 후 그 결과를 기반으로 분류기를 만든 후에, 검증집단을 분류한다. 일반적으로 학습집단 문헌수가 검증집단보다 적으므로 번역처리양은 상대적으로 적어 효율적이다.
- 검증집단 번역: 범주 정보가 부여된 학습집단을 이용하여 분류기를 구축한 후, 이 분류기를 이용하여 범주 정보가 부여되지 않은 검증집단을 원문 언어(source language)로 번역하고 그 결과를 분류한다.

교차언어 학습 방법에서는 두 언어 간의 번역을 위해 기계번역 소프트웨어와 같은 번역기를 사용할 수 있으며, 일반 언어사전을 비롯하여 전문 용어사전까지 다양한 언어 자원을 이용하여 번역할 수 있다. 또한 번역을 위해 병렬 말뭉치나 비교 말뭉치 자원을 이용할 수도 있다.

Bel, Koster와 Villegas(2003)의 실험에서는 단일어만 이용한 자동분류 기법을 CLTC 기법(다국어 학습 방법과 교차언어 학습방법)과 비교하였다. 이 연구에서 교차언어 학습방법은 검증집단의 번역하는 방법과 학습집단에서 주요어만 번역하는 방법을 이용하였다. 이러한 기법들을 Winnow와 로치오 분류기에 적용한 결과, 단일어만 이용한 자동분류가 가장 좋은

성능을 보였으며, 그 다음으로 다국어 학습 방법, 검증집단 번역 방법, 학습집단에서 주요어를 이용한 방법 순으로 성능을 보여주었다.

Rigutini, Maggini와 Liu(2005)는 기계번역을 수행할 때 개입되는 잡음이나, 원문헌과 번역문헌 사이에 존재하는 다른 통계적 특성으로 인한 CLTC의 성능 저하를 방지하기 위해 EM 알고리즘을 사용하였다. 또한 CLTC 기법에서는 학습집단 번역 방법을 사용하였다. 실험 결과, EM 알고리즘과 자질선정을 결합한 방법이 90.6%로 단일어 분류 성능인 94.4%에 근접하는 결과가 나타났다.

Wu와 Lu(2008)는 한 언어의 학습집단을 다른 언어로 번역할 때 적절한 번역어를 찾기 위해 2개 국어 어휘사전(bilingual lexicon)과 EM 알고리즘을 사용하였다. 즉 특정 범주에서 한 단어가 여러 번역어를 가질 때, 최적의 번역어를 선택하기 위해 EM 알고리즘을 적용하였다. 이렇게 생성된 기본 자질들을 나이브 베이즈 클러스터링 알고리즘에 적용하여 분류 성능이 단일어 자동분류와 거의 유사한 결과가 나타났다.

2.3 중간언어 방법

중간언어 방법은 문헌에 나타난 개념이나 용어를 특정 언어에 의존하지 않는 보편적인 표현(representation)으로 번역한다. 이 방법에서는 서로 다른 언어로 된 문헌을 동일 범주로 부여하기 위해 이들이 얼마나 많은 표현을 공유하는지에 따라 결정할 수 있다. 이는 하나의 보편적인 개념에 대해 여러 언어에서 각각 표현한 단어들을 한 용어로 번역함으로써 가능하다(Rigutini, Maggini and Liu 2005).

Amine와 Mimoun(2007)은 학습단계에서 범주를 특징짓는 개념을 포함하는 범주 프로파일을 구축하였다. 즉 문헌에 출현한 용어를 WordNet의 synset으로 매핑하고, 이들 synset 사이에 존재하는 관계를 WordNet의 계층구조에서 찾아냄으로써 문헌의 주요 개념을 식별하였다. 이후 분류단계에서는 검증집단의 문헌을 번역하고 문헌 프로파일을 생성한 후, 이를 범주 프로파일과 유사도를 계산하여 범주에 할당하였다. 실험 결과, 최빈도 synset만 이용한 방법이 더 좋은 성능을 보였다.

Melo와 Siersdorfer(2007)는 번역기를 통해 번역된 실험문헌에 대해 WordNet을 이용하여 ORM(ontology region mapping)을 구축하였다. 여기서 ORM은 관련 개념을 온톨로지 내의 용어 그룹으로 매핑한다. 이렇게 매핑된 관련 개념을 문헌에 출현한 단어와 연결함으로써 그 단어에 적절한 가중치를 부여하거나 문헌에 새로운 자질을 추가하였다.

3. 실험 설계

3.1 실험문헌집단

문서 범주화 실험을 위한 실험문헌집단은 다양하나, CLTC를 위한 실험문헌집단은 극히 드물다. CLTC의 선행연구들을 살펴보면 ILO 말뭉치[1]를 사용한 사례가 있다. 이 말뭉치는

1) 이 말뭉치는 http://www.ilo.org/ilolex/index.htm에서 접근 가능하다.

유엔 노동기구인 국제노동기구에서 만든 규정집, 권고 등을 포함하는 데이터베이스로 영어, 프랑스어, 스페인어의 번역본으로 이루어진 병렬 말뭉치이다. 그 외 다른 선행연구들은 연구자가 직접 뉴스 웹 페이지를 수집하고 가공한 실험문헌집단을 사용하였다. 한국어의 경우, 다른 언어의 번역본이 공존하는 병렬 말뭉치는 있으나 대부분 이들은 범주정보를 제공하지 않는다. 따라서 이 연구에서는 정보검색용 실험문헌집단인 KTSET[2]에서 CLTC를 위한 실험문헌집단을 추출하였다.

KTSET은 국문 및 영문 저자, 서명, 서지사항, 초록, 분류번호, 색인어 등 18개의 접근점으로 탐색할 수 있는 문헌, 질의어, 적합성 정보 등으로 구성되었다. 입력된 데이터는 정보과학회 논문지, 한국 정보과학회 proceeding, 정보관리학회지에 수록된 1,053개의 논문이다. 분류번호의 경우 Computing Reviews에서 작성한 CRCS 번역본을 적용하였다. 한 문헌당 분류번호는 5개 이내로 부여하였다(김성혁 외 1993).

실제 KTSET은 정보검색용 실험문헌집단이므로 자동분류에 적합하도록 몇 가지 규칙을 적용하여 가공하였다. 첫째, 이 연구에서는 기계번역을 사용하기 때문에 적은 오타라도 정확한 번역에 영향을 미칠 수 있어 이를 수작업으로 수정하였다. 둘째, 실험 대상 문헌은 국문과 영문을 동시에 포함해야 하므로, 영문초록이 없는 문헌은 제외하였다. 셋째, 문헌에 범주정보인 분류번호가 부여되지 않은 문헌을 제외하였다. 넷째, 자동분류를 하기 위해서는 범주별로 학습에 필요한 적정 수의 문헌을 필요로 하는데, 이 실험에서는 매회 랜덤하게 추출된 학습집단을 사용하므로 범주별로 학습이 가능한 문헌이 포함될 수 있도록 KTSET의 전체 통계치를 고려하여 그 수를 30개로 정하고, 30개 미만인 문헌을 포함하는 최상위 범주는 실험대상에서 제외하였다. 마지막으로 복수의 분류번호가 주어진 문헌에 대해 첫 번째 분류번호를 제외한 나머지를 제거하였다. 이러한 규칙을 적용하여 구축된 실험문헌집단은 〈표 1〉과 같다. 실험문헌집단의 총 수는 421개이며, 범주 수는 5개이다.

이 연구에서는 CLTC를 위해 기계 번역기 또는 자동 번역기를 이용하였다. 즉 자동 번역기(소프트웨어)를 이용하여 실험대상 문헌을 한글에서 영문, 또는 영문에서 한글로 번역하였다. 이 연구에서 쓰인 자동 번역기로는 웹에

〈표 1〉 실험문헌집단의 범주와 범주별 문헌수

범주 기호	범주 명칭	문헌 수
C	컴퓨터 시스템	58
D	소프트웨어	94
H	정보시스템	124
I	계산방법론	113
X	정보학	32
합 계		421

2) 이 실험문헌집단은 김성혁 외(1993)에서 한국어 정보검색시스템과 자동색인기의 객관적인 성능 평가를 위하여 개발한 실험데이터 컬렉션이다.

서 쉽게 이용할 수 있는 Google 번역(G)과 L사의 일반번역 소프트웨어(L1)와 과학기술논문 전용 번역 소프트웨어(L2)를 이용하였다.

실험문헌집단을 CLTC의 관점에서 설명하면 다음과 같다. 먼저 KTSET에서 범주 정보인 분류기호와 한글 초록을 추출하여 국문 실험문헌집단을 만들었으며, 분류기호와 영문 초록을 추출하여 영문 실험문헌집단을 만들었다. 이렇게 만든 실험문헌집단에 대해 3가지 자동번역기를 사용하여 번역 실험문헌집단(총 6종)을 생성하였다. 즉 한글에서 영문으로 번역된 실험문헌집단인 영문 3종과 영문에서 한글로 번역된 실험집단인 국문 3종을 생성하였다. 따라서 이 연구의 실험문헌집단은 원문 언어로 된 실험문헌집단 2종을 더해 총 8종으로 구성하였다.

이 연구에서는 국문과 영문으로 된 문헌을 사용하므로 각각의 문헌을 대상으로 다음과 같이 자질을 추출하였다. 먼저 국문의 경우에는 주요 자질을 추출하기 위해 형태소 분석을 이용하였다. "21세기 세종계획"에서 제공되는 "지능형 형태소 분석기 2.0"을 이용하여 구문분석을 수행하고 일반명사, 고유명사, 영어 단어로 태깅된 자질을 추출하였다. 영문의 경우 어간 추출과정(stemming)을 거치고 불용어 목록(stopword list)과 비교하여 자질을 추출하였다. 특히 국문인 한글초록에 나타난 영어 단어도 어간 추출과정을 거쳤다. 이는 한글 초록에서 추출된 영문 자질의 경우 그 자체가 좋은 자질로써 가능성을 가지고 있으며, 영문초록에서도 동일한 자질이 추출 가능함으로 CLTC에서 상호이용이 가능하기 때문이다.

일반적인 문헌 자동분류 연구에서는 분류 성능에 영향을 미치지 않고 시스템의 효율성을 저하시키는 자질을 제거하는 자질선정을 수행한다. 하지만 이 연구에서 사용된 분류기인 SVM의 경우 자질 축소에 따른 별다른 성능 향상이 되지 않는 것으로 알려져 있으며, 고차원 자질 공간을 통제할 수 있는 능력이 있다(Joachims 1998; Taira and Haruno 1999). 따라서 이 연구에서는 별도의 자질 축소 실험을 수행하지 않고 전체 용어들을 자질로 선정하였다.

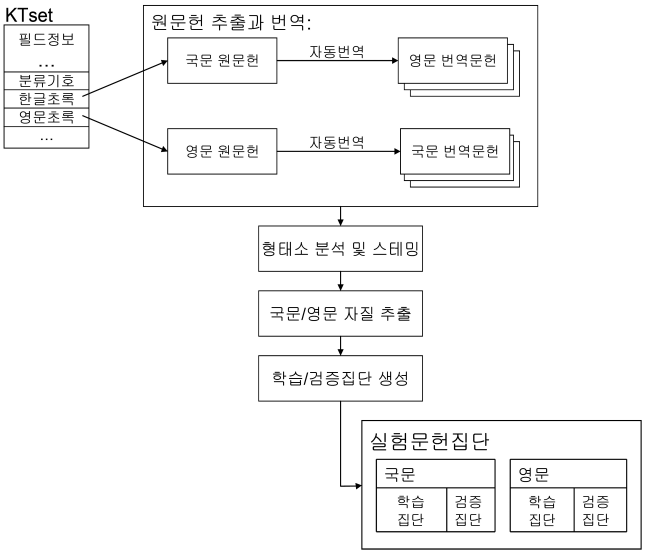
KTSET에서 CLTC에 필요한 실험문헌집단을 구축하는 전체적인 과정을 살펴보면 〈그림 1〉과 같다.

3.2 실험설계

이 연구는 단일어와 다양한 CLTC 기법을 적용한 자동분류 성능을 비교하고자 하였다. 특히 자동번역을 이용하였을 때 여러 CLTC 기법에 따라 문헌 자동분류에 어떠한 영향을 미치는지를 중점적으로 살펴보고자 하였다. 따라서 일반적인 자동분류 환경인 단일어 문헌 자동분류(MLTC) 실험을 기준 성능으로 설정하였다.

CLTC 기법으로는, 앞서 2장에서 기술하였듯이, 다국어 학습 방법과 교차언어 번역 방법을 적용하였다. 각각의 CLTC 방법에서 쓰인 구체적인 학습집단과 검증집단에 대한 종류와 관련 정보는 〈표 2〉와 같다.

먼저 MLTC의 경우 원문언어로 된 학습집단을 이용하여 분류기를 구축한 후, 원문언어



〈그림 1〉 CLTC용 실험문헌집단 구축과정

〈표 2〉 CLTC 유형에 따른 학습/검증집단 종류

유 형		언 어	학습집단(TR)	검증집단(TS)
단일어 자동분류 (MLTC)		국문(K)	TR_K	TS_K
		영문(E)	TR_E	TS_E
교차언어 자동분류 (CLTC)	다국어 학습(PT)	국문(K)+영문(E)	TR_K + TR_E	TS_K + TS_E
	학습집단번역 (TRT)	국문(K)	$TR_{E\to K}$(G) $TR_{E\to K}$(L1) $TR_{E\to K}$(L2)	TS_K
		영문(E)	$TR_{K\to E}$(G) $TR_{K\to E}$(L1) $TR_{K\to E}$(L2)	TS_E
	검증집단번역 (TST)	국문(K)	TR_K	$TS_{E\to K}$(G) $TS_{E\to K}$(L1) $TS_{E\to K}$(L2)
		영문(E)	TR_E	$TS_{K\to E}$(G) $TS_{K\to E}$(L1) $TS_{K\to E}$(L2)
	학습집단 혼합형1	국문(K)	$TR_{E\to K}$(G+L2)	TS_K
		영문(E)	$TR_{K\to E}$(G+L2)	TS_E
	학습집단 혼합형2	국문(K)	$TR_{E\to K}$(G+L1+L2)	TS_K
		영문(E)	$TR_{K\to E}$(G+L1+L2)	TS_E

* TR_K : 한글초록에서 추출된 문헌들로 구성된 학습집단
TS_E : 영문초록에서 추출된 문헌들로 구성된 검증집단
$TR_{E\to K}$(G) : 자동번역기(G)에 의해 영어에서 한국어로 번역된 학습집단
$TR_{E\to K}$(G+L2) : G와 L2의 번역된 문헌을 결합한 학습집단

로 된 검증집단에 대해 분류 성능을 시험하였다. 보다 최적화된 기준 성능을 제시하기 위해 다양한 단어빈도 가중치 공식과 역문헌빈도, 문헌길이 정규화 요소를 자질의 가중치로 적용하여 그 성능을 비교 분석하였다. 또한 실험문헌집단의 번역 없이 서로 다른 언어를 직접적으로 문헌 자동분류에 이용한 사전 실험도 수행하였다.

두 번째로 MLTC와 CLTC의 성능을 비교하기 위해 다국어 학습 방법을 적용하였다. 일반적으로 다국어 학습 방법은 두 개 언어 이상으로 이루어진 학습집단을 결합하여 하나의 학습집단을 형성하고 여러 언어로 이루어진 검증집단을 대상으로 자동분류 실험을 수행한다. 따라서 이 연구에서도 이와 같은 과정을 거쳤다.

번역을 이용한 CLTC 방법으로는 크게 2 가지로 나누어 자동분류 실험을 수행하였다. 하나는 학습집단이나 검증집단을 번역하는 방법(TRT와 TST)이고, 다른 하나는 2개 이상의 번역된 학습집단을 혼합하는 방법을 사용하였다. TRT 방법과 TST 방법은 자동 번역기에 번역 결과에 따라 각각 세 가지 하위 실험으로 나누어 총 6가지의 분류실험을 수행하였다. 이를 통해 두 방법 중 어느 방법이 더 좋은 성능을 가져오는지 비교 분석하였다. 또한 학습집단의 혼합형은 여러 번역기의 결과를 혼합한 방법으로 한 번역기를 사용한 것보다 더 좋은 결과를 가져올 수 있는지를 분석하였다.

각각의 실험에 대해 보다 정확한 분류실험을 위해 랜덤으로 학습집단과 검증집단을 추출하였다. 학습집단과 검증집단의 비율은 3:1이 되도록 하였으며, 랜덤 추출로 인한 오차를 줄이기 위해 모든 실험을 10회 반복 하였다. 또한 〈표 2〉에서 기술한 실험들에 대해 매 횟수마다 같은 학습집단과 검증집단을 이용하여 서로 비교평가가 가능하도록 하였다.

3.3 SVM 분류기

이 연구에서 사용한 SVM 분류기는 구조적 위험 최소화 원리를 이용하여 부정예제로부터 긍정예제를 분리해낼 수 있는 결정면을 찾아내는 분류모형이다(Vapnik 1995). 이 분류기는 기존의 다른 분류기에 비해 가장 좋은 성능을 보여주는 것으로 평가되고 있다(Joachims 1998; Yang and Liu 1999; Cristianini and Shawe-Taylor 2000).

SVM은 크게 선형 SVM과 비선형 SVM으로 나누고 비선형 SVM에서는 커널함수에 의해 만들어지는 비선형 결정함수를 이용하게 된다. 비선형 SVM의 기본적인 커널함수로는 다항식 커널함수, RBF 커널함수, sigmoid 등이 주로 사용된다(Vapnik 2000). 이러한 SVM 분류기를 사용하기 전에 해당 학습집단을 대상으로 몇 가지 결정해야 하는 파라미터가 있다. 그 이유는 이들 파라미터에 따라 분류 성능이 달라질 수 있기 때문이다. 이러한 파라미터 중 대표적인 것이 마진폭과 분류 오류 사이에 타협점을 찾아주는 패널티 파라미터 C와 커널함수의 파라미터이다. 따라서 사전 실험을 통해 이 연구에 사용된 학습집단의 적합한 파라미터를 구하였으며(C=128.0, γ=0.0004882, CV rate=69.20%), 이를 이후에 본 실험에 적용하였다. 또한 이러한 파라미터를 이용한 사전실험에서 문헌 자동분류에 주로 쓰이는 선형 SVM이 더 좋은 성능을 보여 이를 본 실험에

사용하였다. 이 연구에서 전반적인 실험을 위한 프로그램은 Python을 사용하였으며 SVM 분류기는 LIBSVM(Chang and Lin, 2001)을 사용하였다.

문헌 자동분류의 실험결과를 평가하기 위해 각 범주별로 정확도, 재현율, 정확률, 그리고 F1 척도를 사용하였다. 전체 범주의 성능을 평가하기 위해 평균 정확도, 평균 재현율, 평균 정확률, 마이크로평균 F1 척도와 매크로평균 F1 척도를 사용하였다. 정확도는 범주화 분할표에서 옳은 분류에 해당하는 문헌 수를 분류한 문헌 총수로 나눈 것을 의미하며, 각각의 정확도에 대해 평균을 낸 값이 평균 정확도이다. 평균 마이크로평균 F1 척도의 경우 개별 범주에 관계없이 전체 재현율과 정확률을 계산하여 이를 F1척도를 이용하여 계산하는 방법으로 고빈도 범주에 영향을 많이 받는 성능 평가 척도이며 범주화 성능평가를 위해 주로 사용되는 방법이다. 매크로평균 F1 척도의 경우 모든 범주에 대해 F1 척도를 구하고 이를 더한 다음 범주 수로 나누어 평균을 계산하는 방법으로 저빈도 범주에 영향을 많이 받는 성능평가 척도이다(Yang and Liu 1999).

4. CLTC를 이용한 분류 실험 결과

4.1 단일어 분류

이 연구에서는 KTSET에서 추출한 한글초록과 영문초록으로부터 얻어진 자질에 대해 자질값을 구하기 위해 여러 가지 용어 가중치 기법을 사용하였다. 이들을 사용해서 학습문헌과 검증문헌을 표현하고 SVM 분류기를 구축한 후 분류 성능을 확인하였다. 참고로 이 실험은 일반적인 자동 분류실험과 동일하다.

자질값을 표현하기 위해 용어 가중치로 단어빈도(TF), 역문헌빈도(IDF), 문헌길이 정규화 등 세 가지 요소의 결합으로 이루어진다. 이 연구에서도 자질값을 구하기 위해 이들 세 요소가 결합된 용어가중치를 이용하였다. 단어빈도 가중치는 이진 TF, 단순 TF, 로그 TF을 사용하였고, 역문헌빈도는 가장 흔히 사용되는 문헌집단 내 전체 문헌 수를 문헌빈도로 나눈 공식을 사용하였으며, 문헌길이 정규화 역시 가장 자주 사용되는 코사인 정규화 공식을 사용하였다. 각각의 단어빈도 가중치에 역문헌빈도와 문헌길이 정규화 가중치 요소를 적용하여 계산하였다. 적용된 공식은 〈표 3〉과 같다. 이

〈표 3〉 자질값을 위한 용어 가중치 공식

요 소		공 식
단어빈도	이진 TF	1(*if tf*〉0), 0
	단순 TF	*tf*
	로그 TF	1 + log tf
역문헌빈도		*log(N/df)*
코싸인 정규화		$\frac{w}{\sqrt{\sum_{i=1}^{n} w_i^2}}$

들 가중치를 한글초록과 영문초록 원문에 대해 적용하였을 때 한글/영어 실험문헌집단의 평균 분류 성능의 결과는 〈표 4〉에 제시하였다.

〈표 4〉에서 나타난 것과 같이 용어 가중치에 대해 두 언어 모두 로그TF를 사용할 때 가장 우수한 성능을 보였다. 한글 실험문헌집단에서는 로그 TF를 사용하였을 때 마이크로평균 F1이 77.14%로 가장 좋은 분류 성능을 보였으며, 단순 TF, 이진 TF 순으로 성능이 낮아졌다. 이러한 결과는 영어 실험문헌집단에서도 마찬가지였다. 따라서 추후에 모든 실험의 용어 가중치는 로그 TF와 역문헌빈도를 곱하고 이를 코사인 정규화 공식에 적용하였다. 또한 추후에 수행된 CLTC에 대해 로그TF를 사용한 한글/영어 분류 성능을 기준 성능으로 삼았다.

한글과 비교하여 영어 실험의 경우 단일어 분류의 성능이 매우 낮았다. 그 이유는 영어초록에서 추출한 실험집단의 자질 크기가 너무 작은 데서 기인하는 것으로 보인다. 실제 영어 원문의 실험집단의 총 고유 색인어 수는 2,430개로 전체 8개의 실험문헌집단 중에 가장 적은 수 이었으며, 총 색인어 수에서도 가장 적었다. 이는 영문초록을 만들 당시 한글을 영문으로 바꾸면서 한글초록 내용을 줄여 영문초록이 짧아진 것에서 기인하는 것으로 보인다. 따라서 이러한 자질의 감소는 SVM 분류기에서 성능의 저하로 이어진 것으로 보인다. 영어로 번역된 실험문헌집단은 번역기를 통해 축약된 내용을 부연 설명하거나 확장하여 더 많은 자질을 확보한 것으로 보인다.

다국어 학습 방법이나 번역에 기반하지 않고 분류기가 원문언어를 학습하고 대상언어를 분류하여도 적정 수준의 분류 성능을 얻을 수 있다. 그 이유는 두 언어 사이에 공유하는 용어들이 존재하기 때문이다. 이러한 용어는 대개 인명이나 지명 등과 같은 고유명사이거나, 약어 등 이다. 영어와 스페인어 사이에 번역 없이 교차언어 자동 분류의 경우 대략 11%의 성능을 가져왔다(Bel, Koster and Villegas 2003). 이 연구에서도 한글 학습문헌(TR_K)으로 분류기를 구축하고 영어 검증문헌(TS_E)을 분류한 실험과 그 반대의 실험을 수행하여 〈표 5〉와 같은 결과를 얻었다.

결과를 보면 번역과 같은 추가적인 작업 없이 한글 문헌으로 학습하여 분류기를 구축하고 이를 이용하여 영어 문헌을 분류한 결과, 마이크로평균 F1이 50.57%에 달했다. 또한 그 반대인 경우도 43.05%에 달했다. 이는 한글초록에서 나타난 영어 단어만을 이용하여 영어문헌을 분류하거나 그 반대로 영어문헌을 이용하여 한

〈표 4〉 자질값에 따른 단일어 분류의 평균 성능

	단어빈도	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
한글	이진 TF	88.50	71.00	66.02	71.24	67.24
	단순 TF	89.68	76.78	71.37	74.19	73.08
	로그 TF	90.86	80.72	71.52	77.14	73.72
영어	이진 TF	85.79	66.78	61.27	64.48	62.85
	단순 TF	87.66	70.20	65.73	69.14	66.86
	로그 TF	89.07	75.81	66.87	72.67	69.22

〈표 5〉 다른 언어를 이용한 단일어 분류의 평균 성능

TR / TS	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
TR_K / TS_E	80.23	43.40	43.19	50.57	40.49
TR_E / TS_K	77.22	52.25	36.13	43.05	35.92

글초록에 나타난 영어만을 이용하여 얻은 분류 성능이다. 이러한 원인은 앞의 선행연구와 달리 실험에 사용된 문헌이 학술 초록에 해당하므로 한글문헌에서 하나의 개념에 대해 두 언어의 단어를 이용하여 각각 나란히 표기하였기 때문이다. 실험문헌집단에서 한글초록에 나타난 한글과 영어의 색인어 자질을 구체적으로 살펴보면, 총 색인어 수 27,646개 중에 영어 색인어 수가 2,335개로 8.50%에 해당하였다. 하지만 고유 색인어 수로 살펴보면, 총 고유 색인어 수 3,340개 중에 영어 고유 색인어 수는 943으로 28.23%에 해당하였다. 그만큼 영어 색인어의 비중이 더 높아졌다고 볼 수 있다. 또한 대부분 이렇게 병기하는 단어는 핵심 단어이거나 중요한 개념을 포함할 가능성이 크므로 분류자질로서 매우 중요한 역할을 한다고 볼 수 있다.

4.2 교차언어 분류

4.2.1 다국어 학습 방법

다국어 학습 방법을 적용시키기 위해서는 분류에 사용되는 모든 언어가 범주가 태깅된 학습집단을 가지고 있어야 한다. 분류해야 할 검증문헌들이 서로 다른 언어로 구성되었을 때, 한 언어의 학습문헌만 이용하면, 분류 성능이 저하된다. 예를 들어 대부분이 영어로 된 학습집단에 대해 영어와 한글로 된 문헌을 분류한다면, 한글 자질에 대한 분류정보가 없기 때문에 한글문헌에 대한 정확한 분류가 어려워진다. 또한 각각의 언어마다 적정 수준 이상의 학습집단을 사용하면 보다 좋은 분류 성능을 얻을 수 있다. 이러한 배경에서 다국어 학습 방법이 필요하다. 이 연구에서는 한글과 영어 두 언어의 학습집단을 이용하여 분류기를 구축하고, 두 언어로 된 검증집단을 설정하여 CLTC의 다국어 학습 방법을 실험하였다. 그 결과 〈표 6〉을 얻었다.

실험 결과를 살펴보면, 다국어 학습 방법의 마이크로평균 F1 척도의 값이 기준 성능인 77.14%보다 2.66% 작은 74.48%로 비교적 우수한 성능을 보였다. 이 실험에서의 학습집단 대 검증집단의 비율은 다른 실험들과 동일하게 3:1로 하였으며, 학습집단과 검증집단에서 각각의 언어별 문헌 수도 1:1로 동일하게 하였다. 각 언어별 분류 성능은 한글문헌이 약간 우수하였는데, 이는 한글문헌에 출현한 영어 자질이 영어문헌의 출현한 자질의 부족함을 상쇄시킨 것으로 보인다.

〈표 6〉 다국어 학습 방법에 따른 평균 분류 성능

TR / TS	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
TR_K+TR_E / TS_K+TS_E	89.79	75.77	64.76	74.48	65.97

4.2.2 자동번역을 이용한 교차언어 학습 방법

특정 주제영역에서 문헌을 기계학습에 의한 자동분류를 수행하고자 할 때 가장 문제되는 부분이 학습집단의 유무이다. 즉 문헌에 적절한 주제범주를 부여한 학습문헌이 존재해야 한다. 이러한 작업은 대부분은 수작업에 의존해야 한다. 이러한 작업에 도움을 주기 위해 CLTC가 필요하다. 특히, 특정 주제영역에서 다른 언어로 된 태깅 정보가 존재한다면 대상언어로 변환 내지 번역을 통해 문헌을 자동분류 할 수 있다. 이러한 방법 중에 가장 쉽게 사용할 수 있는 방법이 바로 자동 번역기 또는 소프트웨어를 이용하는 방법이다.

실험 결과를 보다 정확히 분석하기 위해서는 이 실험에서 사용한 자동 번역기들의 성능평가에 대한 관련 연구나 해당 업체에서 제공하는 구체적인 정보가 필요하였다. 하지만 이 연구에서는 이러한 정보를 부족한 관계로 부득이하게 이 부분에 대한 분석은 생략하고 향후 연구 과제로 남겨 두고자 한다.

(1) 학습집단 번역 방법

학습집단 번역 방법은 자동 번역기를 이용하여 태깅이 존재하는 문헌집단(학습집단)을 대상언어로 번역한 후, 그 결과를 이용하여 분류기를 구축하고 대상언어의 문헌을 분류한다. 이 연구에서는 3 종의 자동 번역기를 이용하여 한국어와 영어로 된 학습문헌을 번역하고 그 결과에서 자질을 추출하였다. 이렇게 구축된 학습집단을 이용하여 분류기를 추구하고 원문언어로 된 검증집단을 자동분류하여 〈표 7〉과 같은 결과를 얻었다.

먼저 영어 학습집단을 한국어로 번역한 실험($TR_{E\rightarrow K}$ / TS_K)을 보면 G 번역기를 이용한 방법의 마이크로평균 F1 척도가 70.76%로 L사의 다른 번역기들(각각 67.84%, 68.05%) 보다 더 좋은 성능을 가져왔다. 하지만 단일어를 이용한 기준 성능인 77.14% 보다 6.38% 낮은 성능을 가져왔다. 매크로평균 F1 척도 역시 73.72% 보다 5.44% 낮은 68.28%의 성능을 보여 주었다. 또한 한국어 학습집단을 영어로 번역한 후 이를 이용한 실험($TR_{K\rightarrow E}$ / TS_E)에서는 L1 번역기를 이용하여 실험의 마이크로평균 F1 척도가 72.95%로 다른 번역기보다 약간 성능이 좋았다. 하지만 전자와 달리 단일어를 이용한 기준 성능인 72.67% 보다 0.28% 높아 약간 더 우수한 성능을 보였다. 반면, 매크로평균 F1 척도는 기준 성능의 69.22% 보다 3.4% 낮은 65.82%로 훨씬 낮은 성능을 보였다.

먼저 원문언어의 기준 성능 보다 자동 번역

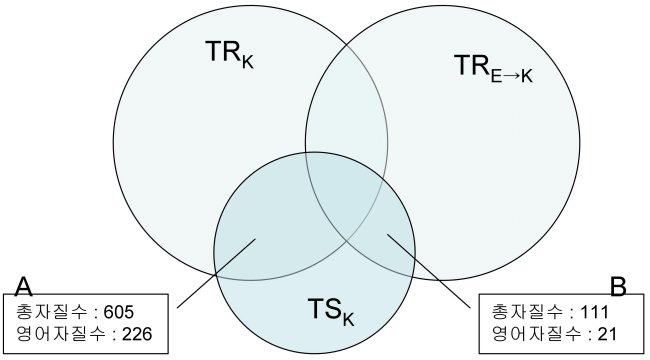
〈표 7〉 학습집단 번역을 이용한 CLTC의 평균 성능

TR / TS	번역기	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
$TR_{E\rightarrow K}$ / TS_K	G	88.30	75.22	65.89	70.76	68.28
	L1	87.81	74.27	62.43	69.52	64.79
	L2	88.00	74.53	62.60	70.00	64.68
$TR_{K\rightarrow E}$ / TS_E	G	89.07	75.51	64.52	72.67	65.94
	L1	89.18	78.70	64.17	72.95	65.82
	L2	88.65	75.77	63.19	71.62	65.28

기를 이용한 학습집단 번역 방식이 성능이 낮은 이유는 크게 두 가지로 분석할 수 있다. 하나는 자동 번역기를 통해 번역된 학습문헌의 자질 수가 원문언어의 자질 수 보다 적기 때문에 분류기가 분류시에 장애를 가진 것으로 보인다. 이러한 원인을 찾기 위해 TR_K과 $TR_{E \rightarrow K}$의 자질집합이 TS_K의 자질집합과 어느 정도 일치되는지를 분석하여 〈그림 2〉와 같은 결과를 얻었다. 그림에서 보면 TR_K와 TS_K의 공통 자질집합에서 $TR_{E \rightarrow K}$의 자질집합을 뺀 'A' 부분의 자질집합의 총 자질 수는 605개로 'B' 부분의 총 자질수의 대략 1/6배 정도 되는 것으로 나타났다. 이것은 영문초록이 국문보다 짧은 실험문헌집단의 특성에서 기인한 것으로 보인다.

다른 하나는 번역을 통한 잡음(noise)이 생겨 자질의 질이 저하되었고 이것이 원문언어의 검증문헌 분류시에 반영된 것으로 보인다. 이는 〈그림 2〉에서 B 부분에 해당하는 자질들을 자세히 살펴보면 알 수 있다. 이 부분에서는 '전략', '전환', '유통'과 같이 주제적 특성이 없는 자질들을 많이 발견할 수 있었다. 이러한 문제점을 해결하기 위해서는 원문 언어에 있는 자질에 대해 적절한 전문용어로의 번역이 필요하다.

이러한 특징은 학습집단 번역 방법에서 자동번역기 사이에 성능 차이의 주된 이유이다. 즉 각 번역기가 영문에서 한글로 번역할 때 전문용어에 속하는 중요한 자질을 그냥 영어 용어로 남겨두거나 정확히 번역해야 함에도 불구하고 그렇지 못하거나 더 나아가 일반용어로 번역하는 오류를 보이기 때문이다. 또는 일반 영어 용어에 대해 한글 용어로 번역을 하지 않는 경우도 또 다른 이유에 해당한다. 실례로 "network(10), net(12), unix(19), system(23), inform(9), 내포(11), 영상(20), 질의(23), 데이터(106)" 등을 들 수 있다. 이와 같은 단어는 DF가 비교적 높으면서, 오직 L1과 L2 번역기에서 많이 출현하였다. 이들 단어는 범주와 상관없이 출현하면서 분류기의 성능을 저하시킬 수 있다. 한글에 나타난 영어 단어는 당연히 중요한 분류자질로서 역할을 해야 하는데 반해, 이 예의 영어 단어는 한글 문헌에서는 대부분 이미 한글화가 되었거나 보편적으로 사용되는 용어로 중요한 자질이 되지 못한다.



〈그림 2〉 학습집단과 검증집단의 공통 자질집합 정보

영어 실험에서는 단일어 분류의 기준 성능과 비슷하거나 약간 우수한 경향을 보였다. 이는 앞서 설명하였듯이 영어 실험문헌집단의 색인어 자질이 너무 적은 것에서 기인하는 것으로 보인다. 실제 고유 색인어 수에서 있어서 영어 원문 실험집단은 2,430개로 G 번역기를 통해 얻어진 실험집단의 3,218개에 비하면 대략 75% 수준으로 이는 고차원 자질을 선호하는 SVM 분류기의 성능 저하를 가져올 수 있다.

(2) 검증집단 번역 방법

문헌 자동분류에서 학습집단을 통해 구축된 분류기는 검증집단에 포함된 문헌을 분류하게 되는데, 실제 환경에서는 학습집단을 제외한 모든 미분류 문헌은 검증집단이 될 수 있다. 이러한 측면에서 본다면 학습집단 보다 검증집단의 규모가 훨씬 크다고 볼 수 있다. 많은 컴퓨팅 자원을 필요로 하는 자동번역을 이용하여 큰 규모의 검증집단을 번역하는 것은 비효율이다. 하지만 학습집단 번역 방법과 검증집단 번역 방법이 성능상의 차이가 있는지 확인해볼 가치는 있다. 검증집단 번역 방법도 3가지의 번역기를 이용하여 번역된 검증집단을 만들고, 이들을 원문 언어 학습집단으로 구축한 분류기를 이용하여 분류하였다. 그 성능은 〈표 8〉에 제시하였다.

실험 결과를 살펴보면, 한국어의 경우 G 번역기를 이용한 검증집단 번역 방식의 마이크로평균 F1이 68.86%로 단일어의 기준 성능에 비해 8.28%로 크게 저하된 것으로 나타났다. 이러한 저하는 매크로평균 F1척도에서도 같은 모습을 보였다. 하지만 영어의 경우에는 학습집단 번역 방법의 결과처럼 단일어 기준 성능에 근접하는 모습을 보였다. 특히 G 번역기를 이용한 검증집단 번역 방식의 매크로평균 F1은 67.11%로 MLTC를 제외하고 가장 우수한 분류 성능을 보였다.

(3) 번역결과 혼합 방법

자동번역기의 경우 번역을 위해 사용한 사전의 종류나 알고리즘의 차이에 따라 다양한 번역 결과를 가져올 수 있다. 이는 자동 번역기의 번역 결과로부터 추출된 자질이 다양함을 뜻한다. 따라서 이 연구에서는 여러 번역기에 의해 번역된 학습집단을 2개 또는 3개를 혼합하여 새로운 학습집단을 만들고 이를 이용하여 분류기를 구축하고 원문언어로 된 검증집단을 분류하였다. 그 결과 〈표 9〉를 얻었다.

〈표 8〉 검증집단 번역을 이용한 CLTC의 평균 성능

TR / TS	번역기	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
TR_K / $TS_{E\to K}$	G	87.54	73.50	62.98	68.86	65.22
	L1	85.71	70.49	56.46	64.29	58.46
	L2	87.54	73.30	61.15	68.86	62.79
TR_E / $TS_{K\to E}$	G	88.80	75.48	65.21	72.00	67.11
	L1	87.62	73.51	62.25	69.05	64.54
	L2	87.47	71.38	60.93	68.67	62.63

〈표 9〉 학습집단 번역결과를 혼합한 CLTC의 평균 성능

TR / TS	평균정확도	평균정확률	평균재현율	MicroAvg F1	MacroAvg F1
$TR_{E\to K}$(G+L2) / TS_K	88.15	74.60	63.37	70.38	65.59
$TR_{K\to E}$(G+L2) / TS_E	89.26	78.03	64.26	73.14	66.08
$TR_{E\to K}$(G+L1+L2) / TS_K	88.04	75.66	62.34	70.10	64.71
$TR_{K\to E}$(G+L1+L2) / TS_E	89.07	77.35	63.44	72.67	65.06

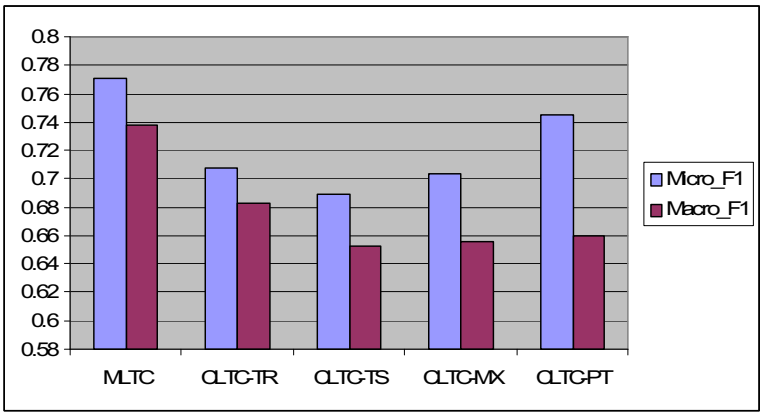
실험 결과를 보면, 학습집단 번역 방식에 좋은 성능을 보인 두 번역기 G와 L2을 적용한 첫 번째 한글 실험은 70.38%로 기준 성능에 비해 여전히 낮은 성능을 보였다. 하지만 같은 두 번역기의 영어 실험은 73.14%로 기준 성능인 72.67%보다 약간 우수한 결과를 보였다. 이 결과는 영어로 대상으로 한 분류실험 중에 가장 좋은 성능이다.

두 언어에서 세 번역기를 사용한 실험들은 두 번역기를 사용한 것보다 모두 약간 낮은 성능을 보였다. 이는 세 번역기를 통해 자질의 크기가 커졌지만, 동시에 범주별로 구분되어야 분포해야 할 자질들이 서로 중복되어 오히려 분류 성능을 저하시킨 것으로 보인다.

4.3 CLTC 방법의 성능 비교

이 연구에서는 단일어 분류실험을 비롯하여 다양한 CLTC의 방법을 동일한 조건에서 수행하였다. 동일한 조건의 의미는 문헌 자동분류 실험에 주로 거론되는 실험변수들을 동일하게 하였다는 뜻이다.

한글과 관련된 다양한 CLTC 방법 중에 가장 좋은 마이크로평균 F1과 매크로평균 F1을 보인 방법을 살펴보면 〈그림 3〉과 같다. 전체적으로 볼 때 CLTC의 분류 성능이 단일어(MLTC) 보다 낮았지만, CLTC 방법 중에서는 다국어 학습 방법(CLTC-PT)이 가장 좋은 성능을 보였으며, 그 다음으로 학습집단 번



〈그림 3〉 한글 CLTC 방법의 분류 성능 비교

역 방법(CLTC_TR), 학습집단 번역 혼합 방법(CLTC_MX), 검증집단 번역 방법(CLTC_TS) 순으로 성능이 저하되었다.

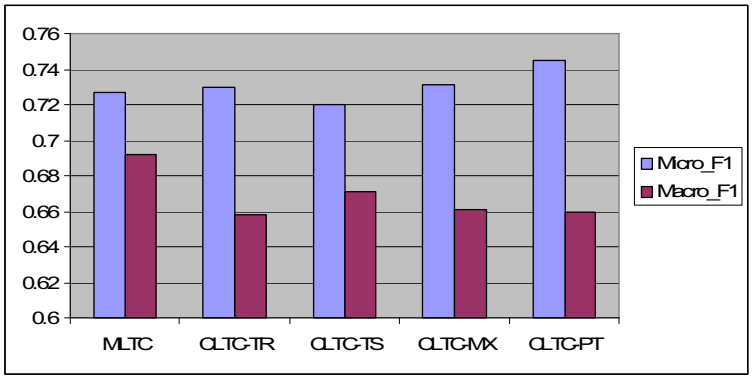
CLTC의 번역을 이용한 방법 중에는 학습집단 번역 방법이 다른 방법에 비해 마이크로평균 F1과 매크로평균 F1 평가 척도에서 좋은 성능을 보여 주었다. 이는 CLTC 방법 중에서는 가장 높은 이용 가능성을 보이는 것으로 해석할 수 있다. 왜냐하면 학습집단 번역 방법이 기계번역 측면에서 효율적이고 일반적인 CLTC 환경에서 쉽게 적용할 수 있는 장점을 갖고 있기 때문이다. 또한 이 방법이 다른 방법에 비해 비교적 좋은 분류 성능을 보이기 때문이다.

영어 실험의 CLTC 방법의 분류 성능을 방법별로 살펴보면, 〈그림 4〉와 같다. 영어의 경우 한글과 달리 CLTC 방법의 분류 성능이 대부분 단일어 성능과 비슷하거나 더 좋았다. 이는 앞서 기술하였듯이 전체 자질의 수에서 기인하는 것이다. 다만 각 방법별로 살펴보았을 때, 한글과 달리 CLTC-MX가 CLTC-TR과 CLTC-TS보다 약간 우수한 분류 성능을 보였다. 이는 CLTC-MX가 자질의 수가 적은 소규모의 실험문헌집단을 대상으로 서로 다른 성격의 번역기에서 생성된 자질의 혼합 또는 데이터 결합을 통해 더 좋은 성능을 얻을 수 있음을 보여 주었다.

5. 결 론

이 연구에서는 기계번역을 이용하여 다양한 자동분류용 실험문헌집단을 만들고, 여러 가지 CLTC 방법을 적용하여 단일어 자동분류의 성능과 비교하였다. 또한 기계번역의 결과를 혼합하여 성능의 개선이 가능한지 실험하였다. 실험문헌집단은 KTSET으로부터 한영초록과 범주정보인 분류기호를 추출하였으며, 분류기는 SVM 분류기를 사용하였다. CLTC 실험을 위해 3 종의 자동 번역기를 사용하여 한영 또는 영한 번역을 수행하여 6종의 실험문헌집단을 추가하였다. 이 연구를 통해 밝혀진 사실은 다음과 같다.

첫째, 기존 연구와 같이 단일어를 이용한 문헌 자동 분류방법이 CLTC보다 더 좋은 성능



〈그림 4〉 영어 CLTC 방법의 분류 성능 비교

을 보여 주었다. CLTC 내에서는 다국어 학습 방법이 가장 좋은 성능을 가져왔다. 그 다음으로 학습집단 번역 방법, 학습집단 번역 혼합 방법 순으로 나타났다. 검증집단 번역 방법이 가장 낮은 성능을 보였다.

둘째, 학습집단 번역 방법이 기계번역 측면에서 효율적이고 일반적인 환경에 쉽게 적용할 수 있으면서 분류 성능도 비교적 좋아 CLTC 방법 중에서는 가장 높은 이용 가능성을 보였다.

셋째, 기계번역을 이용하여 CTLC를 수행하였을 때 번역과정에서 발행한 자질 축소나 주제적 특성이 없는 자질로의 번역으로 인해 성능저하가 나타났다. 이러한 문제를 방지하기 위해 전문 용어사전과 같은 적절한 어휘자원의 사용을 통한 번역을 고려해야 한다. 또한 자질이 수가 적은 소규모의 실험문헌집단을 대상으로 이러한 방법을 적용할 경우, 서로 다른 성격의 기계 번역기를 사용하여 그 결과를 혼합하거나 데이터 결합을 통해 더 좋은 성능을 얻을 수 있었다.

넷째, 분류의 대상이 되는 문헌에 국문과 외국어가 병기되는 경우, 외국어가 중요한 자질이라면 번역을 이용한 CLTC에 많은 영향을 주는 것으로 파악되었다.

이 연구의 제한점은 다음과 같다. 현재 국내 번역기의 평가와 장단점에 대한 구체적인 정보가 부재한 상황으로, 부득이 하게 이 연구에서는 기계 번역기의 직접적인 성능 분석에 따른 깊이 있는 분석을 생략하였다. 또한 향후 과제로는, 우선 이 연구에서 수행한 CLTC 방법을 다른 유형의 문헌 환경과 더 많은 실험문헌집단에 적용하여 검증할 필요가 있다. 또한 자동 번역기의 성능평가 뿐만 아니라 일반사전이나 전문 용어사전 등을 포함한 어휘자원을 이용한 CLTC 방법을 적용할 필요가 있다. 이는 CLTC의 최적화와 범용화를 위해 반드시 필요하다. 마지막으로 이 연구에서는 SVM 분류기를 사용하였는데, 다른 분류기를 적용하여 CLTC에 미치는 영향을 분석할 필요가 있다.

〈감사의 글〉

이 연구에 필요한 실험문헌집단의 한영 또는 영한 번역을 위해 자동번역기를 제공하여 주신 L사에 감사를 드립니다.

참 고 문 헌

[1] 김성혁, 서은경, 이원규, 김명철, 김영환, 김재군. 1994. 자동색인기 성능시험을 위한 Test Set 개발. 『정보관리학회지』, 11(1): 81-102.

[2] Adeva, J., R. Calvo, and D. L. Ipiña. 2005. "Multilingual Approaches to Text Categorisation." *The European Journal for the Informatics Professional*, 6(3): 43-51.

[3] Amine, B. M., and M. Mimoun. 2007. "Word-Net based Cross-Language Text Categorization." *ACS International Conference on Computer Systems and Applications*, 848-855.

[4] Bel, N., C. Koster, and M. Villegas. 2003. "Cross-Lingual Text Categorization." *LNCS*, 2769:

126-139.

[5] Chang, C. and C. Lin. 2001. "LIBSVM : a library for support vector machines." [online]. [cited 2008.08.30]. 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.

[6] Cristianini, N., and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. London: Cambridge University Press.

[7] Gliozzo, A. M., and C. Strapparava. 2005. "Cross language text categorization by acquiring multilingual domain models from comparable corpora." *Proceedings of the ACL workshop on building and using parallel texts*, 9-16.

[8] Joachims, T. 1998. "Text categorization with Support Vector Machines: Learning with many relevant features." *Proceedings of the 10th European Conference on Machine Learning*, 137-142.

[9] Kishida, K. 2005. "Technical issues of crosslanguage information retrieval: a review." *Information Processing & Management*, 41: 433-455.

[10] Melo, G. and S. Siersdorfer. 2007. "Multilingual text classification using ontologies." *Proceeding 29th European Conference on Information Retrieval*, 541-548.

[11] Oard, D. W., and A. R. Diekema. 1998. "Crosslanguage information retrieval." *Annual Review of Information Science and Technology*, 33: 223-256.

[12] Peters, C., and P. Sheridan. 2001. "Multilingual information access." *Lectures on information retrieval*, 51-80.

[13] Rigutini, L., M. Maggini, and B. Liu. 2005. "An EM based training algorithm for Cross-Language Text Categorization." *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 529-535.

[14] Taira, H., and M. Haruno. 1999. "Feature selection in SVM text categorization." *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, 480-486.

[15] Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

[16] Vapnik, V. N. 2000. *The nature of statistical learning theory*. 2nd ed. New York: Springer.

[17] Wu, K. and B. Lu. 2008. "A Refinement Framework for Cross Language Text Categorization." *Information Retrieval Technology 4th Asia Information Retrieval Symposium*, 15-18.

[18] Yang, Y., and X. Liu. 1999. "A re-examination of text categorization methods." *Proceedings of the ACM SIGIR Conference on Research and Development in International Retrieval (SIGIR'99)*, 42-49.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Sung-Hyuk Kim, Eun-Gyoung Seo, Won-Gyu Lee, Myung-Cheol Kim, Young-Whan Kim, and Jae-Kun Kim. 1994. "A Development of the Test Set for Estimating the Retrieval Performance of an Automatic Indexer." *Journal of the Korean Society for Information Management*, 11(1): 81-102.