중한 이메일 자동번역시스템

김운○ 권오욱 오영순 김영길
한국전자통신연구원 언어처리연구팀
wkim1019@etri.re.kr, ohwoog@etri.re.kr, suni@etri.re.kr, kimyk@etri.re.kr

A Chinese-Korean E-Mail Translation System

Yun Jin○, Oh-Woog Kwon, Ying-Sun Wu, Young-Kil Kim
ETRI Natural Language Processing Research Team

요 약

본 논문에서는 중국어의 이메일 특성을 이용한 중한 대화체 자동번역 방법에 대하여 기술한다. 본 논문에서는 중국어와 한국어와 같이 언어 간의 어순이 다르고 이메일과 같이 특정한 도메인의 언어적 자원도 제한적인 특성을 고려하여 중국어 이메일 특성을 이용한 규칙 기반의 번역 방법을 시도하였다. 이를 위해, 본 논문에서는 중국어의 굳어진 표현이 많고, 한글자 단어 많으며, 입력 오류 많고, 청유 및 경어가 많은 이메일 특성 분석을 통해 그에 대응되는 처리 방법을 제안하였다. 그리고, 그 방법의 타당성을 증명하기 위해 규칙기반의 중한 뉴스 자동번역 시스템과 비교 실험을 하였으며, 규칙기반과 통계적 방법의 타당성 실험을 위해 Gmail과도 비교 실험을 하였다. 두 가지 비교 실험 결과, 본 논문에서 접근한 방법이 모두 우수하였으며, 그 타당성을 증명하였다.

주제어: 이메일, 자동번역, 중한

1. 서 론

최근 들어, 인터넷과 같은 통신수단의 발달로 이메일은 이미 오늘날 사람들에게 없어서는 안 될 중요한 교류수단이 되었다. 하지만 언어적 장벽은 이종언어 간의 이메일 교류를 저애하는 걸림돌이 되고 있다. 그 대안으로 이메일 자동번역이 자동번역 연구 분야에서 가장 주목받는 분야로 떠오르고 있다. 구글은 통계적 방법을 이용하여 42개 언어에 대한 임의의 양방향 번역이 가능한 Gmail 번역 서비스[1]를 진행하고 있으며, [2]에서는 일본어 문맥을 이용한 이메일 번역을 시도하였다.

이메일 문서는 개개인에 의해 작성된 문서로써 개인의 성향이 강하게 반영되어 있어 표현이 자유로우며 소재가 다양한 특징을 지니고 있다. 따라서 기술문서와 뉴스에 비해 비문법적인 문장이 흔히 포함되어 있다. 예를 들면, 불필요한 문장부호를 사용하거나, 오타, 약어 등을 많이 포함되어 있다. 이런 특징은 이메일 자동번역 품질을 저애하는 문제점으로 되고 있다. 또한 이메일은 개개인의 의사가 글로 반영되어 있고 상대방에게 자신의 의도와 같은 정보를 전달하는 것이 특징적이다. 이런 특징은 이메일 자동번역에 있어서 가장 중요한 도전으로써 이메일 작성한 사람의 의도를 상대방에게 정확한 전달해야 한다.

사용자 입장에서의 문제점은 얼마만큼의 번역 성능에 도달해야 이종언어 간의 이메일 교류가 가능하며, 언어적 특징에 따른 번역 성능이 현재 어느 정도까지 도달하였는지에 관한 것이지만, 이메일 자동번역 연구 분야에서는 특정 언어 간 해결해야 할 이슈는 무엇이며, 어떤 접근 방법을 필요로 하며 그 해결책은 무엇이며, 앞으로 해결해야 할 과제는 무엇인지가 관심사가 될 것이다.

본 논문에서는 중국어 이메일 특성을 이용한 중한 대화체 자동번역시스템 구축방법에 대해 소개하고자 한다. 이를 위해, 본 논문에서는 우선, 중국어 이메일을 수집하여 그 특성을 분석하였으며, 다음으로, 분석된 특성을 이용하여 규칙기반의 중한 대화체 자동번역 시스템을 구축하였다. 마지막으로, 구축된 시스템의 성능 비교를 평가하기 위해 중한 뉴스 자동번역 시스템과 수동 평가를 실시하였으며, 중한 대화체 자동번역에 있어서 규칙기반의 방법론과 통계기반의 방법론과의 성능 비교를 위해 구글의 Gmail 번역과도 성능비교를 실시하였다.

본 논문의 구성은 다음과 같다. 2절에서는 규칙기반의 중한 대화체 자동번역 시스템에 대하여 간략하게 소개하며, 3절에서는 중국어 이메일의 4가지 특성에 대하여 기술하며, 4절에서는 분석된 특성에 대한 처리 방법을 기술한다. 그리고, 5절에서는 제안한 방법에 대한 타당성을 증명하기 위한 실험에 대하여 기술하며, 마지막 6절에서는 결론을 맺으며 향후 연구를 소개한다.

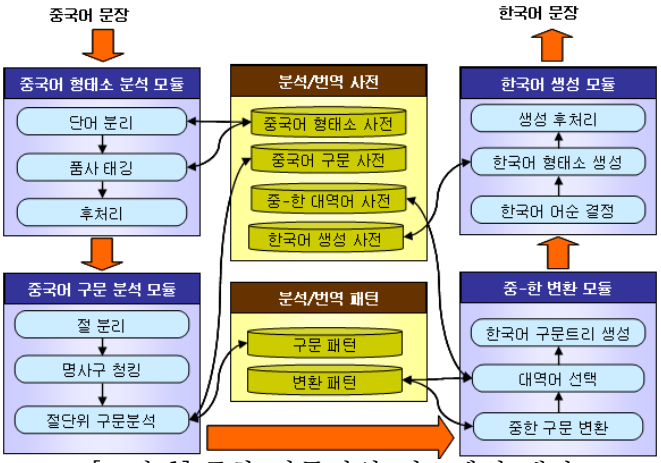
2. 중한 대화체 자동번역 시스템

본 절에서는 중국어 이메일 특성을 이용한 중국어 대화체 자동번역 방법의 이해를 돕기 위해 규칙기반의 중한 대화체 자동번역 시스템 대하여 간략하게 소개한다.

본 중한 대화체 자동번역 시스템은 변환방식(transfer approach)으로 설계된 규칙기반 자동번역 방법으로 구현되었다. 또한, 본 중한 자동번역 시스템은 절 단위로 자동 번역하는 것과 중국어 구문구조와 한국어 구문구조의 차이를 변환패턴이라고 하는 지식 기반으로 해결하는 것을 특징으로 한다.

본 중한 자동번역 시스템은 [그림 1]과 같이 중국어 형태소 분석 모듈, 중국어 구문 분석 모듈, 중한 변환 모

듈과 한국어 생성 모듈로 구성되어 있다.



[그림 1] 중한 자동번역 시스템의 개략도

중국어 형태소 분석 모듈은 중국어 단어분리, 형태소 품사 태깅와 후처리 단계로 나눌 수 있다. 단어분리 단계에서는 띄어쓰기가 없는 중국어 문장에서 사전기반 최장일치방법과 중국어 단어 빈도를 이용한 방법으로 단어를 분리한다. 형태소 품사 태깅은 bi-gram HMM 모델에 기반하여 구축되어 있으며, 후처리 단계에서는 품사 태깅의 오류에 대한 오류 수정규칙 기반의 방법을 이용한다.

중국어 문장들이 대부분 복문인 경우가 많고 이들 관계가 병렬관계가 대부분이어서 구문분석 복잡도를 줄이고 구문분석 정확도를 향상을 위하여, 본 중한 자동번역 시스템은 중국어 형태소 분석 결과를 이용하여 단문으로 분리한다. 중국어의 특성 상, 절 분리는 문장기호인 콤마와 콜론 등을 기준으로 분리하면 거의 정확하다 하지만, 그렇지 않은 경우에 대한 규칙을 설정하여 절분리를 수행한다. 분리된 절단위로, 형태소 분석 결과를 바탕으로 명사구 청킹을 한 후, 구구조 기반의 절단위 구문분석을 수행한다.

중한 변환 모델에서는 중국어 구구조 구문트리를 한국어 의존트리로 변환하며 관련 대역어를 선택하여 한국어 의존 구문트리를 생성한다. 일반적인 중국어 구구조 구문트리가 한국어 의존 구문트리로 변환되지 않는 불규칙적인 구문변환이나 대역어 선택에 애매성이 있는 경우를 해결하기 위하여 변환패턴이라는 번역지식을 이용한다. 변환패턴은 특정 중국어를 키(key)로 가지며, 불규칙 변환정보는 중국어 구문 표현부와 불규칙적으로 변환해야 할 한국어 구문 표현부로 구성된다. 중국어/한국어 구문 표현부는 키에 해당하는 단어를 중심으로 하여 구문노드들과 구문관계를 구문태깅방법(syntactic annotation)을 표현한다. 구문노드는 구문자질 이외에 구문노드의 헤드에 대한 의미정보, 어근정보, 어휘정보 등으로 표현하여 특정한 문맥을 표현할 수 있도록 하였다.

중한 구문 변환 시에, 변환패턴을 참조하여 현재 변환하고자 하는 중국어 구문트리가 변환패턴의 중국어 구문 표현부와 일치하면, 규칙적인 한국어 구문트리로 변환하지 않고 변환패턴에 기술된 한국어 구문 표현으로 트리를 변환한다. 대역어 선택의 애매성이 있는 단어에 대해서도 디폴트 대역어로 가지 않는 문맥을 중국어 구문 표현부로 설정하고 이에 해당되는 한국어 대역어를 한국어 구문 표현부에서 설정하여 대역어 애매성을 해소한다.

한국어 생성 모듈에서는 중한 변환 모듈에서 생성된 한국어 구문트리와 대역어를 이용하여 한국어 어순으로 정렬하고 어미/조사 등에 대한 형태소를 생성하여 한국어 문장을 생성한다. 한국어 생성 모듈의 생성 후처리 과정은 한국어 생성에서 자주 틀리는 오류는 생성 후처리 규칙 DB를 구축하여 오류를 수정한다. 한국어 생성 모듈에서는 부사 위치 선택, 중국어에서 빈번하게 나타나는 양사에 대한 생성 여부 판단, 중국어에서 잘 나타나지 않는 시제에 대한 처리와 분리된 절로 생성된 절간의 연결어미 생성이 주요 문제이며 이를 해결한다.

3. 중국어 이메일 특성 분석

본 절에서는 서론에서 소개한 일반 이메일의 특성을 포함한 중국어 이메일에서 나타나는 4가지 특성 분석에 대해 기술하고자 한다. 주요한 목적은 이런 특성 분석을 통해 중국어 대화체 자동번역의 성능향상을 도모하려는 데 있다.

3.1 굳어진 관용 표현이 많음

뉴스 또는 기술문서는 대부분 구문적으로 정확하게 씌어져 있으며, 구문변환을 거쳐 단어 단위의 대역어로 번역하여도 정확한 번역이 가능하지만, 중국어 이메일에는 중국어 고유의 굳어진 관용어가 많으며, 표현이 다양한 특성을 갖고 있다. 특히, 인사말과 맺음말의 굳어진 표현은 일상적인 이메일뿐만 아니라 비즈니스용 이메일에도 자주 나타나며 이러한 관용적인 표현은 이메일을 주고받는 상대와의 관계, 친분 및 서열에 따라 그 표현이 다양하고 격식도 [표 1]과 같이 한결 자유롭다.

[표 1] 굳어진 중국어 관용적 표현

분류	중국어 예문	한국어 뜻
인사말	好久不见了，你都还好吧？	오래동안 못 보았네요, 당신은 아직 잘 지내죠?
	最近过得怎么样？	최근 어떻게 지내시는지요?
	承蒙关照！	덕분에 잘 지냈습니다.
맺음말	祝好运！	행운을 빕니다.
	祝身体健康！工作愉快！	신체건강하시고, 하시는 일 잘 되시길!
	顺颂 商祺！	축하합니다.

이런 표현 방법의 가장 큰 특징은 표현이 중국어로 굳어져 규칙기반의 번역시스템에서는 정확한 번역을 생성하기 힘들다. 예를 들면, “承蒙关照！”와 같은 경우, 그

대로 번역하면, “관심을 받는다”가 되지만, 실제 의미는 이와 조금 다른 “덕분에 잘 지냈습니다.”가 된다.

3.2 모호성이 많은 한 글자 사용이 빈번

중국어 이메일에는 한 글자 단어 사용이 빈번한 특징을 가지고 있다. 한 글자 단어에는 동사나 명사가 많으며, 그 외에도 호칭어, 방위사, 접속사, 어기조사 등 모호성이 강한 단어들도 포함되어 있다.

[표 2] 한 글자 단어 사용 예문

번호	중국어 예문	한국어 뜻
(가)	你最好上QQ吧，	당신은 QQ에 로그인 하는게 좋겠네요.
(나)	你们元旦放几天假?	설에 몇일 쉬나요?
(다)	所有人脚都起了泡。	모든 사람의 발은 이미 물집이 생겼다.

[표 2]의 예문 중 (가)는 단음절 동사가 포함된 문장으로 이 문장 중 “上”은 동사로도 쓰일 뿐만 아니라 방위사(“위”)로도 쓰이며, 관사(“지난”)로도 쓰인다. 또한, 동사로 사용될 경우, 목적어에 따라 여러 가지 의미를 가진다. 예를 들면, “<산>에 오르다”, “<학교>에 가다”등으로 다양하다. 이 예문에서는 “<QQ>에 로그인하다”의 뜻을 가진다. 예문 (나)와 (다)는 “동사+목적어”로 구성된 두 글자 단어(‘放假’, ‘起泡’) 사이에 어기조사나 수량사를 추가한 경우이다. 이와 같이 동사와 목적어 사이에 다른 문장성분이 삽입되어 한 글자 동사와 한 글자 명사로 각각 분리되면서 의미 모호성이 생겨난다. 예를 들면, “假”의 명사적 의미에는 “휴가”이 외에 “가짜”라는 뜻도 가지고 있다.

사람과 사람사이의 의사소통 정보 교환 등을 하는 만큼 중국어 이메일에는 중국어 특색의 다양한 한 글자 호칭이 많이 등장하고, 한 글자 직함도 자주 쓰인다.

[표 3] 다양한 호칭 예문

번호	중국어 예문	한국어 뜻
(가)	徐总	서총경리님
(나)	李处、刘工	이처장님, 유 공정사님
(다)	小新	신아
(라)	顺顺	순아

예문 (가)에서 '徐总'은 성씨 '徐(서)'와 직함 '总经理(총경리)'의 줄임말인 '总'이 결합되어 '서총경리'를 지칭한 것이고, 예문 (나)는 '李处', '刘工'은 각각 '李+处长(이처장)','刘+工程师(유 공정사)'를 지칭하는 말이며, 예문 (다)에서 '小新'은 '小[1]'와 이름 마지막 글자 '新'가 결합하여 이루어진 호칭이다. 또한 예문 (라)는 ‘顺顺’과 같이 이름 글자가 겹친 호칭도 자주 발생한다. 따라서, 한 글자 단어에 대한 모호성 처리가 필요하며, 한 글자 호칭에 대한 주격처리 또는 경어처리가 필요하다.

3.3 비구문적 입력 오류 많음

중국어 이메일에는 비구문적인 입력 오류가 많은데, 이런 오류에는 문장부호 또는 목적어와 같은 구성 성분을 생략하거나, 불필요한 문장성분을 추가하거나, 구조조사 오용 및 타이핑 오류 등이 포함되어 있다.

[표 4]는 중국어 이메일에 나타나는 비구문적 입력 오류에 관한 예문들이다. 그 중 (가)는 문장 중 목적어가 생략된 경우로써, 구문적 파싱 오류를 유발할 뿐만 아니라, 관련 동사 “养(요양하다)”의 모호성[2](“기르다”, “출산하다” 등 의미가 있음.)으로 인해 부정확한 번역을 유발하게 된다. 예문 (나)는 단어 “可是(정말)”를 중복하여 사용한 예문이다. 이 예문에서 중복 사용한 단어는 접속사 “그러나”의 뜻으로 많이 쓰이지만, 이 예문처럼 드물게 “정말”의 뜻으로 쓰이기도 한다. 마지막으로 예문 (다)는 구조조사를 오용한 예문이다. 구조조사는 중국어에서 2개 이상의 단어 및 구 사이에서 보조적 구성성분과 중심어 사이의 구조적 관계를 표시하는 것으로 “的”, “地”, “得” 등 3가지가 있다[3]. 이런 구조조사는 음독이 같아 흔히 혼용하는 오류를 법하며, 중국어 교육용 코퍼스에서도 나타나는 오류 중 하나이다[4]. 본 예문에서는 “得” 대신 “的"를 오용하였다. 예문 (라)는 의문문 기호를 사용하지 않은 오류이다.

[표 4] 비구문적 입력 오류 예문

번호	중국어 예문	한국어 뜻
(가)	可能养一段时间(身体）就好了。	한동안 휴식하면서 몸 관리하면나아질거예요.
(나)	这一年我们过得可是真实在。	금년에 우리는 정말 충실하게 살았어요.
(다)	你要练的很好。	너는 연습을 잘해야 한다.
(라)	一切都好吧。	모두 잘 지내죠?

이와 같은 비구조적 오류들은 중국어 태깅 뿐만 아니라 구조적 분석 단계에서 구문 실패를 많이 유발하여 그 처리가 필요하다.

3.4 청유형과 경어 표현이 많음

중국어 이메일에는 청유형 뜻을 포함하고 있는 문장이 많다. 이런 청유형 표현은 문맥 상으로 주로 “希望(희망하다)”, “请(바라다)”과 같은 동사에 의해 결정된다. 예를 들면, “缺少文件，打不开，希望速发全，谢谢。(문서 없서 열 수 없으므로 빨리 보내주기 바랍니다. 감사합니

1) 중국어에서 '小'는 한국어의 '군','양'과 맞먹는 바, 성이나 이름 앞에 붙여 자신보다 어린 사람에 대한 친근감을 나타냄.

2) “养”: <가족>을 부양하다. <가축>을 기르다. 요양하다.

다.)" 와 같은 경우다.
중국어 이메일에는 한국어 이메일과 비슷하게 상대방의 직급, 연령, 관계에 따라 예의를 갖추어 경어형으로 표현하는 문장이 많다.
청유/경어 관련 특성은 중국어 이메일에만 국한된 특성은 아니며 많은 이메일 문서에 이 특성이 존재한다 하지만, 이 특성은 뉴스와 같은 문서와 차별되는 특성으로 기술적으로 해결해야 하는 특성이다. 특히 청유형과 같은 경우, 언어마다 다른 접근방법이 필요하다.

4. 중국어 이메일 특성 별 처리방법

위에서 소개한 중국어 이메일 특성을 중한 대화체 자동번역 시스템에 반영하기 위하여 본 절에서는 위의 각 특성 별 처리 방법을 기술하고자 한다.

4.1 굳어진 관용 표현에 대한 처리

본 논문에서는 인사말, 맺음말과 같은 관용적인 이메일 표현들에 대해서는 번역 메모리(TM)를 활용하였다. 이런 관용적인 표현들을 번역 메모리로 저장하고 그 매칭을 통하여 매칭된 경우에는 자동번역을 하지 않고 번역메모리에 있는 한국어를 그대로 생성하였다.
본 중한 자동번역 시스템에서는 번역메모리를 중국어 문장부과 그 대역 한국어 문장부로 구분하였다. 번역메모리의 활용을 높이기 위해서, 중국어 문장부에 들어 있는 문장기호(콤마, 콜론, 세미콜론, 마침표, 물음표, 느낌표)를 없애고 저장하여, 문장기호의 차이에 의한 매칭 저하를 방지하였다. 또한, 번역메모리를 매칭하기 전에 입력 중국어 문장을 문장기호를 기준으로 하여 단순 절 분절을 하여, 분리된 절들을 최장일치로 번역메모리의 중국어 문장부와 매칭을 시도하여 입력 중국어 문장의 시작절에서부터 마지막 절까지 수행한다. 입력 문장의 절 중에서 매칭된 모든 절에 대하여 번역메모리의 한국어 문장부의 한국어로 생성을 하고, 일치하지 않은 절에 대해서는 연속하는 절끼리 결합하여 자동번역을 실행하여 번역결과를 가져온다. 번역메모리에 매칭된 한국어 절과 자동번역된 한국어 절을 단순히 결합하여 최종 한국어 대역문을 생성한다.
예를 들어, 중국어 문장 "A, B, C, D."가 입력문으로 들어오면, 먼저 단순 절 분리를 통하여 A절, B절, C절과 D절로 분리한다. 번역메모리 매칭을 위하여, ABCD, ABC, AB, A 순으로 번역메모리와 매칭을 시도한다 만약 AB가 번역메모리에 있었으면, 그 대역어 $AB^{한국어}$를 AB에 대한 대역 문장으로 저장하고, 다시 매칭된 나머지 부분에 대해서 CD, C, D 순으로 매칭을 시도한다. 만약, D절이 번역메모리에 있어서 매칭된다면 D절에 대한 대역문장 $D^{한국어}$를 저장한다. 번역메모리에 매칭되지 않은 절인 C절에 대해서만 "C," 를 자동번역하여 자동번역 결과 $C^{자동번역한국어}$를 생성하여 저장한다. 그러면 마지막으로 저장된 각 절의 한국어들을 결합하여 "$AB^{한국어}$ $C^{자동번역한국어}$ $D^{한국어}$" 를 생성한다.
본 논문에서는 이메일에 자주 나타나는 문장에 대해 번역 메모리(Translation Memory)를 구축하기 위하여 아래와 같이 패러프레이징(Paraphrasing)을 통하여 TM의 적용범위를 넓혔다.

祝好运！->행운을 빕니다
祝你好运！->행운을 빕니다
祝您好运！->행운을 빕니다
祝大家好运！->행운을 빕니다

4.2 한글자 단어에 대한 처리

한 글자 단어에 대한 처리를 위해 본 논문에서는 고빈도 한 글자 호칭, 직함, 동사에 대하여 그 단어가 포함된 2 글자 단어를 사전에서 추출하여 그 글자 형태가 "성씨+호칭/직함", "동사+명사" 인지를 판단하여 특수 처리한다.
동사인 경우 이에 해당되는 2 글자 단어에 대하여 한 글자씩 분리 될 경우 "동사+목적어" 에 해당되는 경우에 대해서 대응되는 대역어를 부착함으로써, 고빈도 한글자 동사에 대한 대역어 변환 패턴을 수정하였다. 예를 들면, "放假" 라는 2 글자 단어는 "放" 과 "假" 한 글자 단어로 분리 가능하며 또한 "동사+목적어" 형태이므로, "放" 동사 변환패턴에 다음과 같이 "假" 와 같은 목적어가 오면 "방학+하다" 를 생성될 수 있도록 수정하였다.

<te> 放/VV #@# 놓아주/VV; 방목+하/VV;
<cl>=(obj 假/NN #@# 하/VV (obj 방학/NN)

4.3 입력 오류 처리

비구문적 입력 오류에 대하여 오래전부터 연구가 있었다. [4]에서는 비구문적 입력을 타이핑 오류, 구문 성분이 부족한 오류, 구문 성분이 중복된 오류 등 3 가지로 구분하고, N-gram 기법으로 오류를 수정하는 방법을 제안하였으며, [5]에서는 비구문적 오류를 각 도메인으로부터 수집하고, 이렇게 수집된 오류를 수동으로 교정하여 병렬 구문분석 코퍼스를 만들어 유사한 오류를 대응되는 병렬코퍼스에 있는 교정된 문장을 이용하여 수정하는 방법을 제안하였다.
본 논문에서는 중국어의 언어적 특성과 이메일과 같이 구하기 힘든 제한적인 도메인의 특성을 감안하여 위와 같은 방법을 시도하지 않았다. 다만, 의문형 단어("如何(어떻습니까), "吗(~나요)" 와 대답형 단어 "是的(그래), 好的(좋아)" 등 단어 뒤에 문장부호가 따르지 않거나, 틀리면 수정하는 방법을 사용하였다. 또한, 구조조사 '的, 地, 得 '의 오용에 대하여 앞뒤 어휘, 품사, 문장부호 등 문맥 정보를 참고하여 그 분포규칙을 찾아냄으로써 형태소 태깅 후처리 단계에서 자동 교정해 주는 방법을 사용하였다.

4.4 청유형, 경어 처리

요청, 의뢰 등의 내용의 전달 기능을 하는 이메일로서

는 이를 정확히 전달할 수 있어야 한다. 따라서 비명시적 고빈도 청유형 동사를 수집하여 문장 위치 자주 오는 부사 등을 정확히 파악하여 청유형 문장으로 생성해 주도록 하였으며, 경어처리를 위해 한국어 생성 단계에서 모든 문장은 존칭형으로 일괄되게 수정하였다. 이렇게 처리한 목적은 이메일에서 상대방의 직급 연령 등을 파악하기 어려우며 존칭형으로 수정하면 번역된 한국어도 보다 자연스럽기 때문이다.

5. 실험

5.1 실험 데이터

본 논문에서 제안한 중국어 이메일 특성을 이용한 중한 대화체 자동번역 방법의 효과성을 증명하기 위해자체적으로 수집한 250개 중국어 이메일 문서를 대상으로 실험을 하였다. 실험에 사용된 대상 문장은 총 250개(2700문장) 문서 중에서 200문장을 랜덤하게 추출하였으며, 실험에 사용된 평가문장의 평균 어절 수는 28.07 글자이다.

5.2 평가 방법 및 실험 결과

본 논문에서는 중국어와 한국어가 모두 능통한 평가자를 중심으로 5인 수동평가를 실시하였으며, 각 평가자는 다음 [표 5]의 평가 기준에 따라 평가하였다.

[표 5] 자동변역 수동 평가 기준

점수	평가 기준
4	원문의 의미가 그대로 전달될 경우
3.5	원문의 문장 전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1,2단어 대역어가 잘못된 경우
3	원문의 문장 전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 여러 단어의 대역어가 잘못된 경우
2.5	원문의 문장 전체의 분석은 실패했으나, 하나 이상의 동사구가 잘 분석되고 정확히 번역되어 부분적으로 문장의 의미가 전달될 경우
2.0	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 하나 이상의 명사구가 잘 분석되고 정확히 번역된 경우
1.0	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 문장 중에 하나 이상의 단어 또는 한 개의 명사구라도 정확히 번역된 경우
0	원문이 번역문에 그대로 출력된 경우

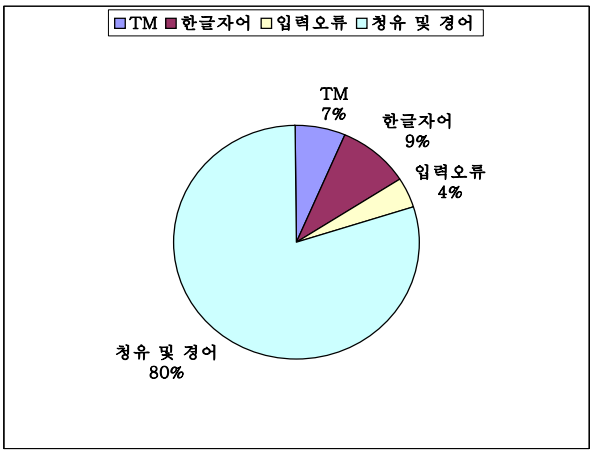
위의 평가 기준에서 알 수 있듯이, 3.0을 포함한 그 이상의 점수는 번역문을 읽었을 경우 전체적으로 그 의미가 이해가 되며, 점수 3.0 이하는 번역문을 읽었을 경우 전체적으로 그 의미가 이해되지 않음을 알 수 있다. 이와 같은 2가지 큰 분류의 기준(1과 0)은 체감번역률로써 본 논문에서 또 다른 평가기준의 하나로 사용하였다.

본 논문에서 제안한 중국어 이메일 특성을 이용한 중한 대화체 자동번역 방법의 유효성을 비교 평가하기 위해 기존 중한 뉴스 자동번역시스템(News CKMT)을 베이스라인으로 정하고 중한 대화체 자동번역 시스템(Dialog CKMT)과 비교 평가하였다(표 6 참조).

[표 6] 기존 시스템과의 번역률 비교

평가 시스템	번역률(%)	체감번역률(%)
news CKMT	72.02	39.00
Dialog CKMT	75.33(+3.31)	49.00(+10)

[표 6]에서 알 수 있듯이, 중국어 이메일 특성을 이용한 중한 대화체 자동번역이 기존 시스템에 비해 번역률 3.31%와 체감번역률 10% 향상되었다. 전체 200 평가문장 중 118개 문장은 성능이 향상되었으며, 77개 문장은 두 버전 간의 성능 변화가 없으며, 나머지 5개 문장은 오히려 성능이 기존보다 나빠졌다. 향상된 문장 중 각 특성별 차지하는 비중은 다음 [그림 2]와 같다.

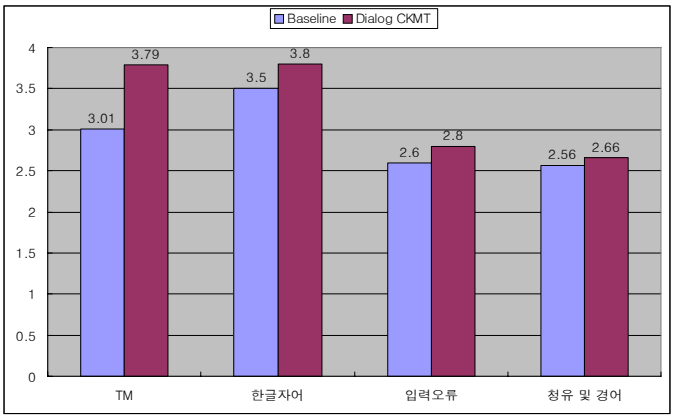


[그림 2 특성 별 성능 향상 비중

[그림 2]에서 알 수 있듯이, 청유 및 경어 처리가 전체 향상된 문장 중에서 가장 큰 비중인 80%를 차지하고 있다. 이는 청유형과 경어처리가 이메일 문장에서 매우 유용하고 효과적임을 알 수 있다. 그 외의 특성들이 차지하는 비중은 다소 차이가 있지만 모두 10%를 넘지 못했다. 그 원인은 이런 오류들이 실제 차지하는 비중이 작은 원인도 있으며, 또한, 본 논문에서 제안한 방법이 아직도 개선할 여지가 많음을 의미한다. 예를 들면, 입력오류와 같은 경우, 실제로 본 논문에서 제안한 문장부호 관련 오류 이 외에도 “你去工商局的时候要带上这个和你的身份证 °(당신이 공상행정국에 갈 때, 이 것과 당신의 신분증을 가져가야 한다.)” 의 예문 중 “时候(~할 때)” 와 같은 단어 뒤에 문장부호가 생략된 입력 오류를 쉽게 찾

을 수 있었다.

본 논문에서는 또 각 특성 별 성능 향상 정도를 비교하였다. 그 결과 [그림 3]과 같이 TM으로 인한 번역률 향상이 가장 눈에 띄었다. 비록 TM 매칭된 문장은 8문장 밖에 안되지만 베이스라인 버전에 비해 번역률이 25.9%나 향상되었다. 따라서, 앞으로 대량의 이메일 수집을 통해 TM의 사이즈를 늘리고 TM 적용범위를 확대할 필요성이 대두 되었다. 이와 상반되게 청유 및 경어처리가 가장 낮은 3.9%의 성능 향상을 보였다.



[그림 3] 각 특성 별 성능 향상 비교

본 논문에서는 또한 규칙기반의 자동번역 방법과 통계기반의 자동번역 방법을 비교하기 위해 구글Gmail 자동번역시스템과도 비교하였다(표 7).

[표 7] 통계적 방법과의 번역률 비교

평가 시스템	번역률(%)	체감번역률(%)
Gmail	54.56	9.50
Dialog CKMT	75.33(+20.77)	49.00(+39.5)

통계적 방법과의 번역률을 비교한 결과[표 3]과 같이 번역률과 체감번역률 각각 20.77%과 39.5% 향상된 것으로 차이가 많이 남을 알 수 있다. 구체적으로 분석해 보면 가장 큰 차이는 한국어 문장의 어순인데, 한일이나 일한 번역과는 달리 중국어와 한국어의 어순이 다르기에 규칙기반 자동번역 시스템은 변환패턴을 구축하여 중국어에서 자연스러운 한국어로 생성 가능하게 하는 반면 통계기반은 이러한 변환 규칙이 없기 때문에 동사의 생성 위치 오류가 많은 특징을 갖고 있었다. 이는 통계기반 자동번역 시스템의 번역률 특히 체감 번역률이 현저히 낮은 가장 중요한 요인으로 작용했다.

또한 통계기반 자동번역 시스템의 번역 문장에는 중국어글자가 그대로 나온다든지, 중국어 단어나 구절이 중국어 병음그대로 출력되어 번역문 이해에 걸림돌이 되었다. 그러나, 원문이 장문일 경우 통계기반 자동번역 시스템의 이러한 단점들이 뚜렷하게 나타나지만 단문인 경우, 일부 번역문장 생성이 비교적 자연스러워 체감 번역률도 높은 경우도 있다. 예를 들면, "感谢您的来信。"와 같은 문장에서 규칙기반의 자동번역 시스템에서는 "당신의 평지에 감사합니다."로 번역되었지만, Gmail의 경우 "편지 주셔서 감사합니다."로 번역되었다. 이 원인은 통계기반의 번역 방법은 코퍼스의 영향을 많이 받기 때문에 중한 병렬 코퍼스가 많이 부족하고, 더욱이 이메일과 같은 한정된 영역의 병렬 코퍼스는 더욱 부족하기 때문에 이와 같은 차이를 보이는 것으로 추정된다.

6. 결론 및 향후 연구

본 논문에서는 중국어 이메일 특성을 이용한 규칙기반의 중한 대화체 자동번역 방법에 대하여 기술하였다. 이메일과 같은 문체가 자유로운 대화체에서는 흔히 통계기반의 접근 방법을 많이 사용하고 있으며, 통계적 방법이 규칙기반 방법보다 우수할거라 생각한다. 하지만, 본 논문에서는 이와 반대로 중국어와 한국어와 같이 어순이 다르고 언어쌍 간 코퍼스 특히, 이메일과 같은 특정 도메인의 언어 자원이 불충분한 특성을 고려하여 통계적 방법보다 규칙기반의 접근방법을 시도하였다. 특히, 접근방법에 있어서 중국어의 이메일 특성을 잘 분석하고, 그 특성에 기반하여 대응되는 처리 방법을 사용하였다.

그 결과, 실험을 통하여 알 수 있듯이 본 논문에서 접근한 방법이 타당함이 증명되었다. 우선, 중국어 이메일 특성을 이용한 방법이 기존에 비해 주목할 만한 성능 향상을 보였으며, 다음으로, 통계적 방법에 비해서도 많은 성능 향상이 있었다.

향후 연구로 보다 많은 이메일 코퍼스를 수집하여 위의 특성에 따른 보다 효율적인 방법론 접근이 필요하며 이런 특성 이 외의 또 다른 특성에 대한 발굴도 필요하다.

7. 참고 문헌

[1] http://www.gmail.com

[2] L. Fais and K. Ogura, "Discourse Issues in the Translation of Japanese Email," In Proceedings of PACLING 200.

[3] X.M. Liu, "Usage Analysis of the Chinese Structural Auxiliary," Issue 12, Modern Chinese.

[4] K.H. Pang, Problems in the way of the Structural Auxiliary Words and Their Application, Journal of Shangqiu Teachers College, Vol. 20, 2004.

[5] Eric S. Atwell, "How to detect gramatical errors in a text without parsing it," Proceedings of the third conference on European chapter of the Association for Computational Linguistics, 1987.

[6] Foster, Jennifer and Carl Vogel, "Good reasons for noting bad grammar: Constructing a Corpus of ungrammatical language," In pre-proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives, 2004, Germany.