ETRI

2013
Electronics and
Telecommunications
Trends

빅데이터 처리 및
분석 기술 특집

빅데이터 기반 음성언어 처리 기술

Big data for Speech and Language Processing

나승훈 (S.H. Na) 언어처리연구팀 선임연구원
정호영 (H.Y. Jung) 음성처리연구팀 책임연구원
양성일 (S.I. Yang) 언어처리연구팀 선임연구원
김창현 (C.H. Kim) 언어처리연구팀 선임연구원
김영길 (Y.K. Kim) 언어처리연구팀 팀장

음성언어 처리 분야는 인간의 자연어 발화를 컴퓨터가 자동으로 이해하고 처리하는 알고리즘을 연구하는 분야로, 자동 통번역, Siri와 같은 음성 대화 시스템, 차세대 인터페이스, 질의 응답 시스템 등 다양한 응용군을 포함한다. 특히, 음성언어 처리 기술은, 최근 빅데이터(big data) 시대를 맞이하여, 방대한 음성/텍스트 정보를 처리하기 위한 필수 기술로 각광받고 있다. 한편, 빅데이터는 그 자체가 거대한 말뭉치 데이터로서 음성언어 처리 기술의 성능을 향상시키는 주된 리소스가 된다. 이에 따라, 최근 빅데이터를 이용하여 음성언어 처리 기술의 성능을 개선시키고자 하는 연구가 활발히 진행되고 있는데, 본고에서는 이들 연구의 배경 및 연구 동향들을 소개하기로 한다.

Ⅰ. 서론

지니톡(GenieTalk), Siri는 최근 스마트폰에서 큰 인기를 얻고 있는 서비스 및 프로그램의 이름이다. 지니톡은 한영/영한 자동번역 서비스, Siri는 음성으로 스마트폰의 기능을 수행하는 대화 시스템으로, 이들의 공통점은 인간의 음성/자연어를 인식하고 처리하는 서비스라는 점이다. 음성언어 처리 분야(speech and language processing)는 바로 이들 프로그램의 기초가 되는 연구 분야로, 포괄적으로는 인간과 컴퓨터 간의 자연어 인터페이스에 관한 주제를 다루고, 자연어를 컴퓨터가 자동으로 이해하고 처리하는 메커니즘을 핵심적으로 연구한다. 자동 통번역, Siri와 같은 음성 대화 시스템, 차세대 인터페이스, 질의 응답 시스템 등 수많은 응용군들이 음성언어 처리 기술에 기반을 둔다. 특히, 빅데이터(big data) 시대를 맞이해 접근 가능한 음성/텍스트 정보가 방대해짐에 따라, 음성언어 처리 기술은 빅데이터 분석/처리를 위한 필수 기술로도 그 중요성을 인정받고 있다.

대부분의 빅데이터 관련 연구는 빅데이터를 어떻게 분석하고 저장하고 관리하는지에 관한 것이다. 그러나, 음성언어 처리 분야에서 빅데이터는 그 자체가 거대한 말뭉치가 된다. 방대한 음성/텍스트 정보는 바로 인간이 발화한 거대한 자연어의 집합체로서, 이들은 음성언어 처리에서 다루는 여러 가지 제반 문제를 해결하는 데 도움이 된다. 이는 현대의 음성언어 처리 기술은 통계적 방법에 기반을 두기 때문이다. 통계에서는 데이터의 규모가 커질수록 확률 모델이 실제 모델과 더 가까워져 관련 확률 값이 보다 정확히 계산될 수 있는데, '빅데이터'를 활용하여 통계치를 획득할 때, 유도된 음성언어 처리 모델이 보다 정교해질 수 있을 것이다.

본고는 빅데이터를 하나의 거대한 리소스로서 음성언어 처리 연구에 활용하는 연구 동향을 살펴보기로 한다. 본고는 빅데이터를 관련 요소 기술을 위한 주된 리소스로 바라보는 관점을 취하므로, 빅데이터를 관리/저장하기 위한 기술에 대한 주제와는 다소 차이가 있다. 다음 장부터 빅데이터에 기반한 음성인식 분야, 언어처리 분야, 자동번역 분야에 대한 연구 분야를 각각 소개하도록 한다.

Ⅱ. 빅데이터 기반 음성인식

모바일 기술 및 클라우드 시스템의 성장으로 많은 IT 업체들이 자연스러운 인터페이스와 편리한 정보검색에 관심을 갖고 있다. 이에 음성인식 시스템을 이용한 자연스러운 인터페이스 및 대화형 정보검색 서비스 요구가 점차 확대되고 있는 상황이다. 대표적 검색업체인 구글도 모바일 검색 서비스를 제공하면서 음성검색 기능을 필수적으로 제공하고 있으며 사용자 데이터를 이용하여 사람이 수행하는 수준으로 발전시키려 하고 있다. 또한 컴퓨터 및 IT 기기의 대표적 업체인 애플도 클라우드 서비스를 내세우며 자연어 음성인식 기반 정보 제공 서비스인 Siri를 소개하여 많은 관심을 받고 있다. 이것은 무제한급 자연어 음성인식 기술을 기반으로 하는 음성 인터페이스 시대가 가까워오고 있음을 나타내는 대표적인 사례로 볼 수 있다.

하지만 고품질의 음성인식 서비스를 위해서는 많은 데이터와 다양한 지식을 활용하여 성능을 개선할 필요가 있다. 음성인식 기술의 보편적 활용에 제약을 주고 있는 해결 과제로는 사용자에 따른 인식률의 차이, 주변 잡음에 따른 인식률 저하, 인식대상 어휘의 제한으로 인한 인식 오류 발생을 들 수 있다. 이 문제들을 해결하기 위해 많은 데이터의 확보와 더불어 이를 활용하는 방법론 및 다양한 지식을 활용하는 음성인식 프레임의 필요성이 증가하고 있다. 음성인식을 위한 많은 데이터 및 다양한 지식은 음향학적 관점 및 언어학적 관점의 두 가지 방향에서 볼 수 있다. 음향학적 관점에서는 화자, 배경 잡음, 마이크로폰 등의 다양한 환경을 나타내는 데이

터 또는 지식을 활용할 수 있고 언어학적 관점에서는 어휘, 문법, 문맥 등을 모델링하기 위한 많은 데이터 및 이런 언어정보를 정확하게 추출하여 지식 정보로 활용할 수 있다. 음성인식 상황에서 두 관점의 지식을 메타 데이터로 표현할 수 있고, 지식 정보의 신뢰적 통계치를 얻기 위한 많은 데이터와 통계적 음성인식 프레임에 결합하는 방법론이 있다면 음성인식의 성능은 크게 개선될 수 있을 것이다.

최근 들어 스마트폰을 중심으로 한 모바일 인터넷 환경과 클라우드 서비스의 확대로 모바일 환경에서의 자연스러운 인터페이스의 수요가 확대되고 있으며, 이로 인해 많은 사람들이 음성인식을 사용함으로써 엄청난 규모의 사용자 로그 데이터를 확보하고 있는 상황이다. 이것은 다양한 배경환경에서 다양한 화자가 말한 다양한 어휘를 확보할 수 있는 것을 의미하며, 무제한급 자연어 음성인식을 위한 데이터 확보의 발판을 마련할 수 있는 계기가 되고 있다. 실제로 구글은 음성검색 서비스를 통해 하루 동안 한 사람이 2년 동안 쉬지 않고 얘기하는 양의 음성 데이터를 수집하고 있으며, 방대한 양의 음성 데이터 수집은 음향학적 정보뿐만 아니라 이 데이터의 전사로부터 다양한 분야의 텍스트 자료 수집을 가능하게 하며, 이를 통해 기존 음성인식업체에서는 시도조차 못해본 100억 개 이상의 문법 구조를 학습하여 음성인식 성능을 개선하는 데 활용하고 있다.

음성인식을 위한 빅데이터의 활용은 이처럼 다양한 환경에서 다양한 화자가 발성한 다양한 어휘, 문법 등을 분석함으로써 무제한급 자연어 음성인식을 위한 음향학적 지식 및 언어학적 지식을 체계화하는 데 중요한 의의를 가진다고 할 수 있다. 따라서 구글의 음성검색과 같은 서비스는 무제한급 음성인식 기술을 바탕으로 서비스를 제공할 뿐 아니라 이를 통해 얻은 빅데이터를 이용하여 음성인식 성능을 점진적으로 개선할 수 있는 선순환 구조에 있다고 볼 수 있다[1]. 구글 이전에 음성 기술 분야의 선두 업체는 뉘앙스였으나, 클라우드 컴퓨팅 기반으로 IT 생태계가 급변하는 동안 이에 대한 대응이 늦어져 모바일 음성인식 시장에서는 구글에 뒤쳐지게 되었으며, 이를 극복하기 위해 애플의 Siri 서비스 등에 음성인식 기술을 제공하기 시작했다. 이것이 의미하는 중요한 점은 음성인식 성능 개선을 위해 실제 서비스를 통한 음성 데이터의 수집이 필요하다는 것이며, 이런 빅데이터에 기반한 음성인식 모델링 기술이 필수적이라는 것이다.

구글은 현재의 음성인식 서비스를 확장하여 2019년경에는 거의 사람이 수행하는 정도의 수준으로 발전시키고자 한다. 이는 모바일 환경에서 엄청난 수의 안드로이드폰 사용자들을 통해 전 세계에서 감당하기 힘들 정도로 엄청난 양의 음성 데이터를 모을 수 있으며, 이를 통해 수집된 음성을 가공하여 성능 개선에 활용하는 선순환 구조를 형성하였기 때문에 가능할 수 있다. 또한 애플의 Siri 서비스에서 추구하는 것처럼 음성인식의 수준을 넘어서 음성 이해의 수준으로 진화해 사용자의 말을 이해하고 그 의도를 판단하여 그에 해당하는 서비스를 제공하는 수준에 도달하는 데도 빅데이터의 영향은 상당할 것이다.

앞에서 살펴본 것처럼 빅데이터를 이용해 음성인식을 위한 정보 및 지식을 체계화하는 것은 음향학적 또는 언어학적 관점에서 정의할 수 있으며, 각각에 대해 빅데이터 활용 및 이를 통한 성능 개선을 소개하고자 한다.

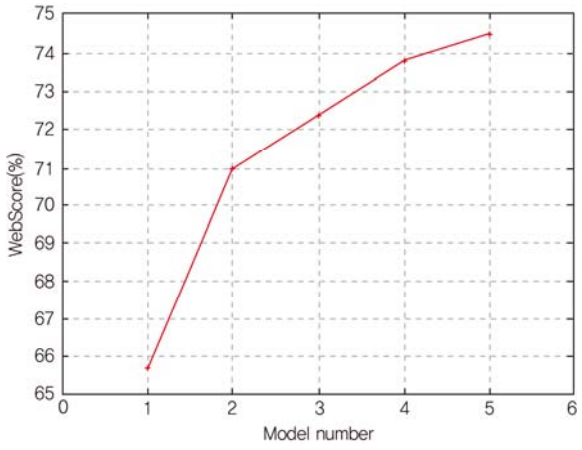
1. 빅데이터 기반 음향 모델

음향 모델링은 불특정 다수 화자의 다양한 발음 특성을 표현하는 것을 목적으로, 대용량의 음성 데이터로부터 통계적 방식으로 대표 패턴을 생성하는 것을 의미한다. 음성인식 시스템이 좋은 성능을 갖기 위해서는 다양한 환경, 화자, 어휘로부터 얻어진 대용량 훈련 데이터로 훈련된 음향 모델이 필요하며, 이것은 실제 음성인식 시스템이 사용되는 채널 환경, 사용자 주변 환경, 사용자 발화 패턴 등을 포함하는 음향학적 관점의 지식 정보

의 활용을 의미한다. 음성인식 서비스를 이용하는 사용자로부터 얻어진 음성 데이터는 실제 사용환경을 가장 잘 반영하는 훈련 데이터가 될 수 있다.

음성인식 시스템의 개발 초기에 있어서는 보편적인 음향 모델을 훈련하기 위해 일반적인 환경에서의 다양한 음성 데이터를 활용하게 된다. 이런 데이터는 주로 다양한 음소적 특성, 화자 특성 등을 반영하기 위한 기본 데이터로서 다양한 잡음이 반영되는 모바일 환경에서의 음성인식 서비스를 위해서는 미흡한 면이 있다. 이를 해결하기 위해서 서비스를 통해 수집되는 데이터를 대상으로 순차적으로 가공하여 음향 모델에 반영하는 것이 필요하다. 빅데이터를 활용하여 음향 모델을 개선하기 위해서는 훈련 가능한 데이터를 선별하는 기능, 미전사 데이터를 훈련에 적용하는 기술 등이 요구되며[2], 이를 통해 음성인식 시스템의 성능을 개선해 나갈 수 있다.

(그림 1)은 구글의 음성인식 서비스에서 음향 모델에 따른 성능 개선의 효과를 나타낸 것이다[1]. (그림 1)에서 모델 번호 1은 모바일 환경을 고려하지 않은 일반적인 데이터를 이용한 것이고, 모델 번호 2는 서비스 환경을 나타내는 1,000시간 데이터를 이용하여 구성한 것이다. 모델 번호 4와 5에서는 5,000시간의 데이터를 이용하여 비교사 학습방식을 도입하여 인식 성능을 개선하였다.



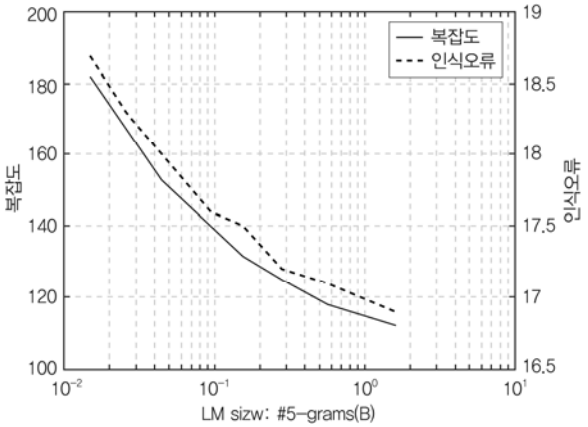
(그림 1) 음향 모델에 따른 구글 음성검색 서비스 성능

2. 빅데이터 기반 언어 모델

상용화된 음성검색 시스템의 개선을 위해서는 문법 구조를 나타내는 언어모델이 요구된다. 대표적인 언어 모델링 방법은 통계적인 방법에 따라 n개의 단어열에 대한 출현빈도를 확률값으로 나타내는 n-gram 기법이다. 정교한 n-gram을 생성하기 위해서는 다양한 코퍼스뿐만 아니라 실제 서비스에서 나타나는 언어 양상을 모델링할 필요가 있다. 이를 위해서는 빅데이터에 이용한 대규모 코퍼스 기반 언어 모델링 기술이 필수적이다.

음성인식 서비스의 경우 서비스 어휘의 수는 기하급수적으로 증가하며, 특정 도메인으로 대상 영역을 한정할 수 없는 특징을 가지며, 이것은 언어 모델의 대용량화와 지속적 확장 기능을 요구한다. 이를 해결할 수 있는 언어 모델링 기술의 최근 동향은 빅데이터 기반 대용량 분산 언어 모델링 기술 방식이 있다. 대용량 분산 언어 모델링 기술은 n-gram 차수의 무제한과 어휘 수의 무제한을 가져올 수 있으며, 미관측 어휘에 대한 언어모델링의 한계를 해결할 수 있는 방법으로서 빅데이터로부터 대규모 텍스트 코퍼스를 얻을 수 있음을 전제로 하고 있다. 최근 자료에 따르면 일반적인 영역에서 20만 어휘 수와 2GB 분량의 언어모델 크기를 갖는 반면, 모바일 웹 영역으로 확장했을 경우, 언어모델 크기는 1.8TB에 다다른다[3]. 이는 단일 서버로 처리할 수 있는 한계를 넘어서는 분량으로 분산 언어 모델링 기술을 요구하게 된다[4].

구글의 경우 자체적인 클러스터링 기술을 이용하여, 무제한의 어휘 수와 대규모 n-gram 개수를 기반으로 하는 언어모델을 제공하고 있다. 해당 클러스터링 기술은 MapReduce라는 프로그래밍 모델과 해당 라이브러리로 빅데이터의 분산 모델링 기술을 적용하여 무제한급의 언어지식을 구축하고 있다. 구글의 음성검색 서비스의 경우 연간 수억 개의 고유의 단어가 키워드로 입력되며, 누적된 검색어 2천억 개 이상을 언어지식 모델에 사용하고 있다. 언어모델은 선정된 어휘와 수집된 텍스



(그림 2) 언어 모델 크기에 따른 구글 음성인식 시스템의 복잡도 및 인식 오류

트 코퍼스를 이용하여 최대 100억 개의 문법 구조를 갖도록 개발되었으며, 이것은 임의의 텍스트 코퍼스를 골라내었을 때 여기에 포함된 문법 구조를 거의 표현할 수 있을 정도로 음성인식 시스템의 복잡도를 대폭 낮추는 효과를 갖게 된다. (그림 2)는 구글 음성인식 서비스에서 언어모델 크기의 증가에 따라 음성인식 시스템의 복잡도 및 인식 오류가 감소하고 있음을 나타낸다[1].

Ⅲ. 빅데이터 기반 언어처리

빅데이터를 활용한 언어처리 연구는 웹이 성장하여 구글과 같은 상용 검색 엔진이 크게 성장한 2000년대 초반 이후부터 집중적으로 연구가 진행되었다. 이들 연구에서는, 통계적 언어처리 기법 중 n-gram 빈도 수에 기반을 두는 오류 교정, 구문 분석에서 애매성 해소 등의 처리 등의 태스크에서, 검색 엔진을 통해 웹으로부터 n-gram 카운트를 어림하거나, 관련 질의어 집합을 정의하여 검색하는 방식으로 성능을 개선시켰다.

1. 웹 기반 모델

언어처리에서 대규모 데이터를 사용해야 하는 논의는 참고문헌 [5]에서 언급되었다. 참고문헌 [5]는 철자 오류 교정 태스크에서, 보다 많은 데이터를 사용할수록 성능이 개선됨을 보여주어, 대규모 데이터 사용 필요성을 강조했다.

대규모 데이터 사용은 웹이 각광받던 시기 이후에 특히 더욱 활발히 연구되었다. 먼저, 참고문헌 [6]은 웹 기반 n-gram 빈도 수가 원래의 코퍼스 기반 n-gram 빈도 수와 높은 상관관계를 가져 획득된 통계치의 안정성을 논증하고, 실제 가상의 의미 중의성 해소에 적용할 때, 웹 기반 빈도 수를 사용한 모델이 적은 규모의 코퍼스를 사용한 모델보다 우수함을 보여주었다.

참고문헌 [7]은 이전의 그들의 연구를 확장하여 웹 기반 모델을 8개의 언어처리 생성과 분석 문제에 적용하여, 성능 평가를 수행하였다. 그러나, 참고문헌 [7]의 연구에서는 웹 기반 모델이 베이스라인보다는 우수하나 기존의 최고의 지도 방식의 시스템의 성능 뛰어넘지 못함을 보여, 웹 기반 모델의 잠재성에 중도적인 입장을 피력했다. 결국, 그들은 웹 기반 모델은 새로운 베이스라인 정도로 인식되어야 한다고 결론을 내렸고, 이들이 최고 성능을 보이는 경우 또한 개별 언어처리 문제마다 다르다고 주장하였다. 그러나, 최근의 연구들은 웹 기반 방식이 효과적이라는 보고가 일반적이다. 예를 들어, 참고문헌 [8]에서는 Google N-gram[1]을 사용하여 어휘 선택 문제에 적용하여 실험한 결과, 코퍼스의 크기가 커질수록 더욱 높은 성능 향상을 가져옴을 보여주었다.

2. 웹 기반 구문 분석 모델

구문 분석 문제에서도 웹 기반 모델의 효과성을 보이는 연구들이 보고되었다[9]-[12]. 먼저 참고문헌 [10]은 영어에서 PP attachment 문제를 위해 웹 검색을 사용하는 방법을 제안하였는데, 그들은 PP attachment 문제 해소에 영향을 주는 여러 가지 질의어의 템플릿을 정의하고, 검색엔진 결과를 분석하여 웹 기반 모델을 근사

1) LDC에서 LDC2006T13로 이용 가능.

화시켰다. 실험 결과, 웹 기반 모델이 일반 모델보다 성능이 우수하다는 것을 확인했고, 웹 기반이 비지도 방식임에도 불구하고 지도 방식의 성능에 근사할 수 있음을 보여주었다.

또한 참고문헌 [11]은 구문 분석 결과가 오류가 있는지를 검사하기 위한 의미 필터로 웹 기반 방식을 제안하였다. 제안 방법은 구문 분석의 n-best 결과에 대해 의미 필터를 적용하여 재순서화하는 것으로, 의미 필터의 적용 결과, 베이스라인의 구문 분석 오류를 크게 감소시킬 수 있었다.

참고문헌 [12]는 기존의 선행 연구들을 확장하여, 구문 분석 전체에서 웹-기반 자질을 활용하는 방식을 제안하였다. 이들은 Google N-gram 코퍼스를 사용하여, 웹 기반 자질을 정의하여 이를 당시 최고의 성능을 보이는 의존 파서 및 구 구조 파서에 각각 적용하였다. 적용 결과, 최고 성능의 의존 파서 및 구 구조 파서에서도, 각각 7%와 9.2%의 구분 분석 오류 감소율을 보여주어, 웹 기반 자질의 유용성을 검증하였다.

요약하면, 빅데이터 기반 언어처리 연구에서는 웹을 코퍼스로 사용하여 n-gram 확률치를 보다 정확히 계산하거나, 언어처리에 도움이 되는 질의어 집합을 정의하고 이로부터 웹 기반 모델을 학습하는 방식으로 진행되고 있다. 최근의 연구에서는, 오류 교정, 어휘 선택, 구문 분석 등 다양한 언어처리 문제에서 빅데이터 기반 모델이 성능 향상에 도움이 된다는 것을 보여주고 있다.

Ⅳ. 빅데이터 기반 자동번역

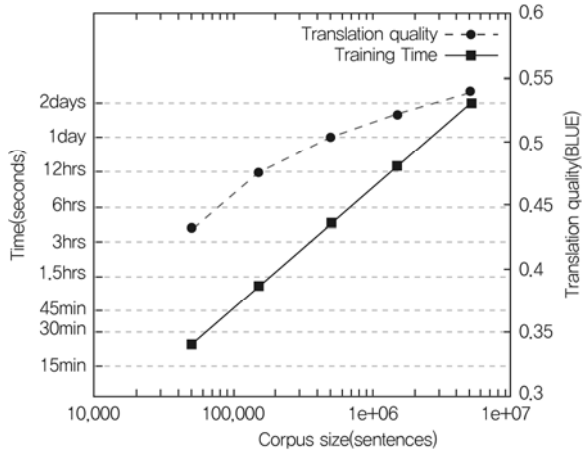
빅데이터를 활용한 자동번역 연구도 최근 들어 활발히 이루어지고 있다. 자동번역 연구에서 빅데이터는 번역 방식의 한 갈래인 통계 기반 자동번역 방법(statistical machine translation)에서 주로 많이 활용되고 있다. 통계 기반 자동번역 방법이란 번역 과정을 수학적으로 단순화시킨 접근법으로, 원문(source sentence)에서 번역문(target sentence)을 생성하는 확률 프로세스를 정의하고 이로부터 가장 높은 확률을 갖는 문장을 최종 번역문으로 제시하는 방식이다.

통계 기반 언어처리와 유사하게, 통계 기반 자동번역 방법에서는 확률 모델 자체가 번역 지식(translation knowledge)이 된다. 확률 모델은 다시 번역 모델(translation model)과 언어 모델(language model)의 두 가지 독립 모델로 나뉜다. 여기서 번역 모델은 번역문이 원문의 내용을 충실히 반영했는지를 계산하는 확률 모델로, 전통적인 대역어 사전이 이에 대응된다. 그리고 언어 모델은 특정 문장이 생성될 수 있는 개연성을 계산할 수 있는 확률 모델로, 자동번역에서는 여러 개의 후보 문장 중 어떤 것이 문법적 그리고 의미적인 관점에서 가장 자연스러운지를 비교하는 데 활용된다.

번역 모델과 언어 모델은 서로 상이한 모델이므로, 이들을 획득하기 위해서는 서로 다른 유형의 말뭉치가 활용되는데, 번역 모델은 원문과 대역문을 모아놓은 자료인 병렬 말뭉치(parallel corpus)로부터, 언어 모델은 목적어의 문장들로 구성된 단일어 말뭉치(monolingual corpus)로부터 구축된다.

1. 빅데이터 기반 번역 모델

빅데이터 병렬 말뭉치를 활용할 때 발생하는 문제는 대규모 데이터를 처리하는 데 드는 소요시간이 크다는 것이다. (그림 3)은 병렬 말뭉치의 크기를 증가시켜 감에 따라 걸리는 소요시간 및 번역 성능 곡선을 보여주고 있다[13]. 그림에서 보다시피, 병렬 말뭉치 규모가 커짐에 따라 번역 성능이 점차적으로 증가하는 한편, 학습에 필요한 소요시간도 함께 증가함을 볼 수 있다. 특히 소요 시간의 증가량이 말뭉치의 규모가 일정 수준 이상이 되면 완화되는 것이 아니라 지속성을 띄고 있다. 빅데이터 말뭉치가 확보되었다고 하더라도 이렇게 학습하는 데 소요되는 시간이 크다면 실제 활용하는 데 어려움이



(그림 3) 병렬 말뭉치 규모에 따른 번역 성능 및 학습 시간[12]

따를 수밖에 없을 것이다.

번역 모델 학습 속도를 획기적으로 개선하기 위한 대안은 Hadoop과 같은 클라우드 분산 아키텍처를 사용하는 것이었다. 대표적으로 구글은 클라우딩 컴퓨팅 환경을 주도했던 업체답게 초기 단계부터 분산 시스템을 자동번역에 적극 활용하였다. 메릴랜드 대학[13], 카네기 멜론 대학[14] 그리고 존스 홉킨스 대학[15] 등 미국 주류 대학팀들도 통계 기반 자동번역 방법에 분산 아키텍처 기술을 도입하기 시작했다.

주목할 만하게, 영국의 에든버러 대학 연구팀[16]은 병렬 말뭉치로부터 번역 모델을 학습하는 과정을 극단적으로 단순화시켰다. 여타 방식과 달리, 참고문헌 [16]에서는, 번역 모델을 미리 구축하지 않고, 주어진 병렬 말뭉치로부터 모든 가능한 대역 구(phrase) 쌍을 색인하고 저장만 해두는 것이 특징이다. 번역 모델의 학습은 입력문이 주어질 때 수행되며 그 과정은 색인된 데이터에 대한 검색을 수행하고 그 결과를 조합하는 단계로 이루어진다. 이러한 실시간 학습 방식에 기반하여, 참고문헌 [16]에서는 약 1TB급 규모의 번역 모델을 매우 효율적으로 구축할 수 있었다. 결과적으로, 당시 알려진 규모에 비해 30배 이상 큰 규모의 번역 모델을 구축할 수 있었다.

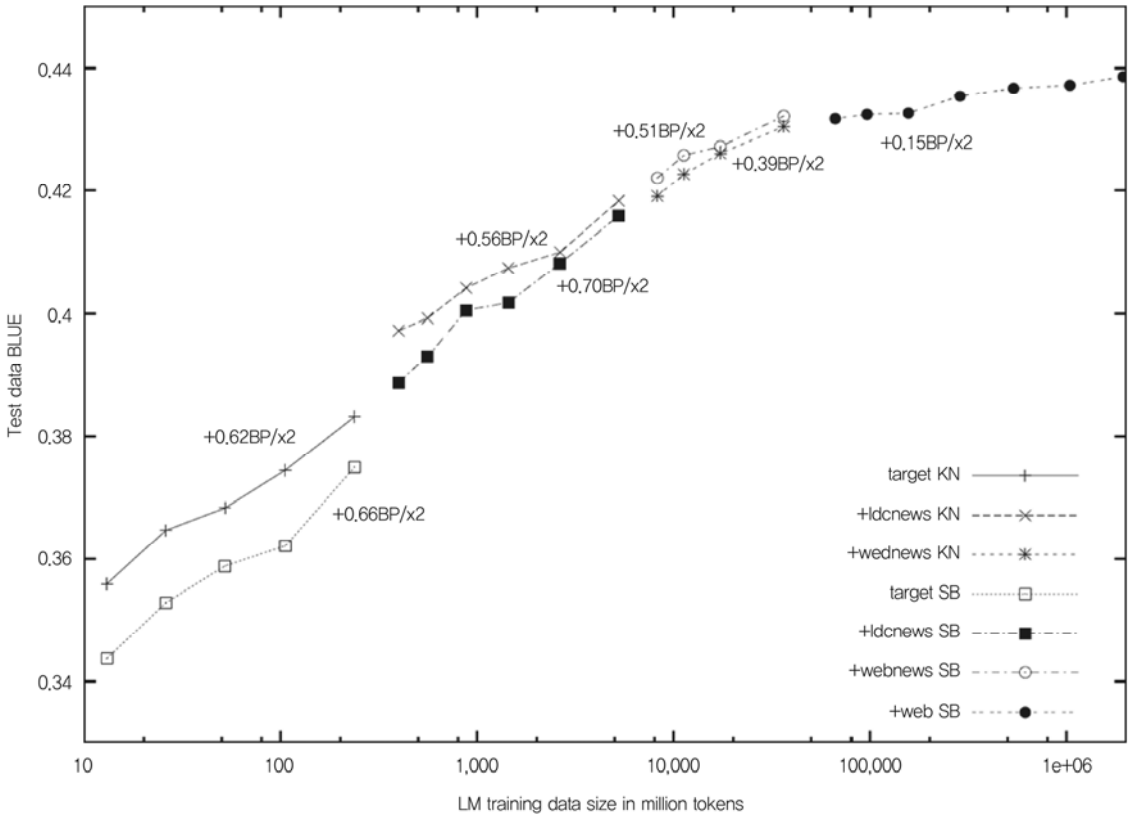
2. 빅데이터 기반 대규모 언어 모델

병렬 말뭉치 외에 단일어 말뭉치를 활용하여 자동번역의 성능을 높이는 연구 또한 활발히 진행되고 있다. 앞서 언급했듯이, 단일어 말뭉치는 언어 모델을 학습하는 데 쓰인다. 빅데이터가 등장함에 따라, 자동번역 연구팀들은 대규모 언어 모델(large language model)을 구축하기 시작했다. 여기서, 대규모 언어 모델이란 '빅데이터' 단일어 말뭉치 사용을 통하여 얻어진 규모화된 언어 모델을 일컫는다.

대규모 언어 모델의 연구는 구글 연구팀이 주도하고 있다. (그림 4)는 2007년도의 구글의 연구에서 얻은 실험 결과이다[3]. 참고문헌 [3]에서 구글은 웹으로부터 약 1.8TB 규모의 언어 모델을 구축하였는데, 여기서 주목할만한 점은 언어 모델의 규모화만을 통해, 번역 성능을 20% 이상 비약적으로 끌어올렸다는 것이다. 통상적으로, 번역 성능을 2~3% 이상 향상시키기가 어렵다는 점을 감안한다면, 당시 연구에서 향상된 성능 수치는 대단한 것이었다. 물론, 1.8TB 규모의 언어 모델은 당시 자동번역에서는 최대 규모 스케일이었지만, 구글의 전체 색인 정보의 크기에 비하면 작은 일부의 규모에 지나지 않는다. 실제 웹의 규모는 훨씬 크기 때문에, 웹의 부분 집합만을 사용하여 획기적인 번역 성능 향상을 가져온 구글의 연구 결과는 타 번역 연구팀들의 많은 관심을 불러일으키기에 충분했다. 구글의 연구 발표 이후, 보다 발전된 형태의 대규모 언어 모델링 방법들이 속속 개발되기 시작했다[17],[18].

3. 개발 사례

최근 들어, 자동번역 시스템 역시 빅데이터 말뭉치를 활용하여 구축되는 사례가 적지 않다. 대표적으로, 구글은 지속적인 투자를 통해 병렬 말뭉치 규모를 비약적으로 늘려왔으며 현재까지 57개 언어에 대해 약 200억 단어급 규모의 현재 세계 최대 규모의 병렬 말뭉치를 보유하고 있다.



(그림 4) 단일어 말뭉치 크기에 따른 번역 성능 개선[3]

빅데이터의 또 다른 활용 사례로 EuroMatrix를 들 수 있다[19]. EuroMatrix는 모든 유럽 언어에 대한 고성능 자동번역을 목표로 하는 다국적 프로젝트로서, 현재까지 각 언어별 평균 약 3천~4천 단어로 이루어진 총 4억 단어 규모의 방대한 병렬 말뭉치를 구축하고 있다. EuroMatrix는 기본적으로 통계 및 규칙 기반 자동번역 방법을 결합한 하이브리드 방식(hybrid method)을 채택하고 있으며, 아직까지 규모 면에서 구글에 미치지는 못하지만 병렬 말뭉치의 규모화를 통해 지속적으로 번역 성능을 높여가고 있는 중이다.

빅데이터는 학술적인 자동번역 연구에서도 적극 활용되고 있다. 미국 표준기술연구소(National Institute of Standards and Technology: NIST)에서 주최하는 번역 평가 대회에서는 약 천만 문장쌍으로 구성된 약 3억 단어 규모의 대규모 병렬 말뭉치를 공개하면서 분야를 선도하고 있다. 또한, 특허 번역의 수요가 높아짐에 따라, 특허 문제에 특화된 대규모의 병렬 말뭉치도 확보되고 있는 상황이다. 일본 정보통신연구소(National Institute of Information and Communications Technology: NICT)에서는 일본어-영어 약 2백만 특허 문장쌍의 병렬 코퍼스를 구축하였고[20], 홍콩시립대학(City University of Hong Kong)에서는 영어-중국어의 경우에 대하여 약 1천4백만 문장쌍의 병렬 코퍼스를 구축하여 사용하고 있다[21].

국내에서는 ETRI가 대규모 데이터에 기반하여 자동번역 시스템의 성능을 개선하는 연구를 추진한 바 있다. 다른 해외 연구팀과 달리, ETRI는 규칙 기반 자동번역 방법(rule-based machine translation)을 채택하여 차별화를 꾀했는데, 이 방식에서는 번역 지식의 규모와 품질이 성능에 영향을 미치는 핵심 요소이다. 이러한 번역

지식을 구축하는 데 드는 비용을 대폭 줄이기 위해서, ETRI는 웹데이터 및 대규모 코퍼스로부터 반자동으로 언어 분석에 필요한 지식을 추출 방법을 개발하여 이를 한중영 대화체 및 기술문서 자동번역 시스템에 탑재하여 사용하고 있다.

Ⅴ. 결론

본고에서 살펴본 것처럼, 빅데이터는 음성언어 처리 분야에서 거대한 리소스로 활용되어, 관련 요소 기술의 성능을 향상시키는 데 큰 도움을 주고 있다. 음성인식에서는 음향 모델링, 언어 모델링 학습을 위해 빅데이터를 사용하여 모델의 정확성을 향상시켰고, 언어처리에서는 어휘 선택 문제, 구문 분석 등의 문제에 대해서 웹에 기반하여 통계 모델을 학습하여 성능을 향상시켰다. 또한, 자동번역에서는 대규모 원시 코퍼스 및 대역 코퍼스로부터, 언어 모델 및 변환 모델을 규모화하여 번역 성능을 대폭 향상시켰다.

물론 이러한 연구에 더 나아가, 실제 상용 서비스 시스템에 빅데이터 기반 음성언어 처리 기술을 안착시키기 위해서는, 빅데이터를 저장하고 검색하는 제반 분산 아키텍처/클라우딩 컴퓨팅 기법을 도입하여, 이를 음성언어 처리의 각각의 확률 모델과 연동시키는 정교한 작업이 진행되어야 할 것이다.

용어해설

통계 기반 자동번역(Statistical machine translation) 번역을 원시어에서 목적어로 변환되는 확률 프로세스로 정의하고, 번역 확률 모델을 병렬 말뭉치로부터 유도하는 자동번역 방법

음향 모델링 음성인식을 위해 음성 데이터로부터 음소 등의 인식 단위를 나타내는 대표 패턴을 생성하는 과정

언어 모델링 음성인식/언어처리 위해 텍스트 코퍼스로부터 단어들의 순서에 기반한 문법 구조를 학습하는 과정

약어 정리

NIST National Institute of Standards and Technology

NICT National Institute of Information and Communications Technology

참고문헌

[1] J. Schalkwyk et al., "Google Search by Voice: a Case Study," *Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics*, A. Neustein, Ed., Springer, 2010.

[2] J. Ma and S. Matsoukas, "Unsupervised Training on a Large Amount of Arabic Broadcast News Data," *ICASSP*, vol. II, 2007, pp. 349-352.

[3] T. Brants et al., "Large Language Models in Machine Translation," *EMNLP*, 2007, pp. 858-867.

[4] A. Emami, K. Papineni, and J. Sorensen, "Large-Scale Distributed Language Modeling," *ICASSP*, vol. IV, 2007, pp. 37-40.

[5] M. Banko and E. Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation," *ACL*, 2001, pp. 26-33.

[6] F. Keller and M. Lapata, "Using the Web to Obtain Frequencies for Unseen Bigrams. Computational Linguistics," vol. 29, no. 3, 2003, pp. 459-484.

[7] M. Lapata and F. Keller, "Web-based Models for Natural Language Processing," *ACM Trans. Speech and Language Process.*, vol. 2, no. 1, 2005, pp. 1-31.

[8] S. Bergsma, D. Lin, and R. Goebel, "Web-Scale N-gram Models for Lexical Disambiguation," *IJCAI*, 2008, pp. 1507-1512.

[9] E. Pitler et al., "Using Web-scale N-grams to Improve Base NP Parsing Performance," *COLING*, 2010, pp. 886-894.

[10] P. Nakov and M. Hearst, "Using the Web as an Implicit Training Set : Application to Structural Ambiguity Resolution University of California at Berkeley," *HLT*, 2005, pp. 835-842.

[11] A. Yates, S. Schoenmackers, and O. Etzioni, "Detecting Parser Errors Using Web-based Semantic Filters," *EMNLP*, 2006, pp. 27-34.

[12] M. Bansal and D. Klein, "Web-Scale Features for Full-Scale Parsing," *ACL-HLT*, 2011, pp. 693-702.

[13] C. Dyer et al., "Fast, Easy, and Cheap : Construction of Statistical Machine Translation Models with MapReduce," *3rd StatMT*, 2008, pp. 199-207.

[14] Q. Gao and S. Vogel, "Training Phrase-Based

Machine Translation Models on the Cloud," The *Prague Bull. Math. Linguistics*, no. 93, 2010, pp. 37–46.

[15] Y. Cao and S. Khudanpur, "Sample Selection for Large-scale MT Discriminative Training," *AMTA*, 2012.

[16] A. Lopez, "Tera-Scale Translation Models via Pattern Matching," *COLING*, vol. 1, 2008, pp. 505–512.

[17] A. Pauls and D. Klein, "Large-Scale Syntactic Language Modeling with Treelets," *ACL*, vol. 1, 2012, pp. 959–968.

[18] Z. Li and S, Khudanpur, "Large-scale Discriminative n-gram Language Models for Statistical Machine Translations," *AMTA*, 2008.

[19] EuroMatrix. http://www.euromatrix.net/

[20] M. Utiyama and H. Isahara, "A Japanese-English Patent Parallel Corpus," *MT summit XI*, 2007, pp. 475–482.

[21] B. Lu et al., "Mining Large-scale Parallel Corpora from Multilingual Patents: an English-Chinese Example and Its Application to SMT," *CIPS-SIGHAN Joint Conf. Chinese Language Proc.*, 2010.