기계번역 성능평가를 위한 핵심어 전달율 측정방안

한국전자통신연구원 음성/언어기술연구센터
유초롱 · 이영직 · 박 준

Evaluation Method of Machine Translation System

Chorong Yu, Youngjik Lee, Jun Park
Speech/Language Technology Research Center, Electronics and Telecommunications Research Institute, Daejeon, Korea

요 약

본 논문은 기계번역 시스템의 성능평가를 위한 '핵심어 전달율 측정' 방안에 대해서 기술한다. 기계번역 시스템의 성능평가는 두 가지 측면으로 고려될 수 있다. 첫 번째는 객관적인 평가로 IBM에서 주창한 BLEU score 측정이나 NIST의 NIST score 측정이 그 예이다. 객관적인 평가는 평가자의 주관적인 판단이나 언어적인 특성을 배제한 방법으로 프로그램을 통해 자동으로 fluency와 adequacy를 측정하여 성능을 평가한다. 다음은 주관적인 평가이다. 주관적인 평가는 평가자의 평가를 통해 번역의 품질을 평가하는 방법이다. 주관적 평가 방법의 대표적인 것으로는 NESPOLE이나 LDC가 있다. 주관적인 평가는 평가자의 정확한 판단으로 신뢰할만한 성능평가 결과를 도출하지만, 시간과 비용이 많이 들고, 재사용할 수 없다는 단점이 있다. 본 논문에서는 이러한 문제를 해결하기 위해, 번역대상 문장에서 핵심어를 추출하고, 그 핵심어가 기계번역 시스템의 수행결과에 전달된 정도를 자동으로 측정하는 새로운 평가방법인 '핵심어 전달율 측정' 방안을 제안한다. 이는 성능평가의 비용과 시간을 절약하고, 주관적 평가와 유사한 신뢰성 있는 평가결과를 얻을 수 있는 좋은 지표가 될 수 있을 것으로 기대한다.

개 요

기계번역 시스템 개발과정에서 성능평가는 큰 비중을 차지하고 있다. 성능평가는 개발과정에서 발생한 변경사항을 적용했을 때의 성능을 비교, 분석하여 개발과정에 반영하기 위해서 사용된다. 기계번역 시스템의 성능평가는 두 가지 측면으로 고려될 수 있다. 첫 번째는 객관적인 평가로 IBM에서 주창한 BLEU score 측정이나 NIST의 NIST score 측정이 그 예이다. 객관적인 평가는 평가자의 주관적인 판단이나 언어적인 특성을 배제한 방법으로 프로그램으로 통해 자동으로 fluency와 adequacy를 측정하여 성능을 평가한다. 다음은 주관적인 평가이다. 주관적인 평가는 평가자의 평가를 통해 번역의 품질을 평가하는 방법이다. 주관적 평가 방법의 대표적인 것으로는 NESPOLE이나 LDC가 있다. 주관적인 평가는 평가자의 정확한 판단으로 신뢰할만한 성능평가 결과를 도출하지만, 시간과 비용이 많이 들고, 재사용할 수 없다는 단점이 있다. 본 논문에서는 이러한 문제를 해결하기 위해, 번역대상 문장에서 핵심어를 추출하고, 그 핵심어가 기계번역 시스템의 수행결과에 전달된 정도를 자동으로 측정하는 새로운 평가방법인 '핵심어 전달율 측정' 방안을 제안한다.

본 논문은 다음과 같은 구성으로 이루어져 있다. 2장에서는 기존에 소개된 기계번역 시스템의 성능평가 방법에 대해서 기술하고, 3장은 '핵심어 전달율 측정'에 대해서 설명한다. 4장에서는 '핵심어 전달율 측정'의 결과를 보이고 그 결과의 분석 내용을 기술하며, 5장에서 결언을 맺는다.

E-mail : crryu@etri.re.kr
E-mail : ylee@etri.re.kr
E-mail : junpark@etri.re.kr

관련연구

본 장에서는 기존에 연구되었던 기계번역 시스템의 성능평가 방법에 대해서 소개한다. 기본적으로 기계번역 시스템의 성능평가를 위한 요구사항은 다음과 같다.

1) Objective(unbiased)
2) Replicable(consistency provides confidence)
3) Indicative(of what it will be useful for)
4) Informative(for developers)
5) Actionable(for TIDES direction decisions)[1]

이러한 요구사항에 기초하여 기계번역 시스템의 성능평가의 연구 노력은 두 가지 방향으로 전개되어 왔다. 하나는 Comprehension Measure를 통한 평가 방법으로 객관적인 방법이 이에 해당한다. 다음은 Diagnostic Measure를 통한 평가 방법이다. 이는 task-oriented된 방법으로 평가자의 판단이 반영된 주관적 평가 방법이다.

1. 객관적인 평가

기계번역 시스템의 객관적인 평가를 위해서는 번역대상 언어와 함께, 정답 셋인 레퍼런스가 필요하다. 정답 셋과 번역결과를 비교하여 객관적인 수치를 산출하기 때문이다. 대표적인 객관적인 평가방법으로는 IBM에서 제안한 BLEU scoring 방법과 NIST의 NIST scoring 방법이 있다.

1) BLEU score

IBM에서 제안한 방법으로 unigram부터 four-gram까지 코퍼스 단위로 scoring한다.

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} w_n \log p_n)$$

$$BP = \begin{cases} 1 & \text{If } c>r \\ e^{(1-r/c)} & \text{If } c<=r \end{cases}$$

BP는 Brevity penalty로서 기계번역 결과 문장의 길이가 정답 셋인 레퍼런스 문장 길이보다 짧은 경우에 penalty를 주기 위한 factor이다. 기계번역 결과 문장 길이가 레퍼런스 문장보다 긴 경우에는 n-gram modified precision measure에서 고려되었기 때문에 무시한다.

2) NIST score

NIST score는 BLEU score에 기반을 두고 n-gram 계산과 Brevity penalty의 계산을 변경하여 scoring하는 방법이다.

3) 객관적 방법의 문제점

기계번역 시스템의 객관적 평가는 각 언어적 특성이나, 기계번역 시스템의 특성, 사용영역 등에 독립적인 평가방법으로써, 프로그램을 통해 자동으로 성능평가가 가능하다. 이는 비용과 시간의 절약을 가져올 수 있지만, 주관적 방법에 비해 신뢰성이 떨어진다.

2. 주관적인 방법

기계번역 시스템의 주관적인 평가는 평가자의 판단에 기반을 둔다. 대표적인 방법으로 NESPOLE이나 TIDES 프로젝트의 일환으로 제안된 LDC 방법, MITRE's NEE scoring 방법 등이 있다.

1) NESPOLE

주관적 평가의 대표적인 방법으로 초기에는 P(Perfect), K(OK), B(Bad)의 3단계로 평가하였다.

P : correct translation and fluent output
K : reasonable translation but disfluent output
B : improper or no translation

현재는 Very good, Good, Bad, Very bad의 4단계로 평가한다. 평가자의 판단에 의한 평가방법이므로 bias되지 않기 위한 여러 제약이 필요하다.

2) MITRE's NEE

키워드기반의 diagnostic 평가방법이다. 레퍼런스 코퍼스에서 추출된 레퍼런스 entity list를 이용해서 score를 측정하는 방법이다. 레퍼런스 문장들로 이루어진 코퍼스에서 키워드를 미리 추출하고 기계번역 결과에서의 키워드들 출현 여부에 따라 scoring하는 방법이다.

3) 주관적 방법의 문제점

객관적 방법에 비해서 주관적 방법은 신뢰할만한 성능평가 결과를 산출해 낸다. 그러나 평가자들의 판단에 기초하기 때문에, 고가의 비용과 시간이 많이 걸리는 단점이 있다. 게다가 각 평가는 일회성으로 그치고 재사용이 불가능하다.

1 TIDES는 Trans-lingual Information Detection, Extraction and Summarization의 약자로 English speaker가 다른 언어에 대한 기본 지식 없이도 중요한 정보를 찾아내고 해석할 수 있도록 하기 위한 목적으로 시도된 프로젝트로서 기계번역 시스템의 주관적인 평가 방법을 제안하기도 했음.

핵심어 전달율 측정방안

1. 개 념

'핵심어 전달율 측정' 방안은 자동으로 성능평가가 가능하면서도, 주관적 평가결과에 가까운 평가결과를 얻기 위해서 제안된 평가방법이다. 우선 번역대상 문장에 포함되어 있는 핵심어들을 선정해내고, 그 어휘가 기계번역 결과에 전달된 정도를 측정하여 성능을 평가한다. 핵심어는 번역대상 문장이 말하고자 하는 의도를 가장 잘 나타낼 수 있는 중요한 키워드로 볼 수 있다. 이러한 핵심어가 기계번역 결과에 잘 전달 되었다면, 문장의 의도가 잘 전달된 것으로 생각할 수 있을 것이다.

1) 특 징

1) fluency와 문장 내에서 단어 순서 무시.
2) 번역대상 문장의 형태소 결과 필요함.
3) 기계번역 결과의 레퍼런스(정답문장) 필요 없음.
4) 대역사전 필요함.

2) 제약조건

1) 핵심어 전달율 측정은 문장단위로 이루어짐.
2) scoring range는 0에서 1사이.
3) 코퍼스 전체의 score는 각 문장의 score 합/문장 수(average)로 계산.

2. 핵심어 전달율 측정방안 동작흐름

Fig. 1은 '핵심어 전달율 측정' 방안의 동작흐름을 나타내고 있다. 1)부터는 각 동작 단계별로 자세하게 기술하고자 한다.

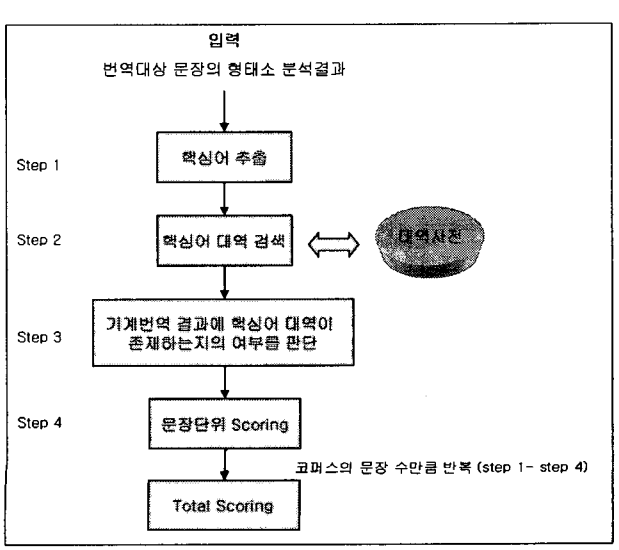

Fig. 1. 동작 흐름.

3. 핵심어 전달율 측정을 통한 평가

1) 번역대상 코퍼스의 형태소 분석

'핵심어 전달율 측정'에서 핵심어를 선정하기 위해서는 번역대상 코퍼스의 형태소 분석 정보를 이용한다. 번역대상 코퍼스의 형태소 분석 결과를 통해, 핵심어로 남길 만한 형태소들을 선정한다.

Fig. 2는 한국어 형태소 분석 결과의 예를 보이고 있다. '핵심어 전달율 측정' 방안은 한-영 기계번역 시스템의 개발과정에서 도출된 성능평가방법으로 본 논문의 실험 환경과 같은 한-영 기계번역 시스템의 경우를 예로 들어 설명하고자 한다.

2) 핵심어 추출

Fig. 2와 같이 형태소 분석 결과를 보고, 핵심어가 될만한 형태소와 그 컨텐츠를 추출한다. 핵심어가 될 형태소는 언어학적인 경험적 지식에 기반을 두고 직관적으로 선별하였다. 핵심어 선별은 국어국문학 전공의 전문가와 기계번역 전문가들로부터 핵심어로 선정될 어휘들의 품사들의 정보를 얻고, 실제 추출 작업은 프로그램을 통해 자동으로 수행하였다.

이를 통해 선정된 핵심어 들의 리스트는 다음과 같다.

◇ 관형사 류 : mma(성상형용사), mmc(수관형사), mmd(지시관형사)
◇ 명사 류 : nbu(단위성 위존명사), nbn(비단위성 의존명사), ncn(비서술성명사), ncpa(서술성, 동작성 명사), ncps(서술성, 상태성 명사), nnc(양수사), nnn(숫자), nno(서수사), npd(지시대명사), npp(인칭대명사), nq(고유명사)
◇ 동사/형용사 류 : paa(성상형용사), pad(지시형용사), pvd(지시동사), pvg(일반동사)
◇ 부사 류 : mag(일반부사), mad(지시부사)

핵심어 리스트에 해당하는 품사만을 남기고 나머지 컨텐츠와 형태소들은 삭제한다. Fig. 3은 Fig. 2에서의 형태소 분석 결과를 기반으로 핵심어 추출결과를 보여주고 있다.

```
기다려/pvg 주/px+서/ep+서/ecs 감사/ncpa+하/xsv+ㅂ니다/ef
방/ncn 열쇠/ncn 여기/pvg+ㅆ/ep+습니다/ef
삼/nnc 층/nbu+이/jp+고/ecc 방/ncn+은/jxt 314/nnn+호실/nbu+이/jp+ㅂ니다/ef
그/mms 사람/ncn+이/jcs 옷/ncn+을/jco 사/pvg+서/ecs 저/npp+한테/jca 선물/ncn+로/jca 줘/pvg+ㅆ/ep+어요/ef
제일/mag 유명/ncps+하/xsm+ㄴ/etm 백화점/ncn+이/jcs 어디/npd+에요/ef
```

Fig. 2. 형태소 분석 결과의 예.

```
기다려/pvg 감사/ncpa
방/ncn 열쇠/ncn 여기/pvg
삼/nnc 층/nbu 방/ncn 314/nnn 호실/nbu
그/mms 사람/ncn 옷/ncn 사/pvg 저/npp 선물/ncn 줘/pvg
제일/mag 유명/ncps 백화점/ncn 어디/npd
```

Fig. 3. 핵심어 추출결과.

3) 대역사전에서 핵심어 대역 검색

2)에서 추출해낸 핵심어들의 적절한 대역을 대역사전에서 검색하여 리스트로 만든다. 한-영 기계번역 결과의 경우, 한국어 핵심어들에 대한 대역 검색은 단어와 품사정보를 이용하여 이루어진다.

4) 기계번역 결과와 비교하여 scoring

3)에서 검색한 대역들이 기계번역 결과에 존재하는지를 비교하여 scoring한다. Scoring은 문장 단위로 이루어지며, 추출한 핵심어들의 대역이 기계번역 결과에 포함되어 있는지의 여부에 따라서 점수를 매긴다. 예를 들어, 번역대상인 한글 문장에서 추출한 핵심어가 세 개인 경우에 기계번역 결과에 핵심어들의 대역아이템이 두 개 존재한다면, score는 $\frac{2}{3}$가 된다.

5) 코퍼스 단위의 scoring

'핵심어 전달율 측정' 방안은 문장 단위로 이루어진다. 전체 코퍼스에 대한 scoring은 위에서 기술한 2)에서 4)의 단계를 코퍼스 내의 문장 수만큼 반복하여 score의 합을 구한 후, 코퍼스 내의 총 문장 수로 나누어 평균값을 구하는 것이다.

핵심어 전달율 측정 실험

ETRI에서는 여행자 영역의 대화체 문장을 대상으로 기계번역 시스템인 PDMT를 개발하고 있다.

PDMT는 Phrase-based Direct Machine Translation의 약자로서, chunk-based SMT와 direct MT방법을 접목시켜 실생활에 응용될 수 있는 시스템 개발을 위해 연구중에 있다. '핵심어 전달율 측정' 방안은 여행자 영역이라는 특수한 상황의 기계번역 시스템의 fluency나 adequacy와 같은 기존의 평가방안의 측정지표들과 차별되는 좀더 간단하면서도 명료한 '의도전달'에 초점을 맞추었다. 좋은 품질의 대역문장을 생성하는 것보다는 여행자가 사용했을 때, 전달하고자 하는 의미가 전달되는 것이 그 목적이라는 데 착안하여 시스템을 개발하였기 때문이다. 이러한 의도전달을 측정기준으로, 번역대상 문장 내의 핵심어들을 추출하고 그 핵심어들의 대역이 제대로 전달되었는지를 측정하는 '핵심어 전달율 측정' 방안이 제안되게 되었다.

'핵심어 전달율 측정' 실험을 위한 환경은 다음과 같다.

◇ 코퍼스 : 여행자 영역의 대화체 한국어 551문장, 그에 해당하는 PDMT 시스템의 번역결과.

◇ 형태소 분석기 : 여행자 영역의 대화체 문장으로 이루어진 대용량 코퍼스를 이용하여 튜닝된 '대화체 번역용 태거'

◇ 한-영 대역사전 : 여행자 영역의 어휘들을 추가하여 튜닝한 한-영 대역사전

이러한 리소스들을 이용하여 테스트 코퍼스에 해당하는 핵심어 전달율을 측정하였다.

```
코퍼스 내의 총 핵심어 수 : 1863
기계번역 결과에 포함된 핵심어 대역의 수 : 815
핵심어 전달율=0.43736
```

핵심어 전달율은 번역대상 코퍼스의 핵심어가 약 43% 정도 기계번역 결과에 전달되었음을 나타낸다. 이 결과를 통한 기계번역 성능평가로서의 가치는 아직 미지수이다. 좀더 많은 기계번역 시스템 결과와의 비교, 주관적 평가결과와의 비교를 통해서 성능 평가 결과의 신뢰성을 획득해야 한다.

1. 고려사항

본 절에서는 '핵심어 전달율 측정' 실험을 통해 발견된 고려사항에 대해서 기술하고, 해결방법에 대해서 생각해 보고자 한다.

1) 유의어 문제

한영 대역사전에서 핵심어에 대한 대역 아이템을 찾아낼 때, 사전에 포함되어 있는 아이템 이외의 유사한 표현들에 대한 고려는 포함되어 있지 않다. 예를 들어, '말해/pvg'라는 핵심어가 존재한다고 가정해보자. MT 결과에는 'tell'이라는 단어가 포함되어 있지만, 대역사전에 '말해'에 해당하는 대역으로는 'say'만이 존재한다. 이런 경우, 같은 뜻을 갖는 유사어휘가 포함된 경우에는 scoring되지 않고 누락되는 경우가 생긴다. 이를 해결하기 위해서는 영어의 유의어 리스트를 작성하고 이를 이용하여 해결할 수 있을 것이다.

본 논문에서는 프린스턴 대학에서 개발한 WordNet 1.7.1을 이용하여 영어 유의어 리스트를 구축하였다. 그리고 이 유의어 리스트를 이용하여 핵심어 전달율을 새로이 측정하였다.

코퍼스 내의 총 핵심어 수 : 1863
기계번역 결과에 포함된 핵심어 대역의 수 : 815
핵심어 전달율=0.43736

유의어 리스트를 포함한 경우에 약 3.2%정도 전달율이 향상되는 것을 볼 수 있다.

2) Weight 가중

현재 scoring하는 방법에서는 문장 특성에 따른 weight의 고려가 없다. 핵심어가 2개인 문장과 10개인 문장이 있다고 가정해보자. 핵심어가 2개일 때, 기계번역 결과에 하나의 핵심어 대역 아이템만 포함되어 있다면 score는 $\frac{1}{2}$이 된다. 핵심어가 10개일 경우에, MT 결과에 5개의 핵심어 대역아이템이 포함되어 있다면 그 score도 $\frac{5}{10}(\frac{1}{2})$가 된다. 이 경우에 두 문장을 똑같이 취급하는 것은 문제가 있다. 이 문제를 해결하기 위해서는 문장당 핵심어 수에 따른 제약조건을 두어 상이한 weight를 주는 방법을 강구해야 한다.

3) 핵심어 추출이 불가능한 경우

간단한 대답이나 감탄사 등으로 이루어진 문장에서는 핵심어 추출이 불가능하다. 이런 경우를 고려하여 핵심어 추출이 이루어져야 한다.

4) 대역사전에 존재하지 않는 핵심어들의 처리

추출된 핵심어 중에서 대역사전에 어휘가 존재하지 않는 아이템들에 대한 처리도 이루어져야 한다. 이는 사전 보강 작업을 통해서 어느 정도 극복할 수 있을 것으로 기대한다.

5) 유사구문에서의 핵심어들의 의미모호성

본 논문의 '핵심어 전달율 측정'에 사용된 핵심어들은 한국어 문장의 형태소 분석 결과를 토대로 추출되었다. 이 핵심어 추출과정에서 영어의 뜻과는 상관없는 한국어적인 표현이나 단어들의 사용되는 유사구문의 경우에 '핵심어 전달율 측정'에 noise로 작용하게 된다. 예를 들어, '예, 손님'과 같은 간단한 문장을 생각해보자. 이 문장에서 핵심어는 '손님/ncn'이 된다. 핵심어 손님에 대한 영어 대역은 'visitor'이다. 그렇다면 기계번역 결과에 visitor가 나와야만 핵심어가 제대로 전달됐다는 결론에 도달할 수 있다. 그러나 이 문장은 대개의 경우에 있어서 'Yes, sir/madam'과 같이 영어로 번역되는 게 자연스럽고 합당하다. 이런 경우에 대해서는, 관용어구 사전 등을 구축하여 해결하는 방법을 고려하고 있다.

6) 동사의 시제변화와 명사의 복수 처리

핵심어로 선택된 동사와 명사들은 한-영 대역사전에 포함된 기본형만을 고려하여 기계번역 결과에 존재하는 지 여부를 판단하여 scoring한다. 동사의 시제 변화와 3인칭 변화, 명사의 복수형 처리 등을 배제한 아주 기본적인 scoring 방법이다. 동사의 변화형과 명사 복수형들의 리스트를 수집하여 대역사전에 추가하는 방법을 강구할 수 있다.

결 론

지금까지 '핵심어 전달율 측정' 방안에 대해서 소개하고 테스트 코퍼스에서의 측정결과에 대해서 기술하였다. '핵심어 전달율 측정'은 번역대상 문장의 핵심어를 추출하여 그 핵심어의 대역아이템을 찾아내고, 그 대역아이템이 기계번역 결과 내에 전달된 정도를 측정하는 평가방법이다. 기존의 기계번역 평가 척도와는 다른 의도전달에 초점을 맞춘 새로운 번역 평가방법이다. 4.1절에서 기술한 바와 같은 고려사항들을 보완하면 보다 신뢰할만한 평가방법으로 자리잡을 것으로 기대한다.

REFERENCES

1) Taro Watanabe, Eiichiro Sumita, Hiroshi G(2003) : *Okuno, "Chunk-Based Statistical Translation", ACL2003 Proceedings, pp303-310*
2) Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu(2001) : *"Blue : a Method for Automatic Evaluation of Machine Translation", IBM Research Report*
3) 이민행 : 기계번역시스템 평가방안 연구사례, 연세대
4) Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics, NIST, 2002. http : //www.nist.gov/speech/tests/mt/doc/ngram-study.pdf
5) TIDES, http : //www.ldc.upenn.edu/Projects/TIDES/

추출된 핵심어	대역아이템 검색	MT 결과	Score
기다려/pvg 감사/ncpa	Wait, Thank you	Thank you for waiting.	0.5
방/ncn 열쇠/ncn 여기/pvg	Room, key, consider	Consider room key.	0.7
삼/nnc 층/nbu 방/ncn 314/nnn 호실/nbu	Three, class, room, 314, room number	Is room room number 314.	0.6
그/mms 사람/ncn 옷/ncn 사/pvg 저/npp 선물/ncn 줘/pvg	That, man, Clothes, buy, i, gift, give	Man is you draw clothes for me I like gift give.	0.7
제일/mag 유명/ncps 백화점/ncn 어디/npd	Great, famous, department store	Where is department store famous great?	1.0

Fig. 4. 핵심어 전달율 scoring의 예.