Translation Quality Estimation Task의 Baseline System 성능 개선

이소현, 이종혁
포항공과대학교 컴퓨터공학과

Translation Quality Estimation Task 개요

❖ Source 문장과 target 문장만을 이용하여 자동으로 문장 번역의 질을 평가

❖ QuEst (baseline system) 성능 향상
→ 번역 품질을 잘 예측하는 데에 도움을 줄 수 있어 기계 번역 기술을 더욱 향상
→ 번역에 대한 post-editing 과정에서 소요되는 시간과 노력을 경감

QuEst (Baseline System)

1. Feature Extraction Module

- Source 문장의 총 단어 수
- Target 문장의 총 단어 수
- Source 문장의 단어들의 길이 평균
- Source 문장의 language model 확률
- Target 문장의 language model 확률
- Target 문장의 총 단어 수 / Target 문장의 unique 단어 수
- Source 문장 단어의 probability 합 / Source 문장의 총 단어 수 (probability threshold: 0.2)
- Source 문장 단어의 probability 합 / Source 문장의 총 단어 수 (probability threshold: 0.01)
- Source 문장의 unigram 중 language model 빈도수가 제1사분위수 이하인 unigram의 비율
- Source 문장의 unigram 중 language model 빈도수가 제3사분위수 이상인 unigram의 비율
- Source 문장의 bigram 중 language model 빈도수가 제1사분위수 이하인 bigram의 비율
- Source 문장의 bigram 중 language model 빈도수가 제3사분위수 이상인 bigram의 비율
- Source 문장의 trigram 중 language model 빈도수가 제1사분위수 이하인 trigram의 비율
- Source 문장의 trigram 중 language model 빈도수가 제3사분위수 이상인 trigram의 비율
- SMT training corpus에 존재하는 source 문장의 unique unigram 개수 / source 문장의 unique unigram 개수
- Source 문장의 구두점 개수
- Target 문장의 구두점 개수

2. Machine Learning Module

- Support Vector Machine (SVM)

❖ 평가 방법

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

Dataset

❖ WMT'12 Quality Estimation (QE) dataset

문장 단위의 영어-스페인어 번역문 및 번역 평가 점수

(Source: 영어, Target: 스페인어, 번역 평가 점수: 1.0 ~ 5.0점 소수점 첫째자리실수, Training data: 1832, Testing data: 422)

Feature 1： Simple Word List 이용

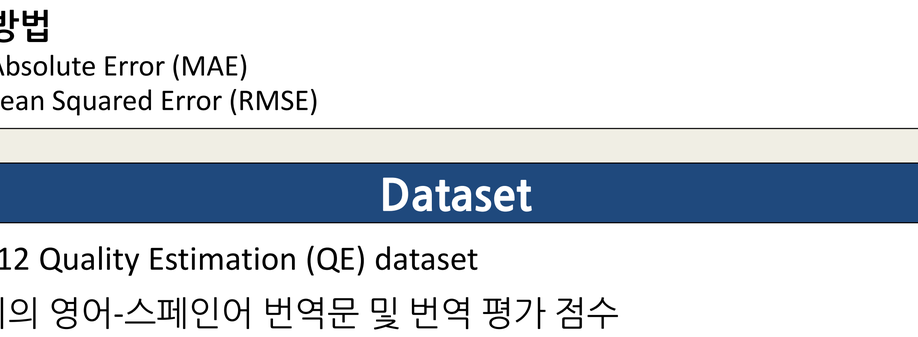
❖ 문장에서의 복잡한 단어의 비율

복잡한 단어일수록 더 많은 의미를 담고 있기 때문에 이들이 잘 유지되는 것만으로도 대체적인 의미가 통할 것이라고 생각

- Resource: Wiktionary 의 1000 basic English words, 번역기를 이용하여 영어 목록에 대응하는 스페인어 단어 목록을 구성
- **Feature7001: Source 문장의 복잡한 단어 개수 / source 문장의 총 단어 개수**
- **Feature7002: Target 문장의 복잡한 단어 개수 / target 문장의 총 단어 개수**
- **Feature7003: Feature 7001의 값 / Feature7002의 값**

Feature 2： Sentence Vector (word2vec)

❖ **Word2vec** (영어, 스페인어 각각에 대하여)



(Mikolov et al., NAACL HLT, 2013)

- Training dataset: Europarl Parallel Corpus ver. 7 (영어-스페인어 parallel corpus)
- Vector dimension: 30
- Dataset에 10번 이상 등장한 단어

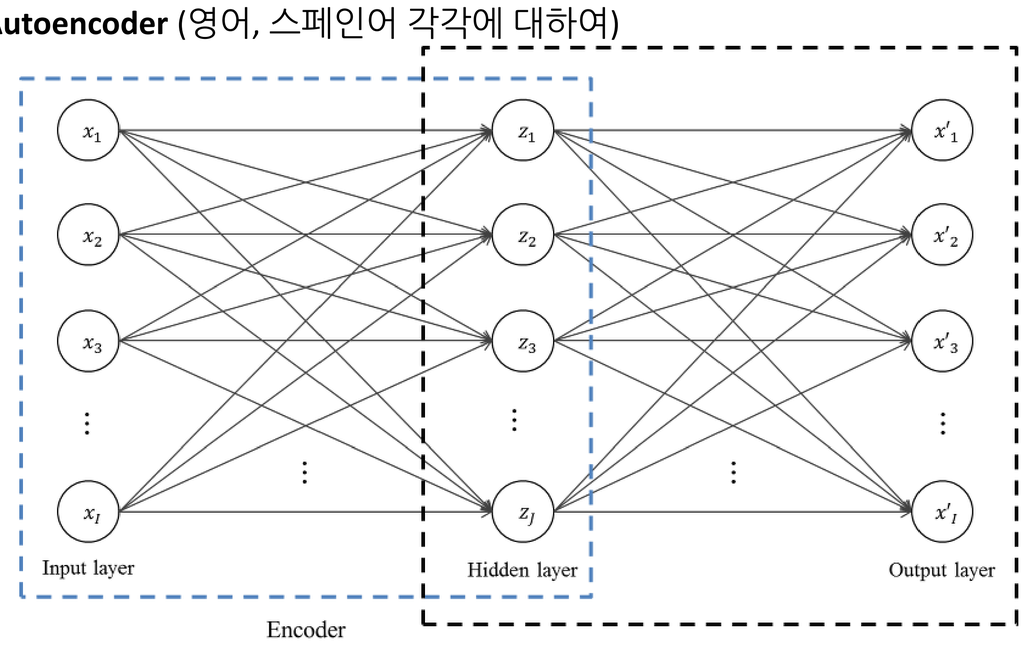
→ 영어 word2vec 총 단어 수: 30802, 스페인어 word2vec 총 단어 수: 48258
(WMT'12 QE training data 총 단어 집합의 약 80.6 %의 단어를 포함)

➔ Output: (문장 단어 수 * word vector dimension)를 dimension으로 하는 vector

각각의 차원들의 값을 feature로 추가하면 학습하기에는 너무 많은 feature 가 생성
→ dimension을 학습하기에 효율적인 값만큼 낮추기 (autoencoder 이용)

Feature 2： Sentence Vector (autoencoder)

❖ **Autoencoder** (영어, 스페인어 각각에 대하여)



- Input, Output: WMT'12 QE training data
- Input size: 가장 긴 문장의 길이 (0으로 padding)
- Training: 20 epochs, Hidden state dimension 4, 8, 16 → **8, 16, 32개의 feature**

Random Lasso

❖ Feature selection method

밀접하게 연관되어 있는 feature 사이에서의 feature selection에 이점을 가진다 [Wang et al. (2011)]

연구 결과

System (num of features)	MAE	RMSE
BL (17)	**0.7070**	**0.8564**
BL + 60 features (77)	0.7176	0.8756
BL + 60 features → RL (54)	0.7163	0.8751
BL + 60 features → RL (54)	0.7225	0.8783
BL + 60 features → RL (25)	0.7251	0.8802
BL + 7001, 7002 (19)	0.7088	0.8588
BL + 7001, 7002, 7003(20)	0.7103	0.8598
SV4 (8)	0.8064	0.9756
BL + SV4 (25)	0.7105	0.8684
BL + SV4 → RL (21)	0.7165	0.8729
SV8 (16)	0.7876	0.9607
BL + SV8 (33)	**0.7052**	0.8605
BL + SV8 → RL (32)	0.7085	0.8663
BL + SV8 → RL (29)	**0.7001**	0.8565
BL + SV8 → RL (25)	**0.7026**	0.8593
SV16 (32)	0.7955	0.9564
BL + SV16 (49)	**0.7062**	0.8583
BL + SV16 → RL (43)	**0.7059**	0.8586
BL + SV16 → RL (33)	**0.7068**	0.8587
BL + SV16 → RL (23)	0.7169	0.8661

결론 및 토론

QuEst에는 존재하지 않는 의미적인 요소를 가진 feature를 추가함으로써 baseline의 성능을 증가시킬 수 있었다. 특히, baseline에 dimension 8의 sentence vector를 feature로 추가하고 random lasso로 29개를 선택하였을 때 가장 좋은 성능을 내었다.

Machine learning module에서 사용한 SVM 기법은 인공 신경망 기법에 비해 비교적 성능이 안 좋을 수 있지만, 빠르게 학습하고 사용할 수 있다는 장점이 있다.

QuEst에서 사용되는 feature은 인공 신경망 기법에도 적용할 수 있으므로, 이번 연구의 결과는 인공 신경망 quality estimation의 성능을 올리는 데에도 도움이 될 것이다.

❖ 보완할 점

- Word2vec: 더 많은 data로 training, word vector의 dimension을 조정해보며 feature 추출을 위한 최적의 dimension을 찾기
- Autoencoder: 더 많은 data로 training, hidden state의 층과 개수를 변형하며 최적의 model을 찾기, RNN encoder로도 dynamic RNN을 이용하여 padding이 필요 없는 encoder를 만들기