A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios

Sunyoung Song

Abstract

- Deep neural networks, huge language model 이 NLP 분야에서 많이 보이고 있음
- 대부분의 모델이 많은 양의 데이터를 필요로 하기 때문에 **low-resource** 상황에서 문제가 있음

- 대표적으로 NLP분야에서 쓰이는 방법으로는 Large-scale에 pre-train 후, downstream-task에 fine-tuning

- **본 논문에서는 마찬가지로 low-resource 상황에서 사용할 방법을 다룸**
  - ex. 전이학습(transfer learning), 데이터 증강(data augmentation), distant supervision

1. Introduction

- 이 논문에서의 **low-resource는 널리 쓰이지 않는 언어이거나, 널리 쓰이더라도 NLP분야에서 연구되지 않은 language**를 의미

- The term “language” in this paper also includes domain-specific language.
  - **low-resource ‘language’는 ‘language’ 뿐만 아니라 low-resource ‘settings’ (domain, task, …)도 포함**
  - 즉, 영어처럼 널리 쓰이는 언어이더라도 데이터가 부족한 domain, task, … 등도 포함
    - ex. 의학

- low-resource 상황을 극복하기 위해 low-resource 상황과 테크닉에 대한 이해가 필요함

1. Introduction

Method	Requirements	Outcome	For low-resource languages	For low-resource domains
Data Augmentation (§ 4.1)	labeled data, heuristics*	additional labeled data	✓	✓
Distant Supervision (§ 4.2)	unlabeled data, heuristics*	additional labeled data	✓	✓
Cross-lingual projections (§ 4.3)	unlabeled data, high-resource labeled data, cross-lingual alignment	additional labeled data	✓	✗
Embeddings & Pre-trained LMs (§ 5.1)	unlabeled data	better language representation	✓	✓
LM domain adaptation (§ 5.2)	existing LM, unlabeled domain data	domain-specific language representation	✗	✓
Multilingual LMs (§ 5.3)	multilingual unlabeled data	multilingual feature representation	✓	✗
Adversarial Discriminator (§ 6)	additional datasets	independent representations	✓	✓
Meta-Learning (§ 6)	multiple auxiliary tasks	better target task performance	✓	✓

Table 1: Overview of low-resource methods surveyed in this paper. * Heuristics are typically gathered manually.

2. Aspects of “Low-Resource”

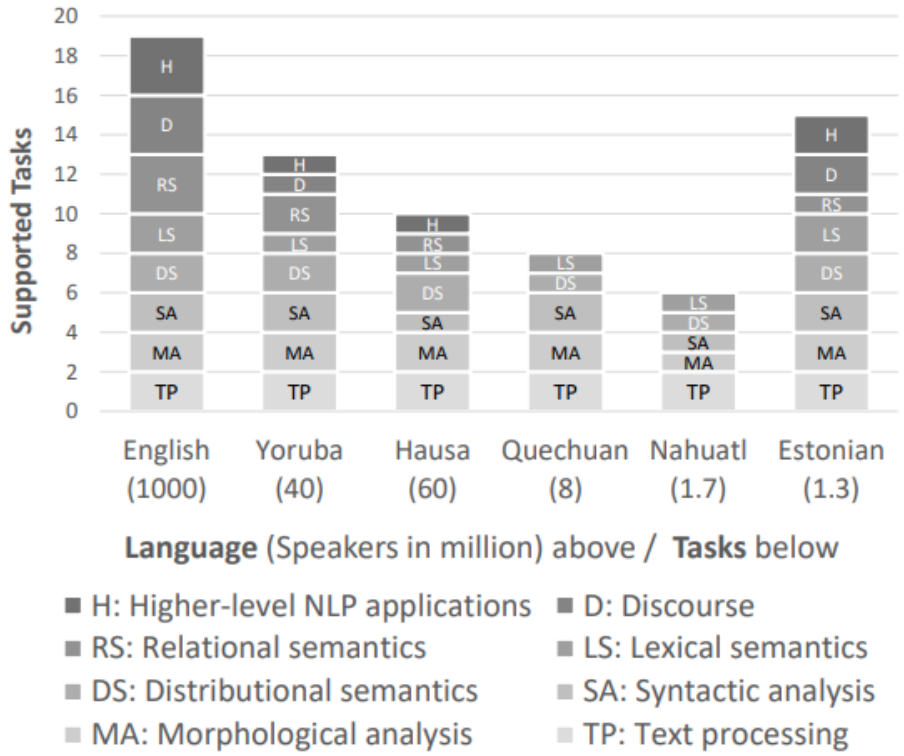


Figure 1: Supported NLP tasks in different languages. Note that the figure does not incorporate data quality or system performance. More details on the selection of tasks and languages are given in the appendix Section B.

3. Dimensions of Resource Availability

- low-resource settings를 범주화 함
  - **(1) availability of task-specific labels**
  - **(2) availability of unlabeled language – or domain-specific text**
  - **(3) availability of auxiliary data**
    - Transfer learning 은 다른 언어나 다른 분야의 task-specific label 활용 가능
    - Distant Supervision은 외부 지식 활용 가능
    - 여러 NLP tools을 사용해 데이터를 만들 수 있음
      - ex. 한국어 감성분석 데이터가 없을 경우, 영어 감성분석 데이터를 번역하여 한국어 감성분석 데이터를 만드는 방법

4. How Low is Low-Resource?

- low-resource의 경계를 나누는 여러 threshold가 있음

- Part-of-speech (POS) tagging : 1000~2000개로 제한하기도 함
- Text Generation task : 350K

  -> 같은 양의 데이터라도 **task에 따라, 사용된 언어에 따라** 성능이 다르기 때문에**, low-resource의 기준도 다름**

5. Generating Additional Labeled Data

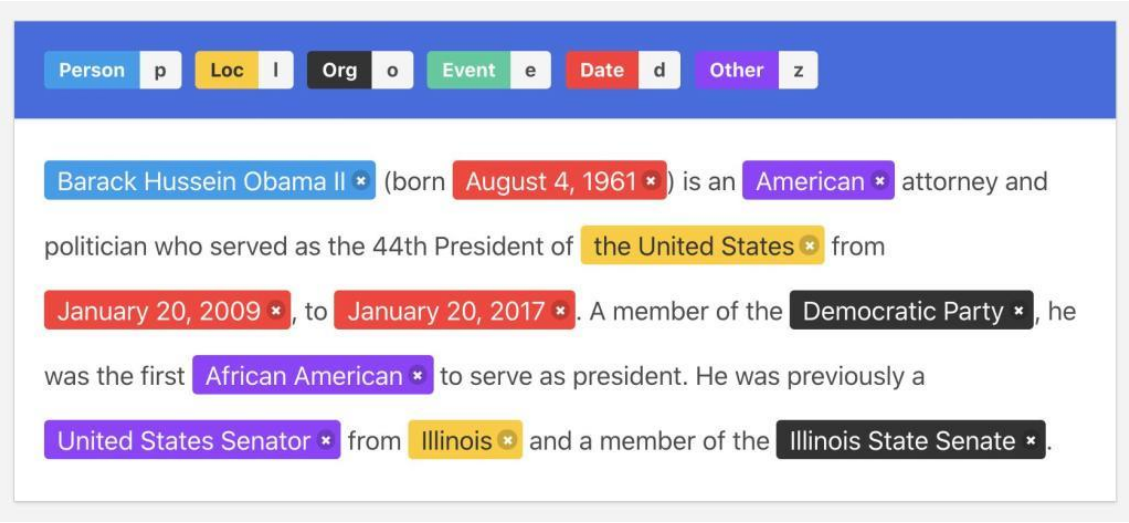
- **(1) Data Augmentation**
  - 기존에 존재하는 labeled data를 이용하여 특징을 변환하여 같은 라벨을 갖는 새로운 data 생성

  - Text Data Augmentation Methods
    - **Token-level Methods**
      - token을 동의어로 대체
      - token을 같은 type의 entities로 대체
      - token을 같은 형태소를 공유하는 단어로 대체
      - token을 맥락을 고려한 language model을 사용해 대체

    - **Sentence-level Methods**
      - **label을 바꾸지 않는 작업**
        - dependency tree 부분 조작
        - 부분 문장 제거하여 문장 단순화
        - Subject-object relation 전환 (ex. 피동형 <- > 사동형)
      - **back-translation (역번역)**
      - **Adversarial methods** : text의 의미를 바꾸지 않는 선에서 input에 간섭을 주는 방식

5. Generating Additional Labeled Data

- **(2) Distant & Weak Supervision**
  - data augmentation과는 달리, unlabeled text를 이용하고 text를 수정하지 않음
  - Unlabeled text에 대응하는 label은 외부 정보 지식(dictionary, gazetteer, …) 을 활용해 (반)자동적으로 얻음
  - NER(Named Entity Recognition)이나 RE (Relation Extraction) 같은 정보 추출에 널리 사용되고 있음 -> 이를 위한 보조 데이터 (auxiliary data)가 잘 마련되어 있음



<NER>

Sentence	Relation
1. **Steve Jobs** and Wozniak co-founded **Apple** in 1976.	*Founder*
2. **Michael Jordan** is an American retired professional **basketball player**.	*Career*
3. **Washington D.C.** is the capital of **United states**.	*CapitalOf*
.....	.....

<RE>

5. Generating Additional Labeled Data

- **(3) Cross-Lingual Annotation Projections**
  - task-specific classifier는 high-resource language를 대상으로 학습됨
  - low-resource에 대해서는?
    1. **Unlabeled low-resource data를 high-resource data로 대응시킴**
    2. **Task-specific classifier을 사용해 대응시킨 high-resource data에 대한 라벨을 얻음**
    3. **그렇게 얻은 label을 unlabeled low-resource data로 project해 low-resource data에 대한 라벨을 얻음**

  - POS tagging(품사 태깅), parsing (구문 분석) task에 사용됨

  *** 주의할 점**
  - Unlabeled low-resource data를 high-resource data로 대응시킬 때, 대응(parallel) 외에 기계 번역 방식을 사용하기도 하는데 low-resource language에서는 잘 작동하지 않을 수도 있다는 문제가 있음
  - 언어에 따라 특정 domain에 따라 (ex. 정치/종교적 domain) 에서는 대응할만한 말(parallel copora)이 없는 경우도 있음

5. Generating Additional Labeled Data

- **(4) Learning with Noisy labels**
  - 완벽하지 않은 모델을 사용해 라벨링 하는 것은 noise가 쌓이게 됨 -> 에러 중첩
  - Distant supervision에서는 이런 문제를 해결하기 위해 2가지의 noise handling 기법 사용
    1. **Noise filtering**
       - 잘못된 라벨일 확률이 높은 학습 데이터를 지워버리는 방법 (completely filtering)
         - Through a probability threshold
         - Through a binary classifier
         - Use of a reinforcement-based agent (강화학습 기반)
       - 노이즈 데이터를 약간 조정해 사용하는 방법 (soft filtering)
    2. **Noise Modeling**
       - Clean label과 noisy label 간의 관계를 파악하기 위한 confusion matrix를 활용하여 noise model을 추가해 noisy label을 clean label distribution으로 shift 하는 방식

- **(5) Non-Expert Support**

6. Transfer Learning

- Unlabeled data를 활용해 language representation을 pre-train시킨 BERT 같은 모델이 최근 많이 쓰이고 있음

- **(1) Pre-trained language representations**
  - input으로는 feature vectors를 넣음
  - **subword-based embeddings**
    - fastText
    - N-gram embeddings
    - Byte-pair-encoding embeddings
    - OOV 문제를 해결하기 위해 original word를 몇 개의 subword로 나눠 조합해 original word를 나타내는 방식
    - Low-resource sequence labeling tasks에서도 좋은 결과를 보임

  - Pre-trained transformer에는 BERT나 RoBERTa 등이 있음
    -> 이런 모델들은 주로 unlabeled data가 많고, task-specific labeled data는 적은 low-resource language를 다룰 때 좋음

6. Transfer Learning

- **(2) Domain-Specific Pre-Training**
  - Specialized domain에 쓰이는 language와 general domain에 쓰이는 language랑 너무 다른 경우, 특정 도메인 자체가 low-resource가 되는 상황이 생김
    - ex. Scientific articles (기호, 이론), Bio-Medical (전문 용어) …

  - 이런 상황이 domain-gap 문제를 일으킴 -> target domain에 adaptation을 통해 문제 해결
    - 즉, language model을 fine-tuning
    - 모델을 domain or task-adaptive한 방식으로 계속해서 unlabeled data를 pre-train할 경우, high-resource든 low-resource든 여러 domain과 task에서 성능이 높아짐을 보인 연구가 있음

  - general domain의 high-resource embeddings를 specific한 target domain의 low-resource embeddings에 활용 가능

- **(3) Multilingual Language Models**
  - Single model을 다양한 언어에 학습시킴
  - Pre-train 과정에서 다양한 언어를 접하기 때문에 다차원 언어를 다뤄야하는 task에 적합

7. Ideas From Low-Resource Machine Learning in Non-NLP Communities

- **(1) Meta Learning**
  - Task-specific한 모델이 아닌 다양한 task에 잘 작동하는 multi-task learning을 목표로 함

- **(2) Adversarial methods**
  - Adversarial Discriminator를 사용하여 모델이 data-specific한 feature representation을 배우는 것을 방지하여, general feature representation을 배우도록 함 (다른 domain에도 잘 활용할 수 있게 하기 위해)

8. Conclusion

- Low-resource NLP에 대한 최근 연구들을 다룸
- 연구자들에게 올바른 방법, 기술 등을 선택할 수 있도록 가이드라인을 줌
- 이런 테크닉을 연구하고 평가할 때는 data availabilty 고려해야 함.
  - 특정 domain이나 특정 language마다 필요한 data 규모가 다르기 때문에 각 상황에 맞게 적절한 테크닉을 사용해야 함

Thank You

감사합니다.