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

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, Dietrich Klakow
Outline

- Introduction
- Low-Resource Settings
- Data Generation
- Transfer Learning
- Ideas from Low-Resource Machine Learning
Introduction

- Current research on NLP mostly focuses on 10 to 20 high-resource languages (e.g., English).

- Thousands of languages with billions of speakers are ignored.

- State-of-the-art NLP models rely on data-hungry deep learning models.
• "Low-resource" learning is broad:
  – Threatened languages (e.g., Yongning Na with 40k speakers and only 3k written, unlabeled sentences).
  – Languages that are widely spoken and seldom addressed by NLP research: more than 310 languages exist with at least one million speakers each.
  – There are 300 languages in Wikipedia.
  – Even for popular languages, some domains might have very little data (e.g., domain-specific language).
Research Motivation

• Overcome the lack of labeled data in low-resource settings by leveraging further sources.
  – Unlabeled data
  – Manual heuristics
  – Cross-lingual alignment

• This presentation: highlighting the underlying assumptions of these techniques.
Aspects of “Low-Resource”

- Lacking data foundation for high-level tasks for low-resource languages.

- Basic tasks for low-resource languages might be possible (tokenization), but training data tends to be of lower quality (compared to English) of very limited size.

- Four American and African languages (with 1.5 to 60 million speakers) are less studied than the Estonian language (with 1 million speakers).

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.
Dimensions of Resource Availability

• Availability of task-specific labels (most prominent).

• Availability of unlabeled language- or domain-specific text.

• Availability of auxiliary data (e.g., knowledge base or gazetteer for distant supervision, machine translation tools).
How Low is Low-Resource?

• For task-specific labels, sometime thresholds are used to define low-resource.
  – For POS tagging: studies use thresholds ranging from 1k to 60k labeled tokens.
  
  – Threshold is also task-dependent; more complex tasks increase resource requirements.
    • For text generation: studies used 350k labeled training instances to frame low-resource learning settings.

  – Defining hard low-resource thresholds for tasks is hard.
    • More work should evaluate low-resource techniques across different levels of data availability for better comparison between approaches.
Dealing with the lack of task-specific labels

- Data Augmentation: using task-specific instances to create more of them.
- Distant supervision: labeling unlabeled data using knowledge.
- Cross-lingual Projections.
Data Augmentation

• Modifying existing data instances with transformations that do not change the label.
  – At word level: replacing words with equivalents, i.e., synonyms, entities of the same type, words that share the same morphology; language models might be used to guide and consider the context.

  – At sentence level:
    • manipulating parts of the dependency tree;
    • removing sentence parts;
    • inversion of subject-object relation;
Paraphrasing through Back-Translation

- Target sentences are back-translated into source sentences.

- Errors in the source side/features do not have a large negative effect on the generated target text, e.g., in text generation tasks (summarization, table-to-text generation).

- Some work also used this for text classification.

*Figure 1: Creating a synthetic parallel corpus through back-translation. First, a system in the reverse direction is trained and then used to translate monolingual data from the target side backward into the source side, to be used in the final system.*
Using Language Models

- Seem to be emerging and might be applicable for finer granularity (token-level) task.
- For text classification: trained a language model by conditioning on a label (e.g., on the subset of the data with the same label). This will generate new data of the corresponding label.
- For sequence labeling: linearizing labeled sentence:

Our work on event detection:

Table 5: Generated sentences by GPT-2 for different datasets. Event triggers are shown in boldface that are surrounded by the special tokens TRG_s and TRG_e generated by GPT-2.
Adversarial Methods

• Often used to find weakness in machine learning models.

• Can be used to augment NLP tasks:
  – Apply small perturbations to the input data that do not change the meaning of the text.
  – Often applied to the word representation levels.
Data Augmentation for NLP

- Not yet found widespread use in NLP (compared to vision).

- Not yet a unified framework that allows applying data augmentation across tasks and languages.

- For NLP, insights of linguistic or domain experts in low-resource settings might be necessary.
Distant Supervision

- Very popular for named tagging and relation extraction.
  - Start with dictionaries of names or knowledge bases (Freebase) and align it with unlabeled text to automatically obtain data.
  - Other heuristics to generate data are possible.
Distant Supervision

- Less popular for other tasks. Other works include:
  - Using dictionary of POS tags for POS tagging.
  - Using simple bag-of-word classifier on a list of seed words and training a deep learning model on its weak supervision (for aspect classification).
  - Using context by transferring document-level label to sentence its sentence-level instances (for sentiment analysis).
  - Using meta-data for text classification.
  - Rephrasing labels in to sentences with simple rules; pre-trained language models then decide the matches of label sentences and unlabeled inputs (i.e., which label sentence mostly likely follow an unlabeled input, -- it was great/bad).
Distant Supervision

- Need reliable knowledge sources and heuristics; might not be available for different NLP tasks and low-resource settings.

- For instance, the large performance between high-resource and low-resource languages for POS is attributed to the lack of high-coverage and error-free dictionaries for weak supervision in low-resource languages.

- Knowledge sources and heuristics require human intervention. Can we just use that effort to label data (e.g., via active learning) instead? The data quality might be higher.
Cross-Lingual Annotation Projections

- Task-specific classifiers are trained for a high-resource language.

- Using parallel corpora, unlabeled low-resource data is aligned to its equivalent high-resource language where labels can be obtained by the above classifiers.

- Projecting the labels (in high-resource texts) back to low-resource languages.

- Applied for: POS tagging, parsing.

- Parallel texts: OPUS, Bible, JW300, or machine translation systems.
Cross-Lingual Annotation Projections

- Require high-quality classifiers for high-resource languages and resources for data alignment.

- High-quality MT systems might not be available.

- Parallel texts tend to be available for specific domains (e.g., political proceedings, religious texts).

- Some potential solutions: only use word translations, bilingual dictionaries, task-specific seed words.
Transfer Learning

- Pre-trained Language Representations
  - Pre-trained word or sentence representations on large-scale unlabeled texts result in improved performance for downstream tasks.
  - Subword-based embeddings (fastText) and byte-pair encoding embeddings addresses out-of-vocabulary issues.
  - Pre-trained word embeddings are published for more than 270 languages.
  - Multilingual pre-trained language models (mBERT, XLM-RoBERTa) leads to significant improvement for multilingual learning.
Pre-trained Language Representations

• How can these methods be helpful for low-resource scenarios where hardware is also low-resources?
  – Can pruning or model compression help?
  – Low- to medium-depth transformer sizes perform better than larger models for low-resource languages.
  – Train models with three orders of magnitude fewer parameters that perform on-par with large-scale models like GPT-3 on few-shot task by reformulating the training task and using ensembling.

• data quality for low-resource, even for unlabeled data, might not be comparable to data from higher-source languages.
  – word embeddings trained on larger amounts of unlabeled data from low-resource languages are not competitive to embeddings trained on smaller, but curated data sources.
Domain-Specific Pre-Training

• The majority of recent language models are pre-trained on general-domain data, such as texts from the news or web-domain, which can lead to a domain gap.
  – Fine-tuning language models on domain-specific data (BioBERT, SciBERT).
  – Fine-tune language models on language-specific data might be better than multilingual pre-training.

• Powerful representations can be achieved by combining high-resource embeddings from the general domain with low-resource embeddings from the target domain.
  – attention-based meta-embeddings (using weighted sums of all embeddings).
  – Using domain-adversarial discriminator.
Multilingual Language Models

- Low-resource languages can also benefit from labeled resources available in other high-resource languages.
  - Combining monolingual representations.
  - Training a single language models for multiple languages (mBERT, XLM-RoBERTa).
- Zero-shot cross-lingual transfer learning.
  - Adding a minimal amount of target-task and language data (10 –100) can lead to significant performance boost.

- The transfer between two languages can be improved by creating a common multilingual embedding space; useful for both standard word embeddings and pre-trained language models.
Multilingual Language Models

• There are still many languages not being covered by current multilingual language models (XLMR only has 100 languages).

• Low-resources languages are not well-represented in mBERT.

Figure 2: Language families with more than 1 million speakers covered by multilingual transformer models.
Ideas from Machine Learning

• Meta-Learning
  – Given a set of auxiliary high-resource tasks and a low-resource target task, meta-learning trains a model to decide how to use the auxiliary tasks in the most beneficial way for the target task.

• Language-Independent Representation Learning
  – Adversarial Training