A Comparative Study of Dictionary-based and Machine Learning-based Named Entity Recognition in Pashto

Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval

New York, NY, USA

Amir Ali (عامر علی)

Outline

- Abstract
- Introduction
- Background of Named Entity Recognition
- Pastho language and Challenge in NER
- Data Collection
- System Architecture
- Result and Discussion
- Conclusion and Future Work
- References

Abstract

- Information Extraction from text
- No research found in Pashto languages
- Build own dataset in the Pashto language
- Extract name, place, and organization
- Method:
  - Dictionary Based NER
  - Hidden Markov Model

Introduction

- Huge information is generated online in text.
- To deal with text, Natural Language processing techniques are used.
- Text Extraction is a way to automatic extraction desired information.
- One subtask of Information Extraction is Named Entity Recognition.
- In this paper, name, place, and organization extract from Pashto lang.

Example

Table 1. An example of named entity in a sentence

دونالډ ټرمپ 'ایران ته د نغدو پیسو ورکړې' له ادعا پر شا شو
Donald Trump backtracks over claim 'of cash flown to Iran'

The above sentence (Table 1) contains two named entities; NER would extract them and assign them a category label as shown in Table 2.

Table 2. Named Entity Classification example

Named Entity	NE Type
(Donald Trump, دونالډ ټرمپ)	Person
(Iran, ایران)	Location

Background

Named Entity Recognition Approaches

1. Rule-Based Approach
2. Machine Learning
3. Hybrid Approach

Rule Based NER Approach

- A rule-based named entity recognition (NER) system is a method for identifying and classifying named entities in text using a set of pre-defined rules.
- The rules specify how to recognize and classify specific words, phrases, or structures in the text as named entities.
- For example, a rule-based NER system might have the following rules:
  - A word that starts with a capital letter and is not a common noun is a named entity of type "person."
  - A word that starts with "The" and is followed by a common noun is a named entity of type "organization."
  - A word that is all capital letters is a named entity of type "location."

Machine Learning Based NER Approach

- These algorithms are trained on annotated data, which is text that has been manually labeled with the named entities and their categories.
- The data is used to learn patterns and correlations between the text and the named entities, which are then encoded in the model as weights and biases.
- Once the model is trained, it can be used to classify new text by applying the learned patterns and correlations to the text.
- Machine learning-based NER can produce high-quality results, but it requires a large amount of annotated data to train.

Hybrid Based NER Approach

- Involve two or more techniques
- Dictionary-based
- Machine learning
- Rule-based
- Hybrid NER can provide high-quality results, but it can also be more complex to implement and maintain than single-technique NER

Pashto Language

- Pashto is a South-Central Asian language.
- It is one of the two official languages of Afghanistan.
- It is also spelled as Pushto and Pashtu.
- About 40-60 million people speak the Pashto language [1].
- Pashto is written from right to left. It uses a variant of Persian Arabic script.

[1] S. Waseeb, —Discussion of challenges and possible solutions in pashto nlp,|| Master's thesis, Technische Universität Berlin, 2016.

Challenges in Pashto NER

- Lack of linguistic resources
- Absence of Capital Letters
- Word Order
- Ambiguity in named entity classes

Data Collection

- Collection of 25 news articles from the BBC Pashto website [2].
  - News-related to sports
  - 7360 tokens out of 1508 are named entities (20.5%)

Table 3: Distribution of named entities

Named Entity	Ratio	Count
Person	9.4%	691
Location	0.9%	66
Organization	10.2%	751

[2] https://www.bbc.com/pashto/

System Architecture

Dictionary-Based NER Approach

- **Data annotation:** Manually Data Annotation
- **Gazetteers:** Three gazetteers like person 3232, location 346, 133 organization.
- **Text Pre-processing:**
  - Input text is prepared for lookup in dictionaries
  - Word Tokenization
  - Computing n-gram (unigram, bigrams, trigrams)
- **Dictionary Lookup:**
  - Lookup techniques are searching or matching
    - Exact matching

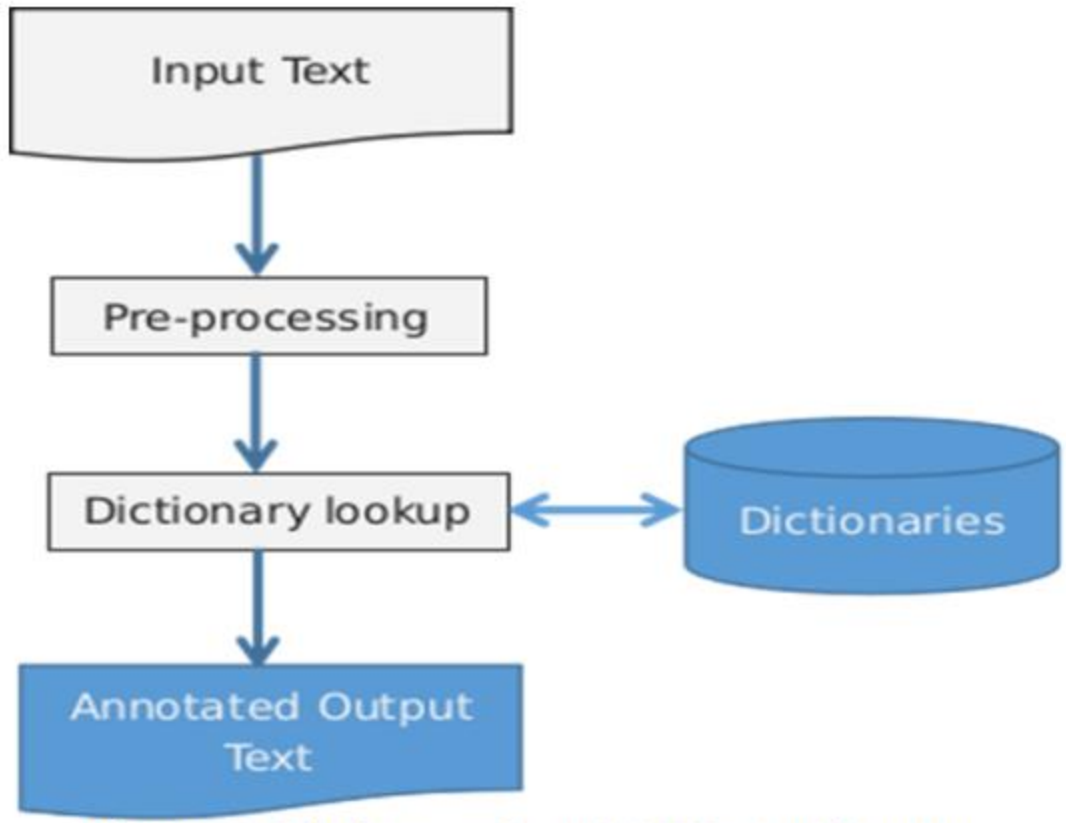

Figure 1: Dictionary-based NER Architecture

HMM-based NER

- HMM is very effective because it calculates the highest probability tag sequence of a given sequence of words.

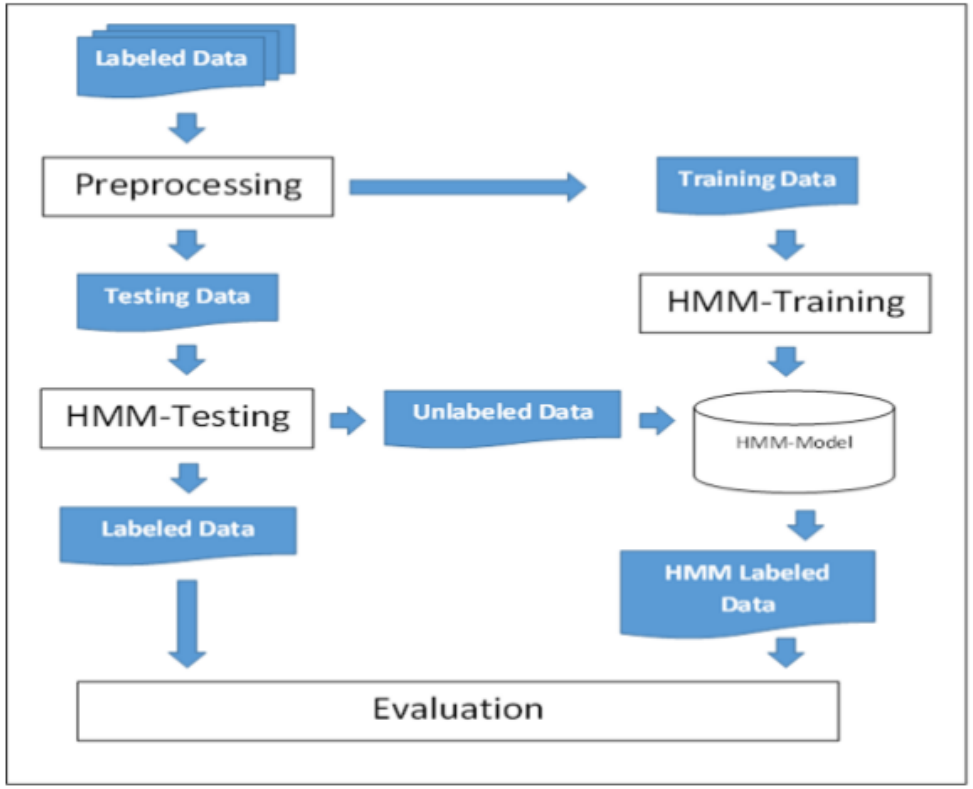

Figure 2. Architecture of HMM-based NER

Result Evaluation

Table 4: Result of Dictionary-based NER

K-Gram	Precision	Recall	F-measure
Unigram	0.37	0.29	0.33
Bigram	0.81	0.47	0.60

Table 5: Result of HMM

Entity Type	Precision	Recall	F-Measure
Person	1	0.65	0.79
Location	1	0.3	0.46
Org	0.98	0.77	0.86
Overall	0.98	0.7	0.82

Table 6: Comparison of Dictionary vs HMM

	Precision	Recall	F-Measure
Bigram	0.81	0.47	0.60
HMM	0.98	0.70	0.82

Conclusion and Future Work

- In this paper author implemented two techniques to extract the name, place, and organization, and HMM performed well compared to the dictionary-based approach.
- In the future, the author aim to extend the corpus by collecting data from other domains, such as politics and economics. It would help in building a better classifier and better error estimation. In addition, it would be interesting to do research on developing a hybrid NER for Pashto.

References

[1] S. Waseeb, —Discussion of challenges and possible solutions in pashto nlp,‖ Master's thesis, Technische Universität Berlin, 2016.

[2] https://www.bbc.com/pashto/

[3] R. Momand, S. Waseeb, and A. M. L. Rai, "A comparative study of dictionary-based and machine learning-based named entity recognition in pashto," in Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, NLPIR 2020, (New York, NY, USA), p. 96–101, Association for Computing Machinery, 2021.

Thank you ☺

Do you have any Questions?