CSE431 PAPER REVIEW

THE PIPELINE PROCESSING OF NLP

Submitted to: Annajiat Alim Rasel, Senior Lecturer, BRAC University

Submitted by:
Name: Hazra Mohammed Ahnaf Faiyaz
ID: 17241014
Group: 15
Task 2

RA: Md. Sabbir Hossain
ST: Mehnaz Ara Fazal

INTRODUCTION

- NLP problems are complex, requiring a systematic approach for effective solutions.
- The essence lies in breaking down these intricate problems into manageable, comprehensible steps.
- Each step in the pipeline process is strategically crafted to address specific challenges within the broader NLP landscape.
- The emphasis is not just on problem-solving but on the methodical resolution of each individual aspect, ensuring a comprehensive and refined outcome.

METHODOLOGIES

Overall:

- Data Collection
- Text Cleaning
- Canonical Form Conversion
- Feature Development
- Modeling and Evaluation
- Implementation and Monitoring

Primary Stages:

- Text Segmentation
- Sentence Segmentation
- Word Tokenization
- Tokenization Challenges

KEY ELEMENTS

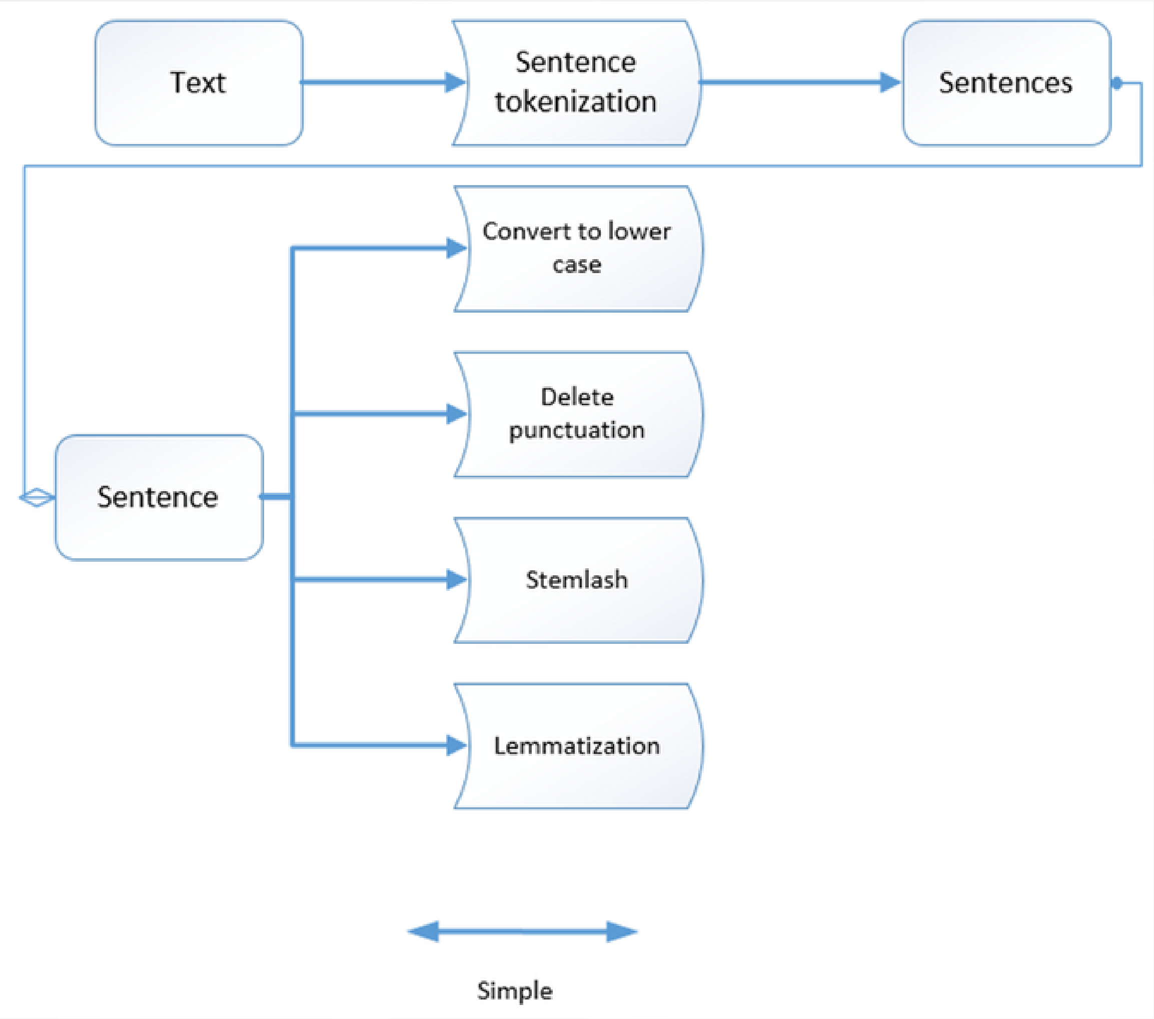

Fig.1 Common primary processing steps for a text fragment

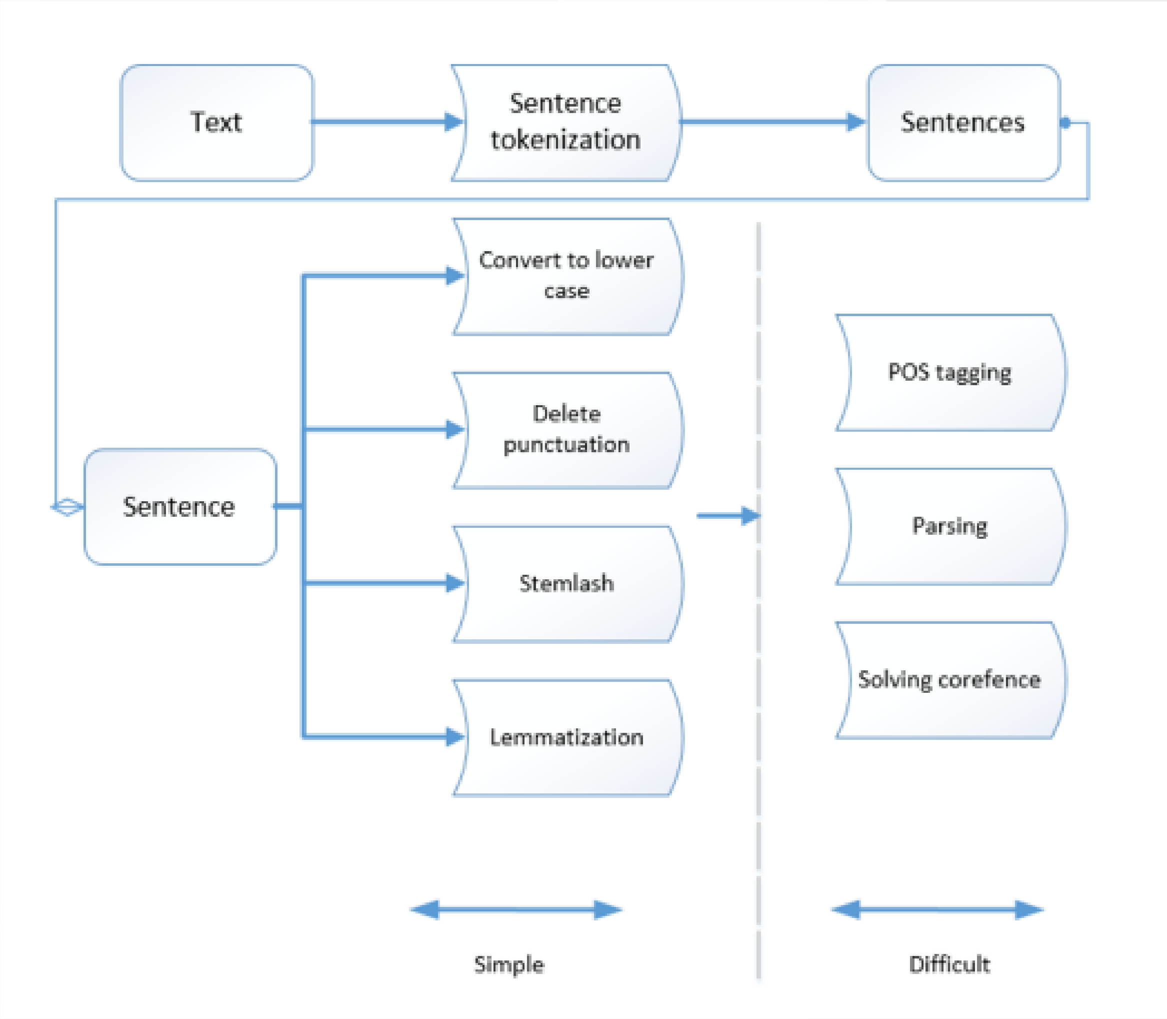

Fig.2 Extended processing steps for a text fragment

LIMITATIONS

- Existing tokenizers may not provide 100% accurate results.
- Findings may not universally apply to all NLP projects.
- NLTK library reliance may have limitations or dependencies.
- Briefly mentions monitoring and improvement without detailed optimization methods.
- Suggests customized tokenizer without elaborating on complexity or resources.
- NLTK-specific code may limit applicability to alternative tools or approaches.
- Numerical citations may become outdated which could cause potential gaps in considering newer developments.
- The study lacks addressing potential ethical considerations or biases in NLP model development.

FUTURE WORK

- Improved Tokenization Techniques
- Comparative Library Analysis
- Enhanced Text Cleaning Strategies
- Ethical Considerations and Bias Mitigation
- Optimizing Model Performance Long-Term
- Cross-Language NLP Implementations
- Automation in Customized Tokenizer Creation
- Real-Time Monitoring and Updating
- Interdisciplinary Collaboration
- Open Source Tool Development
- Longitudinal Study on Temporal Relevance

CONCLUSION

Summary:

The study focuses on NLP system development, emphasizing data collection, rule-based design, and effective text cleaning. It introduces the pipeline process, detailing stages like feature development, modeling, and constant model monitoring.
The primary stages involve text segmentation into words and sentences, highlighting challenges in sentence segmentation and word tokenization. Tokenization imperfections are acknowledged, and NLTK library usage is demonstrated for sentence and word separation.

Closing:

1. Future work suggestions encompass improved tokenization, comparative library analysis, and enhanced text cleaning strategies.
2. Ethical considerations and bias mitigation are recommended for inclusion in future studies.
3. The study could benefit from exploring cross-language NLP applications and investigating automation in tokenizer creation.
4. Real-time monitoring and updating frameworks are proposed for sustained NLP model optimization.
5. Longitudinal studies on temporal relevance and interdisciplinary collaboration are recommended for comprehensive advancements in NLP research.