Nepali NLP – Past, Present and Future

Bal Krishna Bal, PhD
Associate Professor
Information and Language Processing Research Lab(ILPRL)
Department of Computer Science & Engineering
Kathmandu University

Email: bal@ku.edu.np

10/04/2019

Contents

- How and when NLP Research got started in Nepal?
- Past NLP Projects
- Current research state and directions of NLP in Nepal
- The Future

How and when NLP research started in Nepal?

- “Necessity is the mother of invention” – Ancient Greek Philosopher, Plato.
- “From simplified Nepali Typing to an OS” – my write-up in the blog of the Association of Progressive Communications.

  https://www.apc.org/en/blog/simplified-nepali-typing-os
- Madan Puraskar Pustakalya (MPP) wanted to electronically catalogue its collection of books.
- Existing fonts like “Preeti”, “Kanchan” did not support text processing utilities in text processors like “Sort”, and “Find and Replace”.
- There was no uniformity of keyboard mapping of Nepali characters, thus making Nepali typing difficult to the general public.
- Need for developing fonts which enabled data processing facilities for Nepali with simplified keyboard mapping was truly felt.

How and when NLP research started in Nepal?

- MPP undertook the Font Standardization Project assisted by the Ministry of Science and Technology and United Nations Development Project (UNDP) in March 2002.
- This Project led to the inception of Unicode in Nepal, which is an encoding scheme that assigns unique code to every character of standard writing scripts of the world.
- The following were developed as part of the Project:
  - Unicode compliant fonts like Kalimati, Kanjirowa, Thakya Robinson, Samanata etc.
  - Two keyboard layouts, namely, Nepali Unicode Romanized and Nepali Unicode Traditional.
- Up until this time, the many potentials of text processing and computing was limited to typing in Nepal.

Past NLP Projects

- Nepali had been a very low profile language in terms of NLP resources and applications.
- Golden period: 2005-2009
- The PAN Localization Project supported by the International Development Research Center(IDRC), Canada
- The Bhasha Sanchar Project supported by the European Commission.
- First NLP applications and resources were developed:
  - HunSpell based spell checker and MyThes based thesaurus included in the OpenOffice as part of NepaLinux 1.0 in December, 2005.
  - Web-based English to Nepali Machine Translation system, Dobhase, July, 2006.
  - Nepali Text-to-Speech(TTS), Nepali Contemporary Dictionary, sabdakos, Nepali National Corpus(NNC) – 2007
  - Preliminary research works for Nepali OCR
  - Handwriting Recognition System for Nepali

Past NLP Projects

- First NLP applications and resources were developed:
  - Nepali Lexicon
  - Nepali Parts-of-Speech(POS) Tagset
  - Conversion tools
  - The Nepali Stemmer and Morphological Analyzer
  - Nepali Computational Grammar Analyzer
- Summary of the works can be found in the paper: Towards Building Advanced Natural Language Processing Applications – An Overview of the Existing Primary Resources and Applications in Nepali, http://www.aclweb.org/anthology/W09-3424

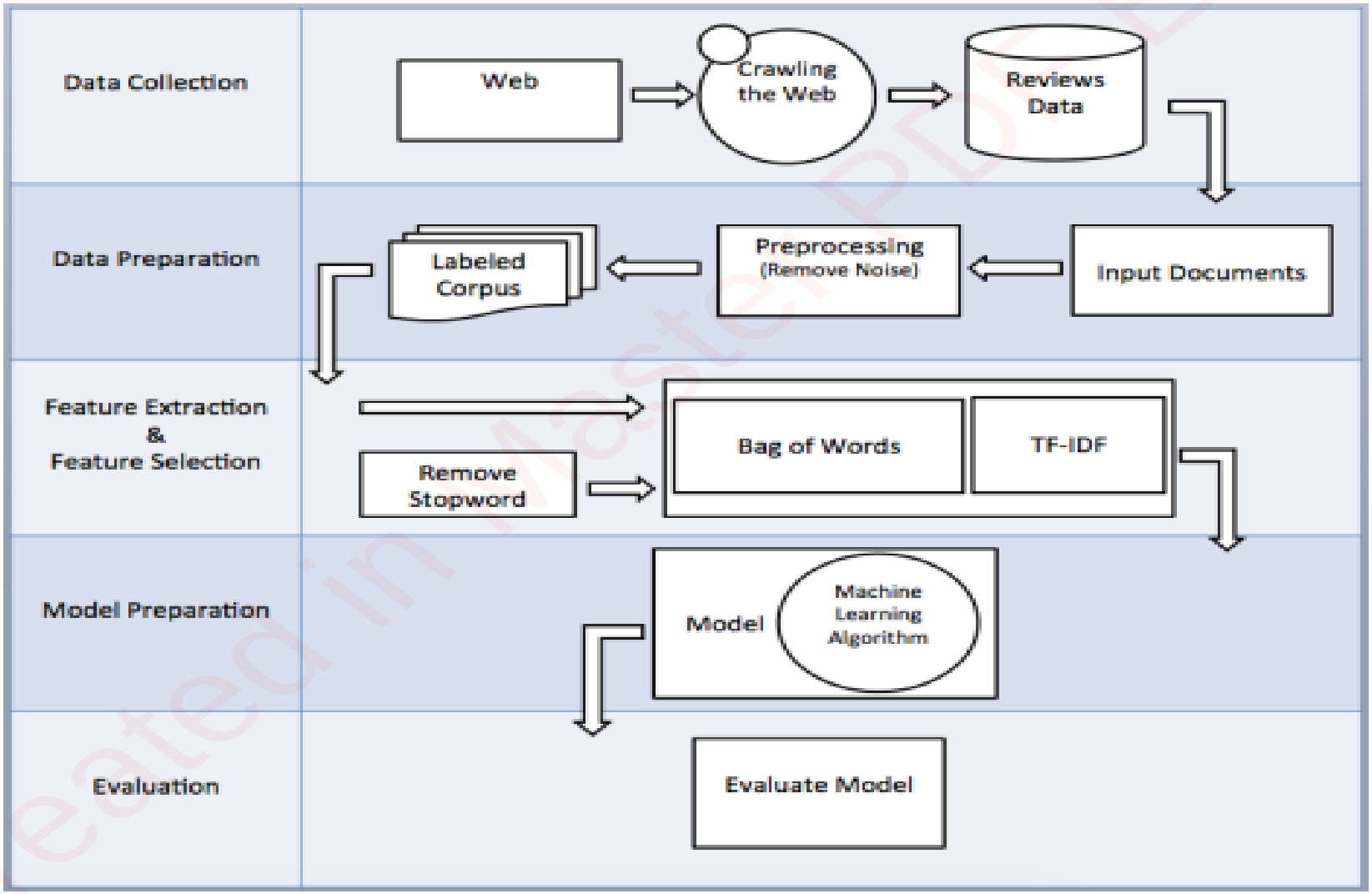
Current research state and directions of NLP in Nepal

- Compared to the research and advancements in other languages, Nepali is still an under-resourced language.
- Nevertheless, it has undergone a lot of developments lately in terms of NLP and language technologies:
  - Google Translate, Support for Nepali along with 100 plus languages.
  - Optical Character Recognition (OCR) in the Google Drive
  - Nepali TTS inbuilt with Google Translate
  - Sentiment Analysis and some advanced research
    - Sentiment Analysis in the word, sentence and document levels
    - Distinguishing between facts and opinions
    - Popularity tracking of public figures in news media

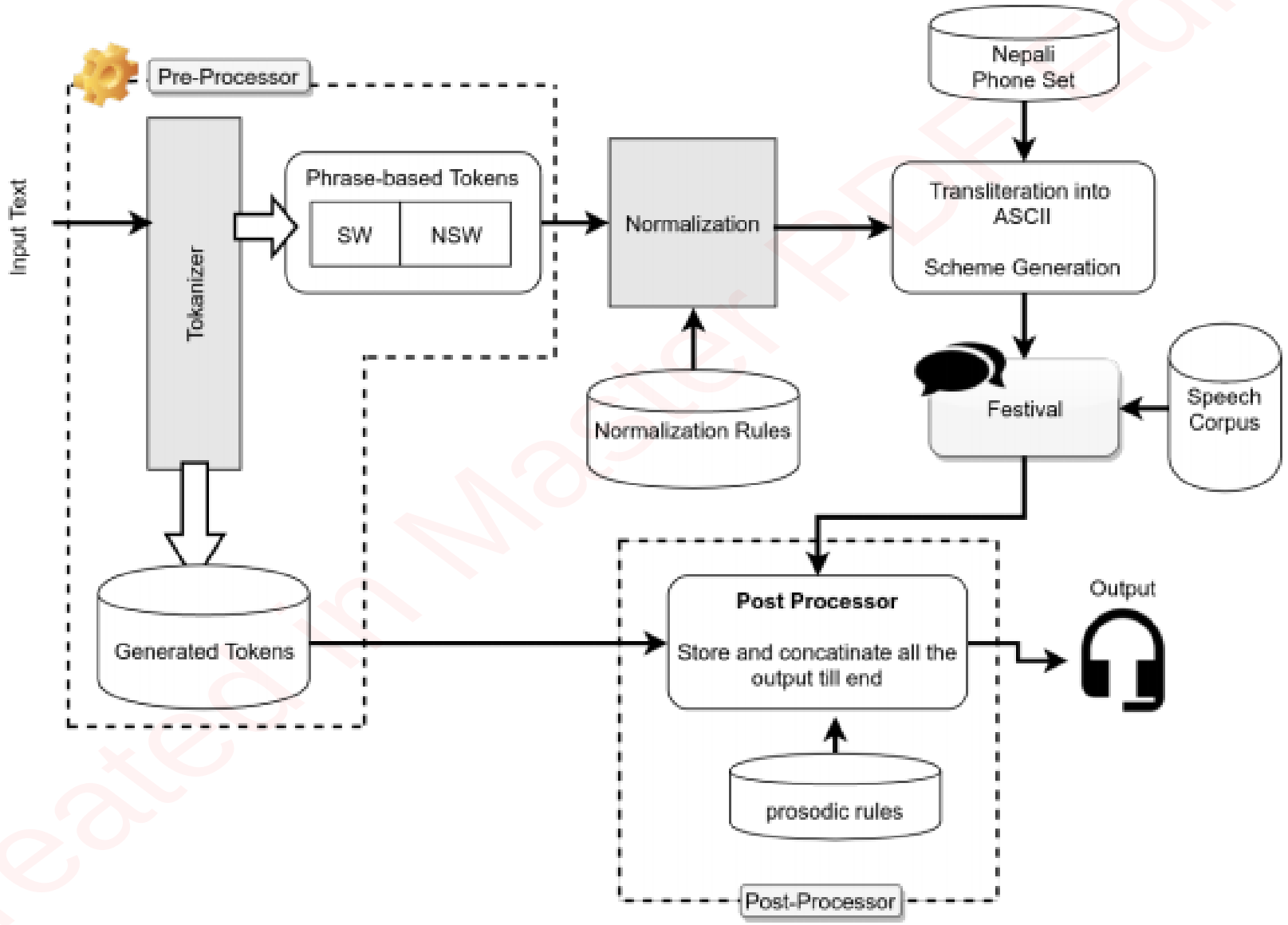
Current research state and directions of NLP in Nepal

- Nevertheless, it has undergone a lot of developments lately in terms of NLP and language technologies:
  - Deep Learning based Machine Translation System for the English-Nepali pair
  - NLP and Applied Research
    - Other currently active Projects:
      - Nepali OCR Project
      - Nepali TTS Project
    - Speech Projects
      - Speech Recognition and Processing
      - Speech-to-Speech Translation

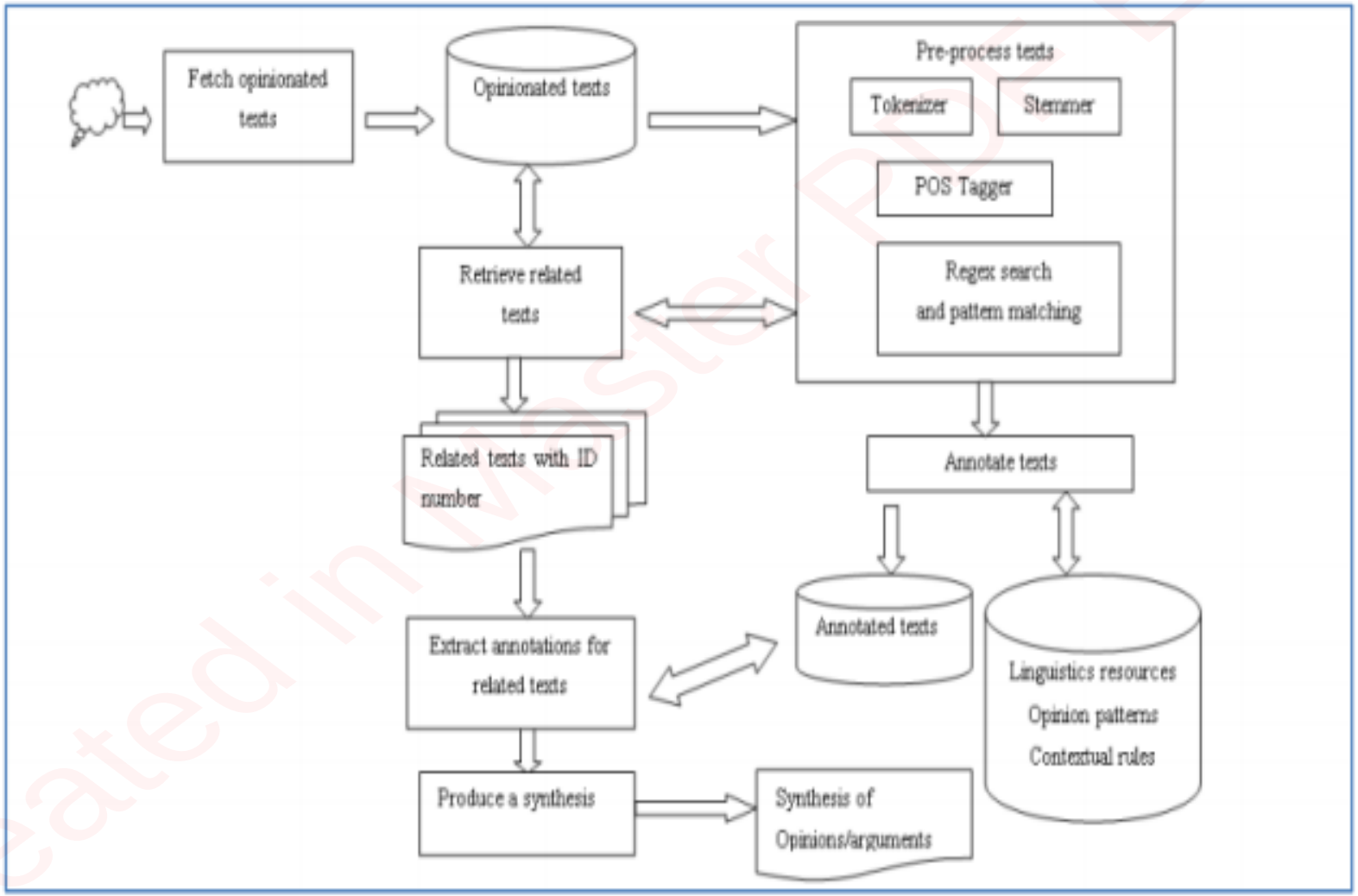
Conceptual Framework – Nepali Sentiment Analysis (Document Level)



Architecture – Nepali TTS



Architecture – News Editorials Analysis



The Future

- NLP has a huge prospect in Nepal
  - Digitization of Nepali content via OCR/HWR
    - Know-Your-Customer (KYC) Forms
    - Census Forms
  - Contextual and Intellectual Voice-based automated services in companies
  - Intelligent Question Answering Expert Systems
    - Agriculture
    - Medical

Thank You

Questions ??