Dzongkha NLP

Tenzin Namgyel, GovTech Agency

Introduction to Dzongkha

Dzongkha (རྫོང་ཁ) is the national language of Bhutan

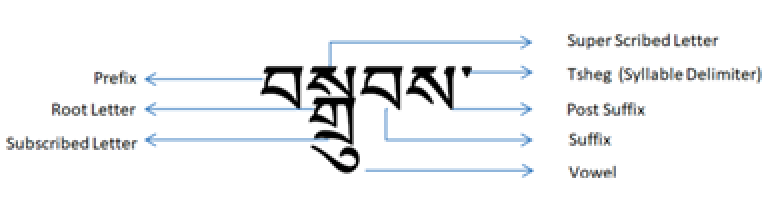
Name literally means "the language of the fortresses" (dzongs)

Dzongkha is a official language used in government, education, and formal settings. Other languages are used as well.

Bhutan is linguistically diverse, with 23 distinct languages spoken. 21 indigenous languages.

Written in Tibetan script introduced by Thonmi Sambhota. The script has 30 consonant and 4 vowel symbols

A Dzongkha syllable can have 1 to 7 characters and most interestingly, up to four characters can stack on one another as shown in the figure below.



Some Unique Features of Dzongkha

- **no word boundary:** “་” is a syllable marker
- **syllabic:** sem: mind; shi: die; semshi: feel sad (not mind die)
- **free word order:** “nga gi apple zayi” and “apple nga yi zayi” both means the same “I ate an apple” (gi is a agentive case marker)
- **infixes:** numbers and modifiers can appear in between the syllables of a word

Dzongkha NLP tools

Spell and Grammar Checker

- working to improve it
- only around 20k instances of training data
- mT5 model was fine tuned (50% accuracy)
- Used script to generate erroneous data
- Using ASR generated text as additional data

Text to Speech

-first version developed as a part of DDF
-new version: mms-tts-dzo fine-tuned with 2665 audio-text pairs
-Mean Cepstral Distortion (MCD)- reduced 4.2 from 7

Language Model

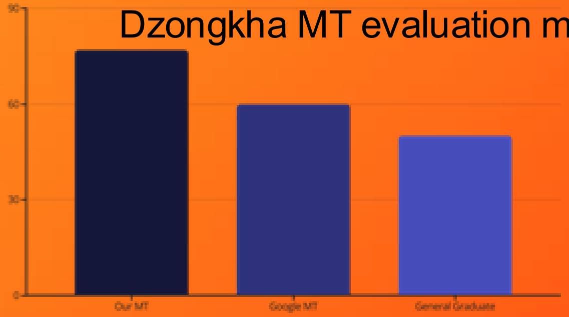
-piloted Dzongkha LLM
-1.2 lakhs training data instances
−Llama 3.1 8b , QWEN2.5 7b, deepseek 7b

Speech Recognition

-First model developed during ASR summer school in 2017 at IITG
-CST and DCDD developed another one as a part of DDF (www.nlp.cst.edu.bt)
-new version: MMS-1B-All Adapter model fine-tuned with 7385 training instances
current word error rate is 0.373

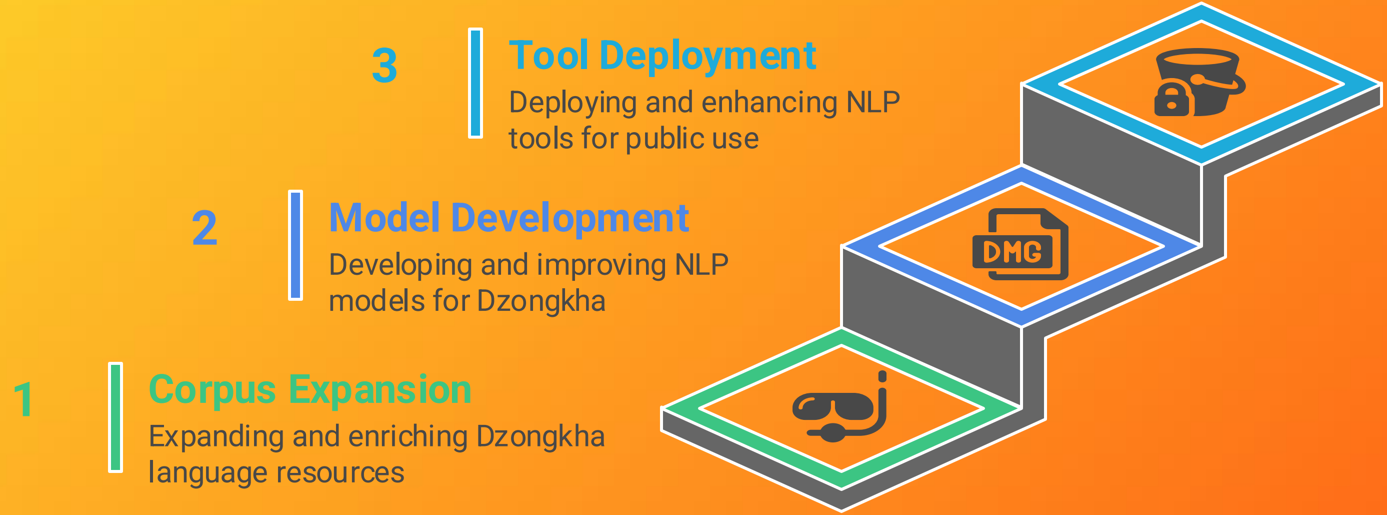
Machine Translation

-first version as part Digital Drukyul Flagship(DDF) www.nlp.cst.edu.bt
- 10 million parallel corpus was developed
- new version done by Govtech Agency currently being tested (NLLBdistilled-600m; Helsinki, Google T5)



Dzongkha MT evaluation my human

Way Forward and Challenges



3 Tool Deployment

Deploying and enhancing NLP tools for public use

2 Model Development

Developing and improving NLP models for Dzongkha

1 Corpus Expansion

Expanding and enriching Dzongkha language resources

Challenges:

- very less corpus and corpus creation is expensive
- need more computational resources but GPU cards are expensive
- new more experts
- limited funding - any support will be appreciated

Try it yourself at:
https://nlp.tech.gov.bt

བཀྲིན་ཆེ།

Thank You!