NLP Course Project

To automate conversion of raw news headlines into grammatically correct form

Objective

To automate conversion of raw news headlines into grammatically correct form

- Ex: Raw headlines: KSERC to evaluate power tariff details
  Grammatically transformed: KSERC is going to evaluate power tariff details

Characteristics of News Headlines

Use Present Simple tense for Past events

Lion escapes zoo

Leave out Aux. Verbs

Four stranded in sudden flood

Leave out Articles

PM hikes Alps for charity

Replace Conjunctions with punctuations

Fire in bakery; hundreds dead

Leave out ‘to say’

Mr Jones: “They’re not taking my house!”

Use infinitives for past events

President to visit France tomorrow

Possible Applications

A model capable of automatic transformation of news headlines can have following usecases -

- Use in e reader applications
- As an accessibility feature in news websites, search engines
- News summary generator

Corpus Preparation

Following changes were made in the corpus provided to make it suitable for model creation -

- Aligned mismatched raw and transformed headlines
- Single instances of raw/transformed data was pruned
- Nouns and their abbreviations were converted to a common form (CM – Chief Minister)
- Punctuations were removed due to their ambiguous use in corpus

Approach **1** - IBM Model 1

→ Model Creation

- Transfer all the files from the folder "*1st_attempt_collabotory*" in a google collab page
- Then run the "*1st_attempt_SMT_headline_to_canonical.ipynb*" file

Approach **1** - IBM Model 1

→ Observation

- Model reached convergence in 35 iterations against suggested 48 iteration due to small dataset (4000 parallel lines)
- BLEU score obtained was 0.47
- Conversion is also not convincing enough on manual inspection. Proceeded with Moses SMT for better results

Approach **2** - Moses SMT

Creating 2 different models using Moses, one as baseline & another with linguistic info

- Model 2 (Moses - PBSMT)

  Correspondences are created between continuous sequences of words in raw and transformed headlines

- Model 3 (Syntax)

  More structure is added to the correspondences by introducing linguistic info like POS tags, sentence structure. MXPOST was used for tagging while Collins parser was used to create parse trees

Approach **2** - Moses SMT

→ Model Creation

- Copy *“moses-model23”* directory present in *“Final_attempt(moses)”* to a UNIX like system
- cd to *“moses-model23”*
- Run the bash script as follows

  $ sudo bash ./run-model23.sh 4 *#where 4 = no of cores available in PC*

Approach **2** - Moses SMT

→ Observation

Based on an example run. Output present in "*Final_attempt(moses)/Output-exampleRun*" directory

Moses – PBSMT

- *BLEU = 0.72*
- *Despite of high BLEU score, the output produced lacks reordering and association like ": ⇔ said"*

Moses – Syntax (Hierarchical)

- *BLEU = 0.70*
- *More acceptable transformation due to presence of POS tagging & parse trees*
- *Lacks in insertion of absent grammatical features like articles, aux verbs*

Approach **2** - Moses SMT

→ Conclusion

Moses – PBSMT

- *Model 2 lacks reordering and association due to insufficient corpus size and lack of any linguistic info*

Moses – Syntax (Hierarchical)

- *Introduction of linguistic info in the form of POS tags and parse trees leads to more acceptable transformation by Model 3. It however still lacks in inserting grammatical features like articles, aux verbs due to insufficient size & poor quality of available corpus*

Future Scope

- Larger parallel corpus, with manually transformed headlines for better results using SMT
- A preceding rule based model before SMT to insert missing grammatical features like ‘to-be’, articles with minimum ambiguity. Heuristic approaches or a probabilistic model may be used to prune multiple transformations in case of conflicting grammatical rules

Thank You

Mentor – Samapika Roy

Anubhav Kumar (17085018), EEE

Arpit Jain (17045027), CHE

Arbabosu Mandal (17095015), ECE