Mid-Semester Report

Index

Dataset

Data Statistics

DATA	COUNT
Wiki Token	39932
Web Token	26508
Wiki Sentence	2310
Web Sentence	2541

POS TAG	
NOUN	28973
VERB	11755
PUNCT	9048
ADJ	2943
ADP	2761
CONJ	2672

POS TAG	
DET	2421
ADV	1591
PRT	1004

Feature Distribution

- Initial Random distribution
- NOUN category dominates and functions as in dump category
- 80:10:10 Feature Distr
- 80:10:10 with Phonological Normalization
- > ***The major problem is the distribution of feature since it is not diverse as the random test set***
  >
  > - *The distribution of train and test sets of distributions of other should be further demonstrated*
  > - *The scores are dropping drastically*
  > - *Should the models used in this dataset*

Feature Normalization

- I have turned semantic features into different POS tags
  - ADP+ComplementType=? -> ADP^ComplementType
  - PRON+PronType=? -> PRON^PronType
- Multiple Polarity around Derivation=Able as
  - ```
    atabiliyor:
    at+VERB+*Polarity=Pos*+Derivation=Able+Polarity=Pos+TenseAspectMood=Prog1+Copula=PresCop+PersonNumber=V3sg
    to
    at+VERB+Derivation=Able+Polarity=Pos+TenseAspectMood=Prog1+Copula=PresCop+PersonNumber=V3sg
    ```
  - To further demonstrate overt morphology
- Should extra feature further eliminated
  - Case=Bare before the derivation the derivation
- > *Main reason why I do this is to capture **overt morphology** which more features can create **a noise to the model***
    ```

Phonological Normalization

- Mostly done however cannot decide on the lemma alternation
    - kitap -> kitab-ı??
- Should be further discussed

Model

- Model is a Transformers architecture with the default hyper-parameters

PARAMETER	VALUE
Vocab Size	15000
Sequence Length	20
Embedding Dimensions	256
Latent Dimensions	2048
Number of (Attention) Heads	8
Batch Size	64
Epoch	5

- Sequence Length at max(wiki_token_length,web_token_length)
- I use Epoch at 5 since I saw that Epoch > 10 (Overfitting)

Evaluation

Metrics

- Partial Match: Calculates in terms of subset of common features
- Exact Match: Literal string Match of sequences
    - 0.1 ~ 0.2
- BLEU Score: Feature-wise edit distance between y_pred and y_true
- Further Analysis
    - edit_distance(lemma_pred vs lemma_true)
    - precision, recall of POS tagging

	PAR-PRE	PAR-REC	EX-PRE	EX-REC	BLEU	POS
Baseline (Random - Token)	0.763	0.774				

	PAR-PRE	PAR-REC	EX-PRE	EX-REC	BLEU	POS
Feature Distribution	0.35	0.35				
Random Dist with Normalization	**0.825**	**0.805**				
Feat Dist with Normalization	0.372	0.374				
Context Embedding (TO-DO)						