Evaluating the Quality of a Corpus Annotation Scheme Using Pretrained Language Models

LREC-COLING 2024 Presentation

Furkan Akkurt[1], Onur Güngör[1], Büşra Marşan[2], Tunga Güngör[1], Balkız Öztürk Başaran[1], Arzucan Özgür[1] and Susan Üsküdarlı[1]

1. Boğaziçi University, Turkey; 2. Stanford University, USA

Introduction

- Pretrained language models increasingly used in NLP
  - New tasks these LMs able to solve in a zero-, one- or few-shot fashion
  - Language resource evaluation as a task
- UD project provides manually created treebanks
  - Annotation differences due to different linguistic theories or simple errors
- Using LLMs to compare 2 consecutive versions of a treebank
  - New method for comparing token-level annotations
  - Code released on GitHub

UD Project and the Turkish BOUN treebank

- UD project has treebanks
  - Treebanks comprise of sentences and their token-level annotations
  - Annotations include word categories, morphological features and dependency relations
- BOUN treebank, introduced in v2.7 with 9 761 sentences
  - 2 distinct versions, v2.8 and v2.11, of concern
  - v2.8 had limitations in expressivity due to differences between UD framework and Turkish features
  - v2.11 addressing representation challenges of Turkish, and
    - solving errors based on lemmatization and morphological features

Annotation example

and comparison

ID	Form	Lemma	POS	Features	Head	Deprel
1	Ali'den	Ali	PROPN	Case=Abl | Number=Sing | Person=3	4	obl
2	oyuna	oyun	NOUN	Case=Dat | Number=Sing | Person=3	3	obl
3	katılmasını	kat	VERB	_	4	ccomp
4	istediler	iste	VERB	Aspect=Perf | Evident=Fh | Number=Plur | Person=3 | Polarity=Pos | Tense=Past	0	root
5	.	.	PUNCT	_	4	punct

v2.8 annotation of the sentence “Ali’den oyuna katılmasını istediler.” (*They wanted Ali to join the game.*)

ID	Form	Lemma	POS	Features	Head	Deprel
1	Ali'den	Ali	PROPN	Case=Abl | Number=Sing | Person=3	4	obl
2	oyuna	oyun	NOUN	Case=Dat | Number=Sing | Person=3	3	obl
3	katılmasını	kat	VERB	Case=Acc | Number=Sing | Number[psor]=Sing | Person=3 | Person[psor]=3 | Polarity=Pos | VerbForm=Vnoun | Voice=Pass	4	ccomp
4	istediler	iste	VERB	Aspect=Perf | Evident=Fh | Number=Plur | Person=3 | Polarity=Pos | Tense=Past	0	root
5	.	.	PUNCT	_	4	punct

v2.11 annotation of the same sentence

LLMs for evaluation

- Adapted to solve highly varied tasks, including evaluation
  - Achieved high correlation with human judgments
- Evaluation of UD treebanks via LLMs
  - Relatively unexplored
  - Assessing annotation quality and accuracy

Method

- Task: comparing annotation schemes of 2 versions of the Turkish BOUN treebank
- LLM input: annotations for a single sentence without surface form, requesting the sentence's original text, including a one-shot example of the task
  - Specifically lemmas, parts of speech, morphological features and dependency relations are provided in natural language.
- LLM output: generated text for the sentence based only on the annotations
  - A sentence is generated by the LLM to be compared with the original sentence.
- Comparison is done on both the character- and token-level.

Prompt example with output

One-shot example left out for space

- Small example of a prompt for the sentence “Ben okula gidiyorum.” (*I’m going to school.*):
  - The following sentences detail linguistic features of a Turkish sentence with lemmas, parts of speech and morphological features given for each token. The sentence has 4 tokens.
  - 1st token's lemma is “ben”, its part of speech is pronoun, its person is first person, its number is singular number, and its case is nominative.
    2nd token’s lemma is “okul”, its part of speech is noun, its number is singular number, its case is dative, and its person is third person.
    3rd token’s lemma is “git”, its part of speech is verb, its voice is active voice, its polarity is positive, its tense is present tense, its aspect is progressive aspect, its mood is indicative mood, its person is first person, its number is singular number.
    4th token’s lemma is “.”, and its part of speech is punctuation.
  - Task is to find the surface form of the sentence.
- and output: “Ben okula gidiyorum.”

Experiments

and results

- Several LLMs used with Poe API
- v2.11 (reannotated version) has a consistent increase of 1.5% character-level accuracy across the board
- GPT-4 produces highly accurate generations, while smaller open-source models, like Llama 2, lack accuracy
  - as open-source LLMs are not trained on Turkish data, and
  - understanding Turkish linguistic features is rare in models.

Results

with GPT-4

- Character-level
  - v2.8: 90.0%
  - v2.11: 91.3%
- Token-level
  - v2.8: 73.8%
  - v2.11: 76.9%

Conclusions

- LLMs used to compare UD treebank quality
  - In experiments, Turkish BOUN treebank v2.11 shows better linguistic representation than v2.8
- Method reveals insights into annotation schemes and contributes to higher quality language resources

References

- UD: universaldependencies.org
- Code: github.com/boun-tabi/eval-ud
- Poe: poe.com