امتحان پایان ترم پردازش زبان طبیعی

زمان امتحان: 2 ساعت

1- **دومعیار اصلی برای یک ترجمه مناسب از جمله مبدا را نام ببرید؟**

a. **جمله انگلیسی** "we can speak Persian" **را به فارسی ترجمه کرده ایم، و جملات زیر کاندیدهای ترجمه هستند. در جدول زیر مشخص کنید که هر یک از معیارهای زیر مناسب هستند یا خیر؟**

	معیار 1 (.....)	معیار2 (.....)
صحبت کنیم ما فارسی میتوانیم		
ما حاضریم مذاکره کنیم		
ما میتوانیم فارسی صحبت کنیم		
حاضر ما عربی رفتیم		

پاسخ:

Fidelity/fluency

	معیار 1 (fluency)	معیار2 (fidelity)
صحبت کنیم ما فارسی میتوانیم	**خیر**	**بله**
ما حاضریم مذاکره کنیم	**بله**	**خیر**
ما میتوانیم فارسی صحبت کنیم	**بله**	**بله**
حاضر ما عربی رفتیم	**خیر**	**خیر**

2- **در یک ترجمه ماشینی آماری از ترازبندیهای زیر در هر مرحله استفاده میشود. لطفا ترتیب مراحل ان را بگویید. یعنی ابتدا کدام تراز بندی را داریم و سپس با چه ابزاری به ترازبندی دیگر میرویم. نام ابزار یا روشی که این تغییر تراز بندی در ان صورت میگیرد را بگویید:**

a. Word alignment
b. Phrase alignment
c. Sentence alignment

پاسخ:

ترتیب ابتدا c **با ابزاری مشابه** Giza++ **(روش** IBM model**) به** c **تبدیل میشود . سپس با روشی مانند** Symmetrizing **به** c **تبدیل میشود**

3- همانطور که میدانید حتی در صورتیکه یک مترجم ماشینی (عصبی یا آماری) را با یک دادگان موازی بزرگ اموزش دهیم، باز هم کلماتی در زمان تست هستند که در vocabulary نیستند (مانند اسامی خاص یا اشتقاقهای افعال یا ...) در رفع این مشکل حداقل 2 راهکار (به غیر از افزایش مجدد دادگان موازی ☺) پیشنهاد دهید؟

پاسخ:

روش اول: استفاده از تگ UNK برای همه کلمات rare در دادگان موازی در نتیجه کلمات دیده نشده، در ترجمه به UNK تبدیل میگردد. در زمان تست کافی است که از alignment بین جملات ورودی و خروجی (که attention ایجاد میکند) استفاده کنیم و ترجمه کلمه UNK که در خروجی داریم را با یک دیکشنری ایجاد کنیم. معمولن اسامی خاص اینگونه ترجمه میشوند

روش دوم: استفاده از BPE که برای اشتقاقها خیلی مناسب هستند.

4- برخی جفت زبانها زبانها مانند "انگلیسی-دانمارکی" یا "انگلیسی-سوئدی" یا "فارسی-اردو"بسیار به هم نزدیک هستند بطوریکه word order کلمات تقریبا یکی است و صرفا تفاوتها در spell کلمات و واژه نامه کلمات هست. اصطلاحا به مترجم ماشینی که ترتیب کلمات را عوض نمیکند، Monotone (یکنوا) مینامند. برای ایجاد یک ترجمه ماشینی عصبی یکنوا چه راهکاری پیشنهاد میدهید؟

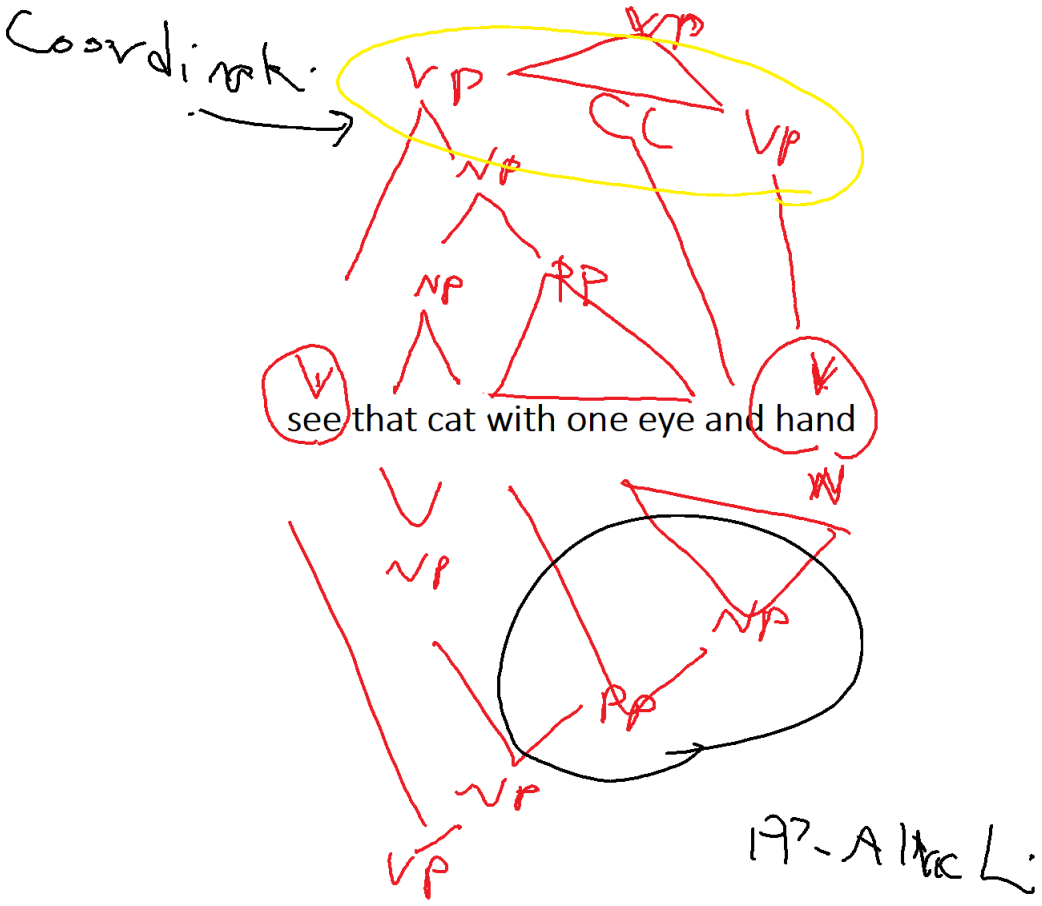
پاسخ: بجای مدل encoder-decoder از مدلهای seq2seq استفاده کنید. مساله اصلی دراین مدلها tokenization است. یعنی چه واحدهایی در نظر بگیریم که وارد مدل seq2seq شود. میتوان واحد را Phase/word/BPE یا حتی character در نظر گرفت. که احتمالا به جفت زبان بستگی دارد. ولی احتمالا بهترین انها BPE خواهد بود

5- دو نوع ابهام نحوی که در تجزیه گرهای نحوی ممکن است اتفاق بیافتد چیست؟ درخت های تجزیه نحوی برای جمله انگلیسی زیر را بکشید و نوع ابهام در هر درخت را نام ببرید؟

see that cat with one eye and hand

پاسخ:

PP-attachment/ Coordination



6- **در مورد تجزیه وابستگی به سوالات زیر پاسخ دهید:**

a. **تجزیه گر وابیستگی افکنشی** (projective) **چیست؟**

b. **بنظر شما ارائه یک تجزیه گر افکنشی ساده تر است یا غیر افکنشی؟ چرا؟**

c. **دو جمله در زبان فارسی مثال بزنید که یکی از آنها تجزیه وابستگی افکنشی** (projective) **باشد و دیگری نباشد؟**

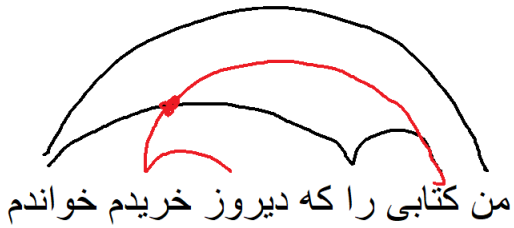
پاسخ:

الف: برخی تجزیه گرهای وابیستگی طوری است که هیچ گره ی فرزندی وابستگی به گره ای خارج از بازه گره پدر (یا اجداد خود) ندارد. در واقع اگر یک گره x در نظر بگیریم صرفا میتواند وابستگی به گره های y داشته باشد که x,y فرزندان مشترک تمام پدران و اجداد x در درخت وابیستگی باشد. اگر کل گره ها در یک خط رسم کنیم و تمام یالهای وابستگی در یک سمت خط باشد، نباید دو یال همدیگر را قطع کنند.

ب: قطعا افکنشی ساده تراست چون اگر محدود فرزندان یک گره پدر مشخص شود، عملا حق انتخاب وابیستگی فرزندان ان گره صرفا گره های فرزند ان گره هست.

ج: مثال افکنشی: من کتابی که دیروز خریدم را خواندم.

غیرافکنشی: من کتابی را که دیروز خریدم خواندم.



7- **ما میخواهیم** query **های ورودی به زبان طبیعی را به یک زبان ساختاریافته** (structured) **برای** query **ها تبدیل کنیم. لطفا توضیح بدهید که چه تسکی در** NLP **ما میخواهیم انجام دهیم. چگونه شما میتوانید برنامه‌ای بنویسید که این تسک رو اتوماتیک انجام دهد؟**

Answer:

- It's semantic parsing.
- If we have enough training data in the form of (natural-language-query, query), then we can train a seq2seq model. If not, then we can also write some rules to convert natural-language queries into the structured language.

8- **در یک** task-oriented dialogue**، ما گاهی نیاز به** query **زدن به یک پایگاه داده پیدا میکنیم. چه تسک یا تسکهایی در** NLP **را احتمالا استفاده میکنیم؟**

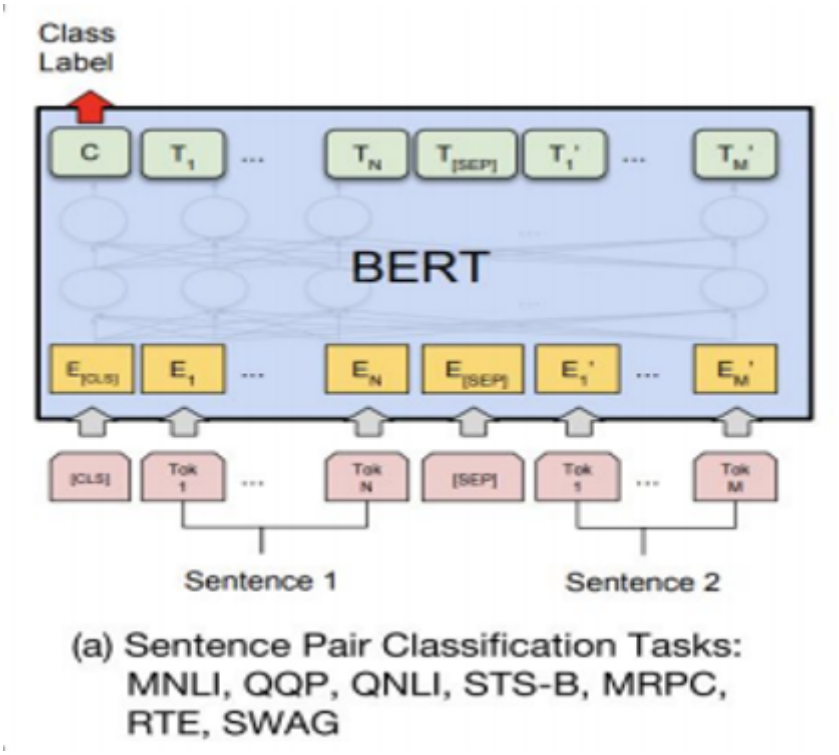
Answer:

- If we use a frame-based architecture for the dialogue system, then "slot filling" is one task. We might also use "semantic parsing" to parse the NL query into SQL or any other query language directly. Another task is “coreference resolution” which might be used here.

9- **اگر ما** paraphrase detection **انجام دهیم (پیش بینی اینکه دو جمله** paraphrase **هستند یا خیر) با استفاده از یک مدل ترنسفورمر، چه انواعی از** attention **استفاده میکنیم؟ یک معماری خوب برای انجام این تسک را با استفاده از ترنسفرمها رسم کنید.**

Answer:

- cross-attention, self-attention



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

10- **چرا در ترنسفورمرها از** self-attention **استفاده می‌کنیم؟ فایده استفاده از** multi-head attention **چه هست؟**

Answer:

To model the context and the dependencies between words in the context.

Each head could represent a specific aspect about the relation between two words, e.g., that a word is the obj of another word. Then, using multi-head, we can model multiple aspects of their relations.

11- **معنای یک کلمه در یک جمله به بافتار** (context) **چپ و راست آن و همچنین معنای عمومی آن بستگی دارد. در کدام مدلها این گزاره صحیح می‌باشد؟**

a) word2vec b) ELMo c) BERT d) Glove

Answer:

b and c

12- **دلیل استفاده از** masked language modeling **برای** train **کردن چیست؟**

Answer:

- To model bidirectional context when doing language modeling. If we don't use masking, then either we have to use unidirectional models or we can cheat by having predicting the seen input.

13- **من یک دیتاست دارم با تعداد زیادی نمونه با این برچسبها** ("non-", entailment") **. چه تسکی قرار است که من انجام دهم؟ چه پیشنهادی به من می دهید برای انجام این تسک؟**

Answer:

- It is called natural language inference or textual entailment.
- If we have enough data, we can fine-tune BERT or other Transformers on the examples we have and perform sentence-pair classification.

14- **اگر من بخواهم که یک مدل خوب برای** sentiment analysis **برای کار در تلفن همراه با حافظه‌ی محدود بسازم، آیا میتوانم از** BERT-large **برای ساخت مدل نهایی خودم استفاده کنم؟ چگونه؟**

Answer:

- Yes, we can do that but we need to do knowledge distillation or other model compression techniques to have a final model with significantly less number of parameters.

15- **برای پاسخ به سوالاتی شبیه به این سوال:**

How many provinces are there in Iran?

چه منابعی را یک سیستم پرسش و پاسخ (QA) ممکن است استفاده کند؟

Answer:

- It can use corpora like Wikipedia and news.
- It could use knowledge bases like Wikidata or Dbpedia.

16- **چگونه می‌توان از** XLNet **یا** RoBERTa **در یک سیستم** IR-based question answering **استفاده کرد؟**

Answer:

- They can encode the question, the documents, paragraphs into vectors.
- Also for finding the exact span in a paragraph, XLNet or RoBERTa can be used similar to how BERT is used. See the class slides on SQUAD.

17- در استخراج اطلاعات (Information extraction)، ما اطلاعات ــــــــــــ را از متن بیرون می‌کشیم. (پاسخ میتواند انگلیسی یا فارسی باشد.)

Answer:

- Factual

18- تفاوت اصلی میان دو تسک named entity recognition و entity linking در چه هست؟ می‌توانید مثالی از هریک بزنید؟

Answer:

NER classifies named entities in text, but entity linking finds their references in a knowledge base. You can find examples in our lecture slides.

19- حداقل دو الگو یا pattern برای پیدا کردن روابط captial_of از یک پیکره به صورت اتوماتیک بنویسید.

Answer:

- capital of [COUNTRY] is [CITY].
- [CITY], the capital of [COUNTRY],

20- فرض کنید که این mentionها را برای عنصر اسمی 100E داریم و یک مدل هم داریم که یک context را گرفته و آن را به یکی از N تا entity type دسته بندی میکند. چگونه ازین مدل میتوانیم برای پیدا کردن تمامی type های 100E استفاده کنیم.

E100 gave her speech at the UN.
Congratulations to **E100** as she picks up her new honorary doctorate.
In her book, **E100** talked about different issues
He had dinner with **E100** yesterday.

Answer: Yes, we can classify each context of E100 with that model and then aggregate the predicted classes to find all types of E100.

Dialogue

21- **در یک دیالوگ ما داریم:**

A: I want to order two Kabab Koobideh.

B: Ok, so two Koubideh.

در اینجا B دارد چه چیزی را میسازد؟

Answer:

Common ground

22- **درست یا غلط عبارات زیر را مشخص کنید:**

a. سیستم الیزا از یادگیری ماشین برای استخراج rule استفاده میکرده است.

b. میتوانیم از gpt برای تولید پاسخ در سیستمهای دیالوگ استفاده کنیم.

Answer:

a- False, b- True

23- **برای دامنه «سفارش غذا از یک رستوران» چه** slot **ها و** type **هایی را میتوان متصور بود؟ برای سادهسازی فرض کنید که** dialogue agent **ما فرض میکند که یک نوع غذا رستوران دارد.**

- food-name or food, type: food
- Number, type: Number
- Address, type: address
- Phone number, type: phone number
- Time, type: time