CMPE 493 - Term Project Part 2
Question Answering

Abdullatif Köksal (abdullatifkoksal@gmail.com)
Rıza Özçelik (riza.ozcelik@boun.edu.tr)
Gökçe Uludoğan (gokce.uludogan@boun.edu.tr)

DEADLINE: May 25, Saturday, 23:59

Acknowledgement

We hugely thank Enes Çakır for coding the tool and enabling a smoother dataset preparation process.

1 Introduction

In this part of the project, you are expected to implement a model that can accomplish two tasks: finding the related paragraph and the answer, given a question. Note that a paragraph is related to a question if and only if the paragraph contains the answer of the question and there exists only one related paragraph for each question.

2 Dataset

For the dataset, you will use the dataset that we have created together in the first part. Whole corpus and 80% of the QA groups are uploaded to Moodle for you to use during training. Note that we have spared 20% of the QA groups for evaluation purposes and will reveal it when the deadline is passed.

The uploaded dataset contains two files: **derlem.txt** and **soru_gruplari.txt**. The former contains the paragraphs and their IDs whereas the latter contains questions, answers and the related paragraph IDs. A sample of both files can be seen in Figure 1 and 2.

3 Evaluation

For the first task, related *paragraph prediction*, we will use accuracy to measure your model's success on the test set. For the second task, question answering, we

```
6001 İnsanın içinde yaşadığı, canlı ve cansız tüm varlıkları içerisinde barındıran yer; ortam veya çevre olarak
adlandırılır. Doğal şartlar altında gelişen olayların oluşturduğu ortama doğal ortam adı verilir. Doğal ortam içerisinde
kayaç ve tabakalardan oluşan, üzerinde dağ, plato, ova, vadi gibi yer şekillerinin yer aldığı ortama taş küre (litosfer);
okyanus, deniz, göl ve akarsu gibi suların oluşturduğu ortama su küre (hidrosfer); yerküreyi çepeçevre saran gazlardan
oluşan ve hava olaylarının meydana geldiği ortama ise hava küre (atmosfer) adı verilir. Doğal ortam sadece cansız
varlıklardan ibaret değildir, burada bitkiler ve hayvanlar da vardır. Bunlar da birlikte canlı küreyi (biyosfer)
oluşturur. Böylece bu dört ortam yeryüzündeki doğal sistemleri meydana getirir.

6002 Coğrafyanın odak noktasında insan vardır. Yukarıda sayılan ortamlarla birlikte insanın yeryüzünde gerçekleştirdiği
bütün faaliyetleri oluşturan ortam ise "beşerî ortam"dır. İnsan eliyle yapılmış bina, yol, köprü gibi unsurlar ile tarım,
sanayi gibi insan faaliyetleri beşeri ortamı oluşturur.

6003 Beşerî sistemler, insanın varoluşundan günümüze kadar uzanan bir süreç içerisinde teknolojinin gelişimine ve nüfusun
artışına bağlı olarak giderek gelişmekte ve daha da karmaşık bir hâle gelmektedir. Örneğin toplayıcılık ve avcılıkla
uğraşan ilk insanlar kaynakların sınırlı olması nedeniyle sürekli göç etmek zorunda kaldı ve küçük gruplar hâlinde
yaşadı. Fakat günümüzden 10 bin yıl kadar önce gerçekleşen Tarım Devrimi ile insanoğlu temel ihtiyacı olan gıda sorununu
çözmeye çalıştı. Böylece hayatta kalmak için gerekli olan temel kaynaklara erişim kolaylaştı. Bu durum yerleşik hayata
geçişi de beraberinde getirdi. Yerleşik hayata geçişle birlikte giderek artan dünya nüfusuna ve büyüyen yerleşmelere
bağlı olarak ekonomik faaliyetler de çeşitlenerek arttı.

6004 Birbiriyle ilişkili olarak işleyen dört temel ortamdan (atmosfer, hidrosfer, litosfer ve biyosfer) oluşan doğal
ortam ve  beşerî ortam birlikte coğrafi ortamı oluşturur. Coğrafi ortam sadece dünyanın yüzeyi ile sınırlı değildir.
İnsanları ve diğer bütün canlı ve cansız varlıkları içine alan; eni, boyu ve derinliği bulunan; üç boyutlu ve coğrafyanın
konusu olan olayların meydana geldiği yerdir.

6005 Coğrafyanın inceleme alanını coğrafi ortam oluştururken coğrafyanın inceleme konusunu insan ve doğal ortam
arasındaki karşılıklı etkileşim oluşturur. Coğrafya, doğal ortamı insan ile birlikte ele alıp değerlendirmesi yönüyle
doğa, fen bilimleri ve sosyal bilimlerden ayrılır. Çünkü coğrafyanın inceleme konusu, coğrafi ortam içerisinde
gerçekleşen insan-doğa etkileşimi, bir başka ifadeyle coğrafi olaylardır.
```

Figure 1: A Corpus File Sample

```
S1715: Haritalarda kullanılan işaret ve renklerin ne anlama geldiğini gösteren bölüme ne denir
S1720: Hangi bölüm haritalarda kullanılan işaretlerin anlamını gösterir
C1293: Lejant
İlintili Paragraf: 6293

S1708: Haritalar kendi içerisinde hangi gruplara ayrılır
C1286: Genel ve tematik
İlintili Paragraf: 6297

S1712: Siyasal haritalar hangi türe örnektir
C1290: Genel
İlintili Paragraf: 6298

S1709: El İdrisi hangi alan üzerinde çalışmıştır
C1287: Haritacılık
İlintili Paragraf: 6305

S1643: Kaşgarlı Mahmud hangi eserinin başına Türk dünyası haritasını eklemiştir
S1644: Başında Türk dünyası haritası olan Kaşgarlı Mahmut eseri nedir
C1242: Divan ı Lügat it Türk
İlintili Paragraf: 6306

S1645: Kitab-ı Bahriyenin yazarı kimdir
C1243: Piri Reis
İlintili Paragraf: 6308

S1710: Doğal ve beşeri sistemleri toplayıp analiz yapılması için kurulan bilgisayar sistemi nedir
S1717: Hangi bilgisayar sistemi doğal ve beşeri sistemleri toplayıp analiz etmek için kurulmuştur
C1288: Coğrafi Bilgi Sistemleri
İlintili Paragraf: 6312
```

Figure 2: A QA Groups File Sample

will use jaccard similarity, since exact match is not always possible. To compute the jaccard similarity, we will use character bigram sets of the prediction and gold truth and average the score of each prediction over all questions. We will normalize both the predictions and the answers by punctuation removal and case folding.

Let us say the first prediction is *Karadeniz'de* and the first gold truth is *Karadeniz*. We will construct two sets $S1 = \{ka, ar, ra, ad, de, en, ni, iz, zd, de\}$, $S2 = \{ka, ar, ra, ad, de, en, ni, iz\}$ and compute the similarity as:

$$J_1 = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{8}{10} = 0.8$$

Assuming that N questions exist in the test set, to calculate the average jaccard similarity over all questions we will compute:

$$J = \frac{1}{N} \sum_{i=1}^{N} J_i$$

4 Submission Details

You are expected to submit a zip file that is named with your group ID in this sheet [1] (<*GID*>.zip). In this zip file, there must a exist a python3 script named with your group id as well (<*GID*>.py). We will run this script with the following command for a group with ID=101:

python3 ./101/101.py test_questions.txt task1_predictions/ task2_predictions/

When this command is executed, your script must output two files that contains your predictions for the test questions for both tasks, separately. Both of these files must be named with the same name: <*GID*>.txt. If your script needs any extra file to run (pickle, stopwords.txt etc.), you can include them in the zip as well.

In this command, first parameter is the path to test questions file and the other two are paths to folders that your predictions must be written to.

Lastly, format of the test questions file will be different from the training questions and there will be one question per line with no question id information. In the following days, we will release the evaluation script we will use for grading alongside a sample test file, so that you can test your submission yourselves in your own machine. We will also announce the library versions we will use during grading to avoid library incompatibilities.

5 Grading

- ~~Dataset Preperation - *35 pts*~~

[1] If your ID is not in this file, reach us as soon as possible.

- Related Paragraph Prediction - *30 pts*
- Question Answering - *20 pts*
- Presentation - *15 pts*
- ~~Bonus - *10 pts*~~.