UQ at the TREC 2019 News Track

Vu Anh Le and Gianluca Demartini

University of Queensland, Australia

1 Introduction

The Data Science research group at the University of Queensland participated to the TREC 2019 News Track by submitting 5 runs for the Entity Ranking task. Our approach is based on entity frequency, sentence similarity scoring functions, and the use of external sources of evidence like Wikipedia.

The results we obtained show that 1) the most effective of our methods makes use of Wikipedia as an external collection to rank entities and that 2) this method can deal well with difficult topics but it should be combined with alternative approaches on a topic-by-topic basis.

2 Methods

In this section we describe the approach used to produce five runs for the TREC 2019 News Track Entity Ranking task.

1. **UQ_count**: The entities appearing in the document are ranked based on the number of time the entity appears in the document (i.e., entity frequency).

2. **UQ_sent**: The entities appearing in the document are ranked based on the mean similarity score computed comparing the sentences containing the entity with the whole document.
3. **UQ_wiki**: The entities appearing in the document are ranked based on the similarity score obtained comparing the Wikipedia representation of the entity and the whole document.

4. **UQ_count_sent**: The entities appearing in the document are ranked based on the entity frequency and the mean similarity score computed comparing the sentences containing the entity with the whole document.

5. **UQ_wiki_count**: The entities appearing in the document are ranked based on the similarity score between the Wikipedia representation of the entity and the whole document combined together with the sentences similarity score and entity frequency.

2.1 Statistics-based Method: UQ_count

The first statistical method we present is *Entity Frequency*, which is based on counting entity occurrences in the news article. After being extracted from the document, entities are filtered to include only entities of type *PERSON, ORG, GPE, NORP* and *FAC*. Then, entities are standardised by removing some stopwords, i.e. *a, an* and *the*, using the spaCy NER tool.

After calculating the frequency of every entity, the list of entities is sorted in descending order of frequency (i.e., most frequent entity ranked first). Finally, the frequency is used as the ranking score for the entities.

2.2 Statistics-based Method: UQ_sent

The second method we present is based on the *Average Sentence Score*, which ranks entities based on the average score of the containing sentences. Each sentence which contains the entity in the articles is scored based on the similarity between the text of the sentence and the whole news article. Then, the score of an entity is calculated using the following equation.

$$SentScore(E_i) = \frac{1}{N}\sum_{j=1}^{N} sim(S_j, Doc) \tag{1}$$

Where E_i is the i-th entity in the document, N is the total number of sentences that contain the entity E_i and Doc is the whole document content. The similarity between the sentence E_i and the article Doc is calculated based on the semantic similarity function of the spaCy NLP library.

2.3 Semantic-based Method: UQ_wiki

For this approach, we use Wikipedia as a source of entity representation [1]. First, we built an entity representation collection based on 5.8 million Wikipedia documents (the English Wikipedia Dump[1]) and then imported it to an Elasticsearch instance to perform query operations. The Wikipedia dump also provides SQL files containing the metadata about all the redirect and disambiguation pages.

Our entity ranking system takes the *Article ID* as the input, performs *Entity Extraction, Entity Disambiguation* and *Summarisation* and returns a list of entities ranked based on the similarity between their Wikipedia representation and the article content.

$$ReprScore(E_i) = \frac{1}{N}\sum_{j=1}^{N} sim(R_j, Doc) \tag{2}$$

Using this architecture, we evaluated two approaches to measure similarity. The first one is comparing the original Wikipedia page and the news article

[1] https://dumps.wikimedia.org/

content while the second approach compares their summaries obtained using the *Aggregation similarity* method presented in [2]. This summarisation method selects the top n sentences with the highest similarity score in the document and combines them into a single paragraph.

2.4 Combined Methods: (UQ_count_sent and UQ_wiki_count)

A limitation of the Wikipedia entity representation similarity method described above is that some of the extracted entities may not have a representation as there is no appropriate Wikipedia page related to the entity. Therefore, we tested four methods that combine together the representation similarity methods and the other two statistics-based methods. Within such combined approach, the overall ranking score is calculated as the product of the component scores as described in the Equation 3.

$$Score(E_i) = max(S_i, R_i)(ln(F_i) + 1) \tag{3}$$

Where S_i, R_i are the SentScore and ReprScore of the i-th entity respectively and F_i is the frequency of that entity.

The entity frequency score is calculated using a logarithmic function to smooth the contribution of the entity occurrence count to the final ranking score. The max function of the SentScore and ReprScore is used to address the problem of entities missing their Wikipedia page. In particular, if there is no page returned from the query to the Wikipedia collection, the SentScore will be used instead to calculate the overall ranking score.

3 Results

The effectiveness of our methods is presented in Table 1. We can see how the most effective approach is *UQ_wiki* which makes use of external evidence. The combined methods do not perform better.

Table 1. Effectiveness of the tested methods. Bold indicates best performing run.

Run ID	NDCG@5	MAP
UQ_count	0.5255	0.5746
UQ_sent	0.3388	0.4805
UQ_wiki	**0.5713**	**0.6288**
UQ_count_sent	0.5222	0.5762
UQ_wiki_count	0.5357	0.5863

Figure 1 shows a per-topic analysis of the effectiveness of our UQ_wiki run. We can observe that it performs better on difficult topics (defined as topics with low best performance).

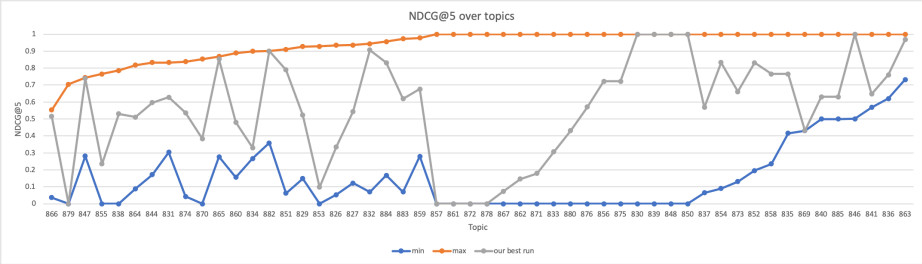

Fig. 1. Per-topic analysis of the effectiveness measured by NDCG@5 our run using the Wikipedia collection as compared to the best and worst performing runs submitted at TREC News 2019.

References

1. Demartini, G., Firan, C.S., Iofciu, T., Nejdl, W.: Semantically enhanced entity ranking. In: International Conference on Web Information Systems Engineering. pp. 176–188. Springer (2008)
2. Ko, Y., Seo, J.: An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Pattern Recognition Letters **29**(9), 1366–1371 (2008)