An Annotation Similarity Model in Passage Ranking for Historical Fact Validation

Jun Araki and Jamie Callan

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Introduction

Background

- Passage retrieval is a core component for question answering (QA) (Tellex et al., 2003; Cui et al., 2005; Ferrucci et al., 2010; Krikon et al., 2012).

Problems

- Many passage retrieval approaches used in QA cannot check linguistic and semantic types annotated in passages at query time (Bilotti et al., 2010; Bilotti et al., 2007).
- NTCIR-11 QA Lab: real-world university entrance exam questions
- We focus on multiple-choice true-false questions on world history.
- The correct answer does not appear anywhere in the corpus.
- Example (the correct answer is 1):

..., most of those who excelled in culture and the arts were those who had passed the Imperial examinations, but in the (2) Ming period, there was a shift toward ...	Introductory text
Question 2. From 1-4 below, choose the most appropriate sentence concerning events that occurred during the period referred to in the underlined portion (2).	Question
1. Japanese silver circulated in China. 2. A Buddhist sect called Zen was created. 3. The play "The Story of the Western Wing (Xixiangji)" was created. 4. The capital was established in Lin'an (present-day Hangzhou).	Answer candidates

- The questions are strongly or weakly dependent on their corresponding introductory text.

Passage Ranking for Historical Fact Validation

Definitions

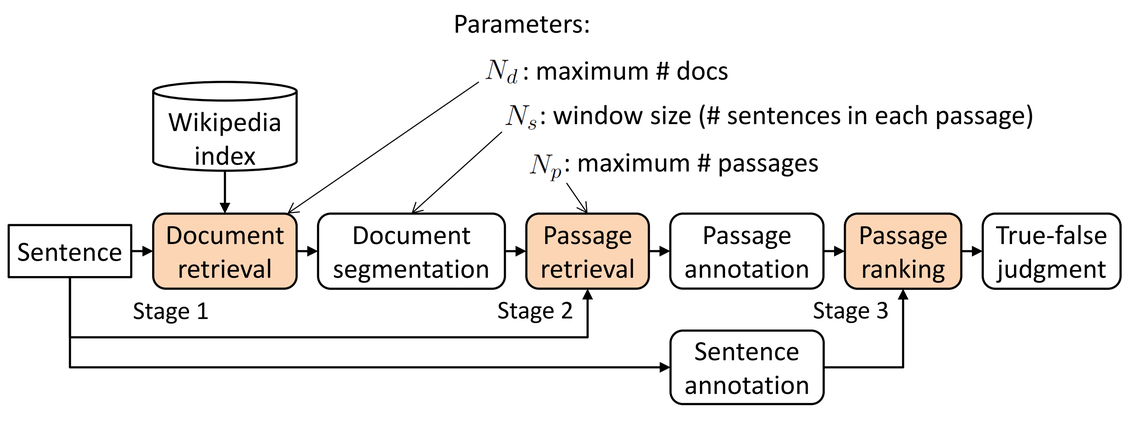
- A **historical fact** is a sentence that tells us historically correct information.
- Example: Japanese silver circulated in China during the Ming period.
- **Historical fact validation** is a subtask to determine whether or not a given sentence is a historical fact.

Assumptions

- We ensure the historical correctness by a reference to information sources that we rely on.
- Wikipedia is abundant of historical facts, and highly likely to cover historical topics of questions in the exam corpus.

Architecture: three-stage passage ranking

- Idea: If a system takes a given sentence as a query (**historical hypothesis**), and retrieves and ranks a passage (**historical evidence**) with a reasonably high score, then the system regards the sentence as a historical fact.



- We use English Wikipedia (2014-02-03 dump) as source data.
- We use the TF-IDF similarity in stage 1 and 2.

Annotation Similarity Model

- Idea: **Making use of annotations of arbitrary types** to boost up answer-bearing passages at query time in an unsupervised manner.
- Motivating example: Cook explored Oceania during the 18th century.

  A passage from http://en.wikipedia.org/wiki/History_of_Oceania:

  > James Cook explored the Pacific islands and the east coast of Australia in the 18th century.

- We intend that an annotation similarity score gives a small amount of similarity adjustment to the bag-of-words similarity score.
- Given sentence *s* and passage *p*, the final score is:

$$score(s,p) = sim_{BOW}(s,p) \times sim_{ANN}(s,p)$$
$$= TF\text{-}IDF(s,p) \times (1 + \alpha\ sim(G_s, G_p, T_c))$$

- For the annotation similarity score, we incorporate the **annotation graph** model in (Bilotti et al., 2010) and **vertex/edge overlap** (Papadimitriou et al., 2010).

Algorithm 1 Annotation similarity model.

Input: $G_1 = (E_1 = \{(te_1)\}, R_1 = \{(tr_1)\}, T)$
Input: $G_2 = (E_2 = \{(te_2)\}, R_2 = \{(tr_2)\}, T)$
Input: $T_c \in T$
Output: annotation similarity score

1: $E'_1 \leftarrow \{(te_1)\}$ where $te_1 \in T_c$
2: $E'_2 \leftarrow \{(te_2)\}$ where $te_2 \in T_c$
3: $R'_1 \leftarrow \{(tr_1)\}$ where $tr_1 \in T_c$
4: $R'_2 \leftarrow \{(tr_2)\}$ where $tr_2 \in T_c$
5: **return** $2\frac{|E'_1 \bigcap E'_2| + |R'_1 \bigcap R'_2|}{|E'_1| + |E'_2| + |R'_1| + |R'_2|}$

Notations
- *G*: an annotation graph
- *E*: a set of elemental annotations
- *R*: a set of relational annotations
- *T*: a type system
- T_c: a subset of types in *T*
- *te*: an element type in *T*
- *(te)*: an elemental annotation
- *tr*: a relation type in *T*
- *(tr)*: a relational annotation

Experimental Results

- Date set: 36 historical facts from 26 true-false questions
- Experimental conditions: $N_d = 1000, N_p = 10, N_s = 3, \alpha = 0.1$
- T = {part-of-speech, named entity, dependency, semantic argument}, provided by Stanford CoreNLP and ClearNLP

T_c	P@1	MRR
(Baseline)	0.3611	0.4609
Named entity (person)	0.3889	0.4801
Dependency (nsubj, dobj)	0.3889	0.4755
Semantic argument (A0, A1)	0.3611	0.4639

P@1 is the percentage of historical facts where an answer-bearing passage is ranked at the first position. Mean reciprocal rank (MRR) is given as:

$$MRR = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{rank(q)}$$

- Named entities (person type) gain the best improvement; names of historic figures are a key element to amplify TF-IDF effects.
- Semantic argument annotations are sparse; a sentence and a passage barely have the same argument structure over the same tokens.

Conclusion

Novelties of this work

- We proposed a passage ranking model that can incorporate annotations of any type along with traditional retrieval models.
- The model improved passage ranking for QA on world history with named entity and dependency annotations.

Future work

- Refining the model so it can benefit from a combination of different annotations, including WordNet synsets and temporal relations
- Implementing a true-false judgment component for building an end-to-end world history QA system