Supplemental Material:
Who did What: A Large-Scale Person-Centered Cloze Dataset

Abstract

We include pseudocode for generating questions (Alg. 1) and multiple choice answer sets (Alg. 2).

Data: an article A
Result: either $null$, if no question can be formed from A, or a cloze question q and the deleted person named entity (NE) t.
$s \Leftarrow$ the first sentence of the article A.
if ***not*** $10 \leq |s| \leq 120$ **then** return $null$;
$E \Leftarrow$ The set of person NEs e in s such that e contains no more than three words. Named entities sharing a word with an earlier named entity are deleted.
if $|E| < 2$ **then** return $null$;
$T \Leftarrow$ constituent parse tree of s
for $e \in E$ *starting from the end of* s **do**
  $b \Leftarrow$ The node in T for person NE e.
  while $b.category \in$ {`NP`, `NNP`, `NNPS`} *and* $b.head = e$ *and no element of* $b.descendant$ *has category* `SBAR` **do**
    $b \Leftarrow b.parent$
  end
  if *no element of* $b.descendant$ *has head word "and"* **then**
    return (q, e) where q is the cloze question formed from deleting b from s.
  end
end
return $null$

Algorithm 1: Question Formation. Named entities are recognized by the Stanford NER system and parse trees are generated by the Stanford PCFG parser. Here $X.category$ is the syntactic category of parse node X, $X.parent$ is the parent-node of the node X, $X.descendant$ is the set of descendants of X and $X.head$ is the head word of X.

Data: a pair (q, e) returned by Algorithm 1.
Result: either $null$, if no appropriate passage can be found, or a passage a and multiple choice answer set C.
$p \Leftarrow null$
for $a \in RankedArticles(q, e)$ **do**
  $C \Leftarrow$ The set of person NEs in a different from e and not in q. Named entities appearing as sub-part of an earlier named entity are deleted.
  if $2 \leq |E| \leq 5$ **then** return (a, C);
end
return $null$

RankedArticles$(q, e)\{$
$A_r \Leftarrow \emptyset$
for $t \in \{1, 3, 7, 14\}$ **do**
  $A \Leftarrow Articles(q, e, t)$
  for $a \in A$ **do**
    if $isValid(a, q)$ **then**
      $A_r \Leftarrow A_r$ followed by a
    end
  end
end
return $A_r\}$

Articles$(q, e, t)\{$
Result: articles containing the person NE e, published within t days of the article from which q was taken, and ranked by Apache Lucene.$\}$

isValid$(a, q)\{$
a is a valid passage for q if the following hold:

- no sentence in a shares more than 78% of its words with the question q.
- no sentence in a contains the sequence of five words to the left of the blank in q, and similarly for the sequence to the right.
- a contains at least one of the person NEs in q. (All person NEs in q are different from e. Two named entities are considered the same if they share some words.) $\}$

Algorithm 2: Passage Selection