UWaterloo: Intent discovery with anchor text

UNIVERSITY OF WATERLOO

John A. Akinyemi and Charles L.A. Clarke

University of Waterloo, Waterloo, Canada

{jakinyem,claclark}@uwaterloo.ca

NTCIR INTENT Task Problem

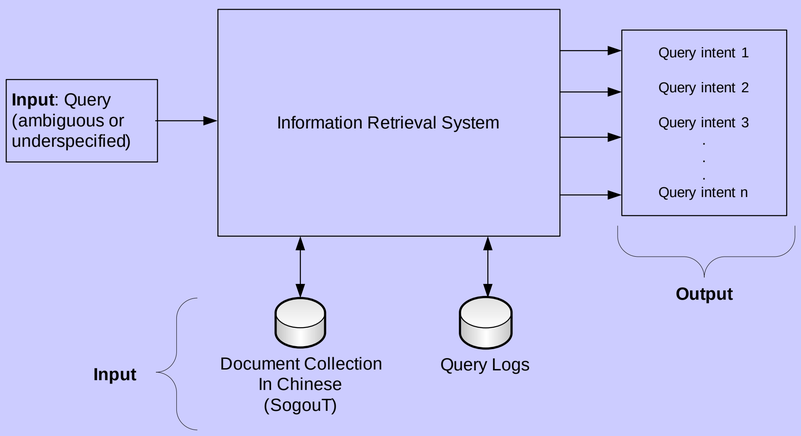

Figure 1: The Problem

The goal of the NTCIR INTENT Task [3] is to obtain diverse intents for provided queries from the provided test collection.

Method

- We explored anchor text and anchor link information for discovering diverse query intents.
- An anchor text is considered to have implicit diverse intents if it hyperlinks more than one target document.
- We mined implicit intents using link information between anchor texts and their corresponding target documents.

Corpus Pre-Processing

- Extract anchor text information from corpus.
- Convert the Chinese characters in anchor texts into their UTF-8 equivalent.
- Segment the anchor text UTF-8 representation into both unigram and bi-gram tokens.
- Index tuples of anchor text tokens where a tuple contains ⟨source document, anchor text (unigram and bi-gram UTF-8 character encoding), target document⟩ as a document unit.

Query Processing

- Segment queries into their unigram and bigram UTF-8 equivalents.

Retrieval Method

- Taking the provided queries as input, retrieval was done using a passage retrieval function [1, 2] rather than a traditional document retrieval function because:
  - ◇ anchor texts are short compared to an average document size, and
  - ◇ Passage retrieval is suitable for short documents.
  - ◇ Query terms occurrence and their close proximity in an anchor text are incorporated in the passage retrieval scoring function.
- We retrieved anchor texts and their target documents.
- For retrieved target documents having additional anchor text, we further retrieve their additional anchor texts.
- Rank anchor texts.
  - ◇ Let $t_1, ..., t_n$ represent an anchor text such that t_1 is the first term in the anchor text and t_n is the last term.
  - ◇ Anchor text scoring function takes as input:
    (i) the total number of query terms q_t,
    (ii) the ratio of the number of unique query terms q_t^u in an anchor text, and
    (iii) the total number of terms in the anchor text n.
  - ◇ The scoring function is given by:

$$score = |q_t| \cdot \frac{q_t^u}{n} \quad (1)$$

- Eliminate duplicate and noisy anchor texts
- Using Equation 1, we rank all retrieved anchor texts
- Ranked anchor texts become our `UWat-S-C-1` run.

Anchor text clustering

- Our `UWat-S-C-2` run takes as input the `UWat-S-C-1` run and groups anchor texts having very strong relationships on the *documents–anchors graph*.
- If two or more anchor text edges are connected to a particular target node, we put all the anchor texts in the same cluster.
- Thereafter, the highest scoring anchor text in each cluster is selected to represent the cluster.
- The selected clusters are ranked based on their scores and are submitted as our `UWat-S-C-2` run.

Submissions

Table 1 shows the official evaluation result of our submissions. In all cases, `UWat-S-C-2` outperforms `UWat-S-C-1`.

	@10	@30
UWat-S-C-1	0.239	0.324
UWat-S-C-2	0.332	0.494

Table 1: Result of submissions

Conclusions

- We have demonstrated that anchor text usage for intents discovery is promising.
- The much better performance of our `UWat-S-C-2` run against the `UWat-S-C-1` run also indicate the utility of anchor links as a reasonable criteria for clustering similar anchor texts and by extension similar documents.
- We envisage that a combination of our method and intent discovery that utilizes user interaction data extracted from query logs will produce better quality result.
  - ◇ We leave this as a future work.

References

[1] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting Redundancy in Question Answering. In *SIGIR*, pages 358–365, 2001.
[2] C. L. A. Clarke and E.L. Terra. Approximating the top-m passages in a parallel question answering system. In *CIKM*, pages 454–462, 2004.
[3] R. Song, M. Zhang, T. Sakai, M. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT Task, NTCIR-9 Proceedings. Tokyo, Japan. NII.