UWaterloo at NTCIR-9: Intent discovery with anchor text

John A. Akinyemi and Charles L.A. Clarke
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
{jakinyem, claclark}@cs.uwaterloo.ca

ABSTRACT

This paper describes our submission to the Intent Discovery task at the NTCIR-9. By treating the source and target documents of anchor texts as nodes, we utilized the anchor texts between the nodes as edges in a *documents–anchors graph* representation of the corpus. We extracted and indexed anchor links information from the provided SogouT corpus. Using the queries, anchor texts are retrieved from the index. Other anchor texts that link to the target documents of retrieved anchor texts are also retrieved. All the anchor texts are ranked and grouped to eliminate duplicates and near duplicates.

1. INTRODUCTION

The goal of the NTCIR INTENT Task [3] is to obtain diverse intents for provided queries from the provided test collection. The task organizers provided the SogouT collection which consists of Chinese text corpus, a query log of user interactions associated the corpus and queries that task participants are required to provide diverse intents for. We explored query intents discovery using anchor text and anchor link information. When an anchor text that appears in different locations in a source document or various source documents hyperlinks more than one target document, the anchor text is considered to have implicit diverse intents. Considering "*windows*" as an anchor text that hyperlinks various target documents related to Microsoft company, operating systems, software updates, and replacement windows; these variety in the target documents indicates the diversified intents that are possibly derivable from the "*windows*" query. We mined these implicit intents using links information between anchor texts and their corresponding source and target documents.

From the provided corpus, we extracted anchor texts, the source documents containing them and their target documents. The Chinese characters are encoded into their UTF-8 equivalent. We crudely segmented the anchor texts UTF-8 representation into their unigram and bigram tokens. Tuples of ⟨source document, anchor text, target document⟩ were indexed as units of documents.

The provided queries were also segmented into their unigram and bigram UTF-8 equivalents. Using a passage retrieval function on the index with the queries as inputs, we retrieved anchor texts and their target documents. For all retrieved target documents that have additional anchor text edges, we further retrieve all the additional anchor texts. All the anchor texts are ranked and grouped to eliminate duplicates and remove noisy anchor texts from the anchor text list. We submitted two runs (`UWat-S-C-1` and `UWat-S-C-2`) for evaluation.

2. EXPERIMENTAL DETAILS

We chose to use an established passage retrieval algorithm, which was originally developed as an initial retrieval step in a question answering system by Clarke et al. [1, 2] rather than a traditional document retrieval function because the anchor text and link information are short compared to an average document size. Passage retrieval is suitable for short documents. Occurrence of query terms as well as their close proximity in an anchor text are incorporated in the scoring function in the passage retrieval algorithm.

2.1 Anchor text scoring function

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. Our anchor text scoring function takes as input (i) the total number of query terms q_t and (ii) the ratio of the number of unique query terms q_t^u in an anchor text and the total number of terms in the anchor text n. The scoring function is given by:

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

Using Equation 1, we rank all retrieved anchor texts. The highest scoring anchor texts are submitted as our `UWat-S-C-1` run.

2.2 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.

3. SUBMISSIONS

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

4. CONCLUSIONS

We have demonstrated that anchor text usage for intents discovery is promising. The much better performance of our `UWat-S-C-2`

runid	I-rec			D-nDCG			D#-nDCG		
	@10	@20	@30	@10	@20	@30	@10	@20	@30
UWat-S-C-1	0.239	0.324	0.327	0.249	0.246	0.193	0.244	0.285	0.260
UWat-S-C-2	0.332	0.494	0.511	0.336	0.389	0.315	0.334	0.442	0.413

Table 1: Official Evaluation Result

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.

5. REFERENCES

[1] Charles L. A. Clarke, Gordon V. Cormack, D. I. E. Kisman, and Thomas R. Lynam. Question Answering by Passage Selection (MultiText Experiments for TREC-9). In *TREC*, 2000.

[2] Clarke, Charles L. A. and Terra, Egidio L. 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.P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT Task, NTCIR-9 Proceedings. Tokyo, Japan. NII. [To appear].