Refining Methodologies for the INEX 2013 Snippet Generation and Tweet Contextualization Tracks

Carolyn J. Crouch, Donald B. Crouch, Swapnil Nawale,
Mihir Atmakuri, Kiran Bushireddy, Sameer Kulkarni

Department of Computer Science
University of Minnesota Duluth
Duluth, MN 55812
(218) 726-7607
ccrouch@d.umn.edu

Abstract. This paper describes our current experiments in snippet generation and tweet contextualization. These experiments are based on work reported in 2011 [2] and 2012 [1] and represent refinements of those earlier techniques. Four of our snippet generation runs produced top-ranked results in the INEX 2012 competition; these serve as the basis for our 2013 experiments in snippet generation. Our 2013 tweet contextualization run produced a top-ranked result as well. The methodologies employed and the results obtained are described below.

1 Introduction

In earlier years of the INEX competitions, major tracks focused on the retrieval of focused elements. As described in [2], we developed a simple method for producing focused elements that, when applied to the individual 2009 and 2010 Ad Hoc tasks, produced a result that fell in each case into the ranks of the top ten. (The details may be seen in [2].) Our retrieval of good focused elements—i.e., elements which when evaluated are competitive with those in the top-ten highest ranked results for that task—is described very briefly below. We use these focused elements as a basis for snippet generation in subsequent experiments.

To retrieve good focused elements in response to a query, we use article retrieval to identify the articles of interest along with dynamic element retrieval [3] to produce the elements and then apply a focusing strategy to that element set. (Dynamic element retrieval builds the document tree at execution time, based on a stored schema representing a pre-order traversal of the document created when it is parsed and a terminal node index of the collection.) *Lnu-ltu* term weighting [8], designed to deal with differences in the lengths of vectors, is utilized with inner product to produce a rank-ordered list of elements from each document. To produce focused elements, we use a focusing strategy to remove overlap. For example, the *correlation strategy* chooses the highest correlating element along a path as the focused element, without restriction on element type, whereas the *child strategy* chooses the terminal element along a path as the focused element, ignoring correlation. Our system is based on the Vector Space Model [7]; basic functions are performed using Smart [6].

2 Snippet Generation (INEX 2012 and 2013)

One goal for this year centers on the feasibility of generating snippets based on focused elements. To facilitate this goal, we first apply a focusing strategy (correlation, child) to focus the element set and then select the highest ranking focused element as the basis for the snippet representing that document. (We refer to this as the raw snippet.) The snippet is then generated using a snippet refinement algorithm, which rearranges the sentences in the final snippet based on a simple scoring mechanism. Method 1 is based on the number of unique query terms in the sentence, and method 2 is inspired by BLEU [5]. The basic framework is as follows.

For each document in the 2012 reference run, a ranked list of focused elements is generated using one of the two focusing strategies. The text of the elements is merged, and the resultant sentences are ranked based on one of the two scoring mechanisms. Three of the resulting runs produced a top-ten result, namely, (1) child with scoring method 1, (2) child with scoring mechanism 2, (3) correlation with scoring method 1. These runs received ranks 1, 7, and 10, respectively. One other run, at rank 3, was based on the text of the article, rather than its focused elements. Our 2013 INEX run was not evaluated due to late submission, but it is based on these four approaches.

3 Tweet Contextualization (2013)

Our tweet contextualization experiments use the Indri and Lucene search engines. Indri is used for a primary indexing of the Wikipedia corpus; it retrieves a small set of relevant documents for each query. Lucene is used for hashtag term splitting. Hashtags can have multiple terms without specification of word boundaries; we use a word segmentation algorithm by Norvig [4] for this purpose. We perform sentence retrieval on the documents retrieved by Indri and rank them by checking n-gram overlap between the sentence and query terms. Top-ranked sentences are combined to form a 500-word summary. The run described herein ranked at 7 in the official ranking and at 6 with respect to readability. Future work focuses on increasing the accuracy of hashtag term splitting and improving the readability of the summaries.

4 Conclusions

We conclude from the results of the task evaluation that our approach to snippet evaluation is soundly based. Additional investigation is needed to gain perspective on how good snippets are generated. The tweet contextualization experiments appear to be well underway.

References

1. Crouch, C., Crouch, D., Chittilla, S., Nagalla, S., Kulkarni, S. and Nawale, S.: The 2012 INEX snippet and tweet contextualization tasks, *CLEF 2012*, Rome, Italy (2012).
2. Crouch, C., Crouch, D., Acquilla. N., Banhatti, R., Chitilla, S., Nagalla, S., Narenvarapu, R. Focused elements and snippets. In: Geva, J., Kamps, J., Schenkel, R.: (eds) *INEX 2011*, LNCS 7414, 295-299 (2012).
3. Crouch, C.: Dynamic element retrieval in a structured environment. *ACM TOIS* 24(4), 437-454 (2006).
4. Norvig, P.: Natural language corpus data. Beautiful Data, 219-141 (2009).
5. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Assoc. for Computational Linguistics*, Philadelphia, PA, 311-318 (2002).
6. Salton, G., (ed.): The Smart Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall (1971).
7. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Comm. ACM 18(11), 613-620 (1975).
8. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21-29, Zurich, Switzerland (1996).