The 2012 INEX Snippet and Tweet Contextualization Tasks

Carolyn J. Crouch, Donald B. Crouch, Sai Chittilla,
Supraja Nagalla, Sameer Kulkarni, Swapnil Nawale

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

Abstract. This paper reports on our current experiments involving the Snippet and Tweet Contextualization Tracks of the 2012 INEX competition. Most of this work in snippet generation extends our earlier (2011) approach, described in [4], which produced a top-ranked result. The source of the snippet in these experiments is the top-ranked focused element(s) of the document in question. Another approach is based on using the document itself as the source of the snippet. Having identified the source, the snippet is then generated based on simple basic methodologies described herein. We also describe our experiments in tweet contextualization, a new track for INEX in 2012.

1 Introduction

In both 2009 and 2010, major tracks in the INEX competition centered upon the retrieval of what were referred to as *focused* elements. A focused element is by definition non-overlapping. We were able, as described in [1, 2, 5], to produce a methodology the results of which would rank in the top 10 for all the focused tasks of 2009 and 2010. This approach, described in detail in [4], is recapped briefly below.

To produce good (i.e., highly ranked) focused elements in response to a query, we first perform a document retrieval to identify the articles of interest. Our system is based on the Vector Space Model [7]; basic retrieval functions are performed by Smart [6]. To produce the set of elements corresponding to each article, we use an approach which we call flexible or dynamic element retrieval—Flex for short. (See [3] for details.) Flex allows us to produce the document tree, bottom up, at run time, based on a schema representing the structure of the document, generated during parsing, and a terminal node index of the collection. *Lnu-ltu* term weighting [8] is utilized with inner product to produce a rank-ordered list of all the elements of the document with respect to the query.

These elements are overlapping, so we now apply a focusing strategy to produce the non-overlapping elements of with the document tree. In the 2012 experiments, we use two focusing strategies, namely, the correlation strategy, which chooses the high-

est correlating element along a path as the focused element (without restriction as to element type) and the child strategy, which chooses the terminal element along a path as the focused element (regardless of correlation). The result is a rank-ordered list of focused elements associated with each document.

If we were to select one element that best represents the content of a document with respect to the query, one might do worse than to consider the highest ranking focused element(s) of that document. That element may not prove to be the best choice—it is certainly only one of many choices—but it appeared to us to be a viable source for snippet generation. Our snippet results in 2011 were based on this premise; one result placed in the top 10 of the official rankings despite our failure to produce a clean, easily readable version in each case.

2 INEX 2012: Snippet Generation

The snippet generation algorithms used in our 2012 experiments are similar in basic strategy, varying only in terms of the source of the snippet (focused elements or article), focusing strategy (correlation or child), and ranking algorithm (the first based on a simple function of the number of query terms in the snippet and the second a BLEU approach based on the number of query vs snippet n-grams, as applied to the sentences extracted). One experiment uses the text directly from the article as the snippet. We are currently awaiting INEX evaluation so as to enable assessment of the snippets.

3 INEX 2012: Tweet Generation

Our tweet conceptualization experiments use Indri to retrieve a small set of documents for each query. The corresponding sentences are ranked with respect to their similarity to the query based on several simple approaches, including word n-grams. A 500 word summary is then constructed using the top-ranked sentences in rank order. Two of these runs, evaluated earlier this year, rank 1 and 2 out of 33 in the official ranking with respect to average scores. These early results appear informative but would clearly benefit from increased readability. Future work is directed at this task and at related issues of interest that have not yet been addressed.

References

1. Acquilla, N.: Improving results for the INEX 2009 and 2010 Focused tasks. M.S. Thesis, Department of Computer Science, University of Minnesota Duluth (2011). http://www.d.umn.edu/cs/thesis/acquilla.pdf
2. Banhatti, R.: Improving results for the INEX 2009 Thorough and 2010 Efficiency tasks. M.S. Thesis, Department of Computer Science, University of Minnesota Duluth (2011). http://www.d.umn.edu/cs/thesis/banhatti.pdf
3. Crouch, C.: Dynamic element retrieval in a structured environment. ACM TOIS 24(4), 437-454 (2006).

4. Crouch, C., Crouch D., Acquilla, N., Banhatta, R., Chittilla, S., Nagalla, N., Navenvarapu, R: Focused elements and snippets. In: Geva, et al. (eds). Focused Retrieval of Content and Structure, LNCS 7424, Springer, (2012). [to appear]
5. Narendravarapu, R.: Improving results for the INEX 2009 and 2010 Relevant-in-Context tasks. M.S. Thesis, Department of Computer Science, University of Minnesota Duluth (2011). http://www.d.umn.edu/cs/thesis/ narendravarapu.pdf
6. Salton, G., (ed.): The Smart 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).