Experiment on Combining Sources of Evidence for Passage Retrieval [*]

Alexander Gelbukh,[1] NamO Kang,[2] SangYong Han [2+]

[1] National Polytechnic Institute, Mexico
gelbukh@gelbukh.com, www.Gelbukh.com
[2] Chung-Ang University, Korea
kang@archi.cse.cau.ac.kr, hansy@cau.ac.kr

Abstract. Passage retrieval consists in identifying short but informative runs of a long text, given a specific user query. We discuss the sources of evidence that help choosing likely high-quality passages, such as relevance to the user query and self-containedness. These measures are different from the traditional information retrieval procedure due to the use of the context of the passage.

1 Introduction

Unlike full document retrieval—a traditional task of information retrieval—passage retrieval task [2, 5] consists in identifying in a long document (or collection of long documents) short text runs relevant for a specific user query, which—unlike in question answering task—do not allow a simple factual answer.

In this paper we discuss the parameters affecting selection of such passages from the text of the document. We intentionally do not give any specific formulas since our experimental result do not yet allow us to reliably argue in favor of a specific way of calculation of these parameters.

2 The Method

Our algorithm consists in the following steps:

- Preprocessing,
- Candidate passage generation,
- Assessing various properties of each candidate passage,
- Combining the obtained scores for each property in a single value overall score.

Then the passages are presented to the user in the order of obtained scores. The steps of the algorithm and the specific quality measures are described in the following subsections.

[*] This research was supported by ITRI of the Chung-Ang University. Work was done when the first author was on Sabbatical leave at Chung-Ang University. [+]Corresponding author.

A. Gelbukh, N.O Kang, S.Y. Han. Experiment on Combining Sources of Evidence for
Passage Retrieval. Computational Linguistics and Intelligent Text Processing
(CICLing-2005), Lecture Notes in Computer Science, N 3406, Springer-Verlag, 2005,
pp. 617–620.

Preprocessing Currently we only apply tokenization and stemming. In the future, anaphora resolution would be desirable, as well as resolution of other types of coreference, including hidden anaphora.

Generating Candidate Passage We select as possible candidates all text windows containing from 5 to 1000 words. At this stage, we do not care of their suitability, since unfit candidates will be ruled out later on, so the lower and upper limits on the passage length were motivated only by efficiency considerations. In the future, generation of possible candidates can be made in a more intelligent manner.

In case of retrieval over a multiple-document collection, we generate all such candidate windows for all the texts in the collection.

Scoring the Candidates Each candidate window is evaluated independently. The properties to be assessed are related with two main requirements, which often are contradicting:

- Desired passages should be *relevant*, i.e., should not contain irrelevant words,
- They should be *self-contained*, i.e., contain enough context to be understandable.

Note that the generated candidates are overlapping, so that for any passage (shorter than 1000 words) there are other passages containing it. With a proper balance between the two contradicting requirements, preference is given to slightly longer passages that contain the same relevant information but are more self-contained. However, too long passages receive too low relevance score and are ruled out.

Below we present the specific criteria used in our current experiments. Other criteria can be added in the future.

Scoring the Relevance for the Query To score relevance, we use known information retrieval techniques, considering each passage as an independent document. In contrast to the usual information retrieval task, however, we have access to the global context of the document, which helps disambiguation as well as more accurate weighting of the importance of keywords.

To evaluate the relevance of a given passage for the user query, we use vector space similarity measure [1]. Note that this measure gives lower scores to longer passages containing more words irrelevant for the query. With this, of nested candidate windows, a shorter window containing more keywords from the query would be preferred. On the other hand, since the vector measure uses frequencies, a longer passage that contains a greater number of relevant words can receive greater score.

To allow the latter effect, the weights of the words from the query should be set to a much greater value than that of the terms not appearing in the query. This allows preferring a window with, say, 10 extra irrelevant words, to include an additional relevant word. This is a parameter that can be adjusted to control the desired size of the passages, i.e., the preference of relevance over completeness of the results.

Another factor affecting the keyword weighting is IDF weighting known from information retrieval. For usual documents, IDF weighting is measured over the whole collection, so that all documents in the collection contribute equally to the IDF weights. In case of passages, they are in linear context of the surrounding text. On the one hand, this can help in word sense disambiguation and anaphora resolution (which we do not tough upon in this paper). On the other hand, this allows for more accurate calculation of IDF value. Namely, in addition to the usual IDF, which is inversely proportional to the number of documents in the collection containing the word in question, we use a document-related value, which is inversely proportional to the number of occurrences of the given word in the paragraphs of the same document. We scale this additional value by the distance (in paragraphs) from the given passage, so that it decreases exponentially with the distance.

The reason for this additional value is that even if a word is not very frequent in general language or in the whole collection, in can be frequent in the given (long) document, and in this case it expresses an idea that is probably already known to the user, thus not contributing to the information value of the extracted passages. However, if the word is used in the same document far from the passage in question, it can express a different idea, which does not affect the given passage.

Scoring the Self-Containedness The requirement of self-containedness implies that desired passages, in particular, should not contain logical references or dependencies on outside of the passage. We use heuristic knowledge-poor approaches to asses the suitability of the passage. In particular, to assure the absence, or to minimize the number of, references outside the given passage, we prefer the passages that:

- Lay at the boundaries of structural units of the text, e.g., beginning of a chapter,
- Represent thematic threads of the text and do not have many thematic relations with the neighboring sentences.

Accordingly, we score higher the candidate windows that lay at the boundary of structural units of the document. Namely, we give an additional bonus to the windows lying at the beginning of a unit, and a smaller bonus to those at the end of the unit. Chapter boundaries are more important than section boundaries, which in turn are more important than paragraph boundaries, in the sense of a greater bonus. As a simplification, one can choose to consider only complete sentences, thus ensuring some degree of self-containedness, but losing short sub-sentences of complex sentences that otherwise would be good candidates.

Similarly, we boost the scores of the candidate windows that show high degree of internal interrelatedness between words and low degree of relatedness between their words to the words in surrounding context. The relatedness with the context is less dangerous at the end of the passage than at its beginning. Indeed, a passage related to the preceding context is likely to develop on the ideas explained earlier, and thus is likely not to be understandable out of context. On the other hand, if a passage is related to the following context, this may indicate that the ideas introduced in the passage are developed later on, but the passage still should be understandable without this continuation.

The idea of linguistic word relatedness is that, say, *teacher* is related to *school* and is not related to *sleep*; there exists a number of word relatedness measures suggested in literature **[**2, 4**]**. To determine the contribution of every specific pair of running words in the text to the inner interrelatedness of the paragraph or to the relatedness of the paragraph to the context, we scale the linguistic relatedness of the corresponding words by an exponentially decreasing function of the distance between them.

Combining the Scores We combine the partial scores in a multiplicative manner, so that a candidate window that is either irrelevant (even if very comprehensible) or incomprehensible out of context (even if very relevant) is not presented to the user. The combination can be tuned to give preference to shorter (more relevant) or longer (more self-contained) candidates.

Experimental Results We are not aware of any standard evaluation procedure for passage retrieval. We evaluated the results by manual inspection of the answers to sample questions. For example, the top three passages retrieved for the query *wars between England and France* from *A Child's History of England* by Charles Dickens, 164,772 words, are the following:

- *The Queen's husband who was now mostly abroad in his own dominions and generally made a coarse jest of her to his more familiar courtiers was at war with France and came over to seek the assistance of England. England was very unwilling to engage in a French war for his sake but it happened that the King of France at this very time aided a descent upon the English coast.*
- *As his one merry head might have been far from safe if these things had been known they were kept very quiet and war was declared by France and England against the Dutch.*
- Same as 1 plus: *Hence war was declared greatly to Philip's satisfaction and the Queen raised a sum of money with which to carry it on by every unjustifiable means in her power.*

As one can see, the lack of semantic processing (ignoring the word *between* in the query) results in some passages in fact unrelated to the query, like the second passage in the table. Elements of meaning understanding can be a topic of future work.

3 Conclusions

Linear positioning of candidate text windows in the context and their variable length allow for estimating some characteristics of passages different from those used in traditional information retrieval, e.g., topical relatedness to the context. However, specific formulas for estimating of such parameters and the ways to tune their combination to obtain optimal results are the topics of our current and future research.

References

1. R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
2. C. L. A. Clarke, G. V. Cormack, T. R. Lynam, E. L. Terra. Question Answering by Passage Selection. In *Advances in Open Domain Question Answering,* Kluwer, 2004.
3. G. Hirst, D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In: C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, Cambridge, MA: The MIT Press, 1998.
4. S. Patwardhan, S. Banerjee, T. Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (CICLing-2003), LNCS N 2588, Springer, 2003, p. 241–257.
5. G. Salton, J. Allan, C. Buckley. Approaches to passage retrieval in full text information systems. ACM SIGIR Research and Development in Information Retrieval, 1993, 49–58.