Feeding the Second Screen: Semantic Linking based on Subtitles (Abstract) *

Daan Odijk
d.odijk@uva.nl

Edgar Meij
edgar.meij@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam
Amsterdam

ABSTRACT

Television broadcasts are increasingly consumed on an interactive device or with such a device in the vicinity. Around 70% of tablet and smartphone owners use their devices while watching television [11]. This allows broadcasters to provide consumers with additional background information that they may bookmark for later consumption in applications such as depicted in Figure 1.

For live television, edited broadcast-specific content to be used on second screens is hard to prepare in advance. We present an approach for automatically generating links to background information in real-time, to be used on second screens. We base our semantic linking approach for television broadcasts on subtitles and Wikipedia, thereby effectively casting the task as one of identifying and generating links for elements in the stream of subtitles.

The process of automatically generating links to Wikipedia is commonly known as *semantic linking* and has received much attention in recent years [3, 6, 7, 9, 10]. Such links are typically explanatory, enriching the link source with definitions or background information [2, 4]. Recent work has considered semantic linking for short texts such as queries and microblogs [6–8]. The performance of generic methods for semantic linking deteriorates in such settings, as language usage is creative and context virtually absent.

While link generation has received considerable attention in recent years, our task has unique demands that require an approach that needs to (i) be high-precision oriented, (ii) perform in real-time, (iii) work in a streaming setting, and (iv) typically, with a very limited context.

We propose a learning to rerank approach to improve upon a strong baseline retrieval model for generating links from streaming text. In addition, we model context using a graph-based approach. This approach is particularly appropriate in our setting as it allows us to combine a number of context-based signals in streaming text and capture the core topics relevant for a broadcast, while allowing real-time updates to reflect the progression of topics being dealt with in the broadcast. Our graph-based context model is highly accurate, fast, allows us to disambiguate between candidate links and capture the context as it is being built up.

Our main contribution is a set of effective feature-based methods for performing real-time semantic linking. We show how a learning to rerank approach for semantic linking performs on the task of real-time semantic linking, in terms of effectiveness and efficiency. We extend this approach with a graph-based method to keep track of context in a textual stream and show how this can further improve effectiveness. By investigating the effectiveness and efficiency of individual features we provide insight in how to improve effectiveness while maintaining efficiency for this task. Additional contributions include a formulation of a new task: semantic linking of a textual stream, and the release of a dataset[1] for this new task, including ground truth.

*The full version of this paper will appear in OAIR 2013 [12].

DIR 2013, April 26, 2013, Delft, The Netherlands.

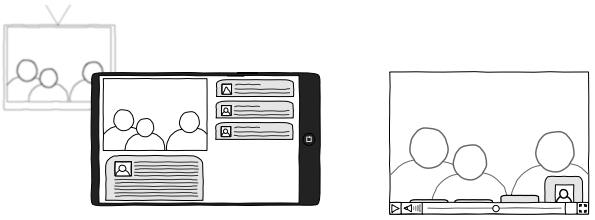

Figure 1: Sketches of a second screen (left) and an interactive video player (right) showing links to background information, synchronized with a television broadcast. Links pop up briefly when relevant and are available for bookmarking or exploring.

Real-Time Semantic Linking. Our approach to real-time semantic linking consists of a retrieval model that is based on how links between Wikipedia articles are created. Our method for real-time link generation consists of three steps: link candidate finding, ranking and reranking. In this retrieval model, each Wikipedia article is represented by the anchors that are used to link to it within Wikipedia. The first, recall-oriented step is aimed at finding as many link candidates as possible. Here, we produce a set of link candidates that each link to a Wikipedia article. To this end, we perform lexical matching in the subtitles of each constituent n-gram with the anchor texts found in Wikipedia.

The second step is to rank the link candidates in L. In particular, we can use statistics on the anchor text usage. We consider the prior probability that anchor text a links to Wikipedia article w:

$$COMMONNESS(a, w) = \frac{|L_{a,w}|}{\sum_{w' \in W} |L_{a,w'}|}, \qquad (1)$$

where $L_{a,w}$ denotes the set of all links with anchor text a and target w. The intuition is that link candidates with anchors that always link to the same target are more likely to be a correct representation than those where anchor text is used more often to link to other targets. We consider these first two steps our baseline retrieval model.

The third step is aimed at improving precision using a learning to rerank approach, that was effective on similar tasks [5, 8, 10]. For link candidates many ranking criteria are in play, making learning to rerank particularly appropriate. We use a set of lightweight features (based on [8]), that can be computed online. These 26 features

[1]The dataset will be shared upon publication of [12]; it consists of subtitles for 50 video segments, with more than 1,500 manually annotated links.

Table 1: Semantic linking results with classification time. Significant differences, tested using a two-tailed paired t-test, are indicated ▲ ($p < 0.01$); the position indicates whether the comparison is against line 1 (left most) or line 2 (right most).

	Average classification time per line (in ms)	R-Prec	MAP
1. Baseline retrieval model	54	0.5753	0.6235
2. Learning to rerank approach	99	0.7177▲	0.7884▲
3. Learning to rerank + context	108	0.7454▲▲	0.8219▲▲

are organized in four groups based on their source: textual anchor, target Wikipedia article and anchor+target. This set includes simple textual features, link probability measures and visitor statistics for a Wikipedia article. The full set of features is listed in [12].

We use a decision tree based approach as it has outperformed Naive Bayes and Support Vector Machines on similar tasks [8, 10]. We choose Random Forests [1] as it is robust, efficient and easily parallelizable.

Modeling Context. Link generation methods that rely on an entire document are not suited for use in a streaming text context as such methods are computationally expensive. What we need, instead, is a method to model context that can be incrementally updated and allows for easily computing features for link candidates.

We model the context of a textual stream as an undirected graph. The graph reflects the content of the textual stream and encodes the structure. This results in a smaller distance for things mentioned together. Furthermore, nodes for Wikipedia articles that are mentioned more often, will have more anchors connecting to them, making them more central and thus more important in the graph.

To feed our learning to rerank approach with information from the context graph we compute a number of features for each link candidate. First, we compute the degree of the target Wikipedia article in this graph. To measure how closely connected a target is, we compute degree centrality. Finally, we measure the importance of a target by computing its PageRank [13].

Experimental evaluation. To measure the effectiveness and efficiency of our proposed approach to semantic linking, we use the subtitles of six episodes of a live daily talk show. The subtitles are generated during live broadcast by a professional and are intended for the hearing impaired. From these subtitles, video segments are identified, each covering a single item of the talk show. Our data set consists of 5,173 lines in 50 video segments, with 6.97 terms per line. The broadcast time of all video segments combined is 6 hours, 3 minutes and 41 seconds.

In order to train the supervised machine learning methods and evaluate the end result, we need to establish a gold standard. We have asked a trained human annotator to manually identify links that are relevant for a wide audience. A total of 1,596 links have been identified, 150 with a NIL target and 1,446 with a target Wikipedia article, linking to 897 unique articles, around 17.94 unique articles per video segment and 2.47 unique articles per minute.

Results and Discussion. An overview of the results is shown in Table 1. First, we consider the performance of our baseline retrieval model. Line 1 in Table 1 shows the scores for the ranking baseline. The recall oriented link candidate finding step produces 120,223 links with 42,265 target articles, including 771 known targets that are in the ground truth (a recall of 0.8595). With this many link candidates, there is a clear need for ranking. The ranking baseline achieves reasonable effectiveness scores; these numbers are comparable to the literature, while leaving room for improvement.

The results for our learning to rerank approach (Line 2) show that it can be highly effective and significantly improve over the retrieval baseline. We can achieve this high effectiveness at an average online classification time of less than 100 milliseconds, making the learning to rerank approach efficient and suited for usage in real time. The results for the learning to rerank runs with context features added are listed in line 3. Compared to the learning to rerank approach, we are able to achieve significantly higher performance.

Conclusion. Motivated by the rise in so-called second screen applications we introduced a new task: real-time semantic linking of streaming text. We have created a dataset for this task. We have shown that learning to rerank can be applied to significantly improve an already competitive retrieval baseline and that this can be done in real-time. Additionally, we have shown that by modeling context as a graph we can significantly improve the effectiveness of this learning to rerank approach. This graph-based method to keep track of context is especially well-suited for the streaming text, as we can incrementally update the context model.

Acknowledgments

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the BILAND project funded by the CLARIN-nl program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), and the Netherlands eScience Center under project number 027.012.105.

REFERENCES

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[2] M. Bron, B. Huurnink, and M. de Rijke. Linking archives using document enrichment and term selection. In *TPDL '11*, pages 360–371. Springer, 2011.
[3] P. Ferragina and U. Scaiella. TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM '10*, pages 1625–1628. ACM, 2010.
[4] J. He, M. de Rijke, M. Sevenster, R. van Ommering, and Y. Qian. Automatic link generation with Wikipedia: A case study in annotating radiology reports. In *CIKM '11*, pages 1867–1876. ACM, 2011.
[5] M. Larson, E. Newman, and G. Jones. Overview of VideoCLEF 2009. In *CLEF '09*, pages 354–368. Springer, 2010.
[6] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning semantic query suggestions. In *ISWC '09*, pages 424–440. Springer, 2009.
[7] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Mapping queries to the linking open data cloud: A case study using DBpedia. *J. Web Semantics*, 9(4):418–433, 2011.
[8] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM 2012*, pages 563–572. ACM, 2012.
[9] R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242. ACM, 2007.
[10] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM '08*, pages 509–518. ACM, 2008.
[11] Nielsen. In the U.S., tablets are TV buddies while ereaders make great bedfellows, May 2012. `http://bit.ly/L4lf9E` [Online; accessed May 2012].
[12] D. Odijk, E. Meij, and M. de Rijke. Feeding the Second Screen: Semantic Linking based on Subtitles. In *Open research Areas in Information Retrieval (OAIR 2013)*, Lisbon, Portugal, 2013.
[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.