A Plan for Ancillary Copyright: Original Snippets

Martin Potthast[1] Wei-Fan Chen[2] Matthias Hagen[3] Benno Stein[2]

[1]Leipzig University
martin.potthast@uni-leipzig.de

[2]Bauhaus-Universität Weimar
<first>.<last>@uni-weimar.de

[3]Halle University
matthias.hagen@informatik.uni-halle.de

Abstract

The snippets that web search engines generate for their result presentation are extracted from the retrieved web pages, reusing pieces of text that match a user's query. Copyright owners of the retrieved web pages are typically not asked for usage rights. This long-time practice now faces increasing backlash from news publishers, legal action, and even new legislation in Germany and Spain: the so-called ancillary copyright for news publishers. This copyright law restricts the fair use of intellectual property of news publishers, allowing them to raise claims for monetary compensation when their text is reused, even within snippets. If passed at the EU level, ancillary copyright could severely impact future information system development. This paper promotes a "technological remedy", namely, to synthesize *true original* snippets without text reuse.

1 Introduction

An organic search result for a keyword query on a web search engine is typically displayed as title and URL along with a brief excerpt of the respective page, showing selected pieces of text that contain keywords from the query, the *snippet*. Snippets guide users in deciding which of the pages on a search results page to visit, if any. Since snippets are extracted from the found web pages, they form a kind of text reuse. Reusing a third party's text is governed by copyright laws and typically requires written consent. The operators of web search engines have been exempt from this regulation under fair use laws. These exemptions are currently being reconsidered.

In recent years, news publishers have raised claims for compensation from search engine companies for snippets generated from their articles. Their argument is as follows: search engines and news aggregators earn money based on the publishers' intellectual property, and, since snippets are informative, they may prevent users from visiting the related news article, depriving them of ad revenue. While no one forces the publishers to have their articles indexed, they also claim to be left with no alternative to the de facto monopolist on most search markets, Google. The fact that search engines nowadays aim at answering certain queries directly on search results pages, often based on content lifted from third party web pages, does not serve to deescalate the dispute: every query answered directly by a search engine takes away traffic from the web pages it indexes, undermining the ad revenue model which funded the creation of apparently useful pieces of information in the first place. Following this line of argumentation, publishers successfully lobbied for political support: the so-called ancillary copyright for news publishers has been passed into law in Germany and Spain. Despite the German version still exempting individual words or "smallest text snippets,"[1] Google instantly demanded free-of-charge usage rights from all major German publishers, delisting those who did not agree, whereas the Spanish law[2] caused the shutdown of Google News in Spain.[3] While the European Union—amidst a fierce public debate among stakeholders both in favor as well as opposed—deliberates an ancillary copyright for all of its members and all kinds of information systems (not only search engines), Google News has recently been redesigned worldwide: the new version does not show snippets anymore.[4] Figures 1 and 2 contrast the new with the old layout.

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

[1] https://www.gesetze-im-internet.de/urhg/__87f.html (German)
[2] https://www.boe.es/boe/dias/2014/11/05/pdfs/BOE-A-2014-11404.pdf (Spanish)
[3] https://europe.googleblog.com/2014/12/an-update-on-google-news-in-spain.html
[4] https://www.blog.google/topics/journalism-news/redesigning-google-news-everyone

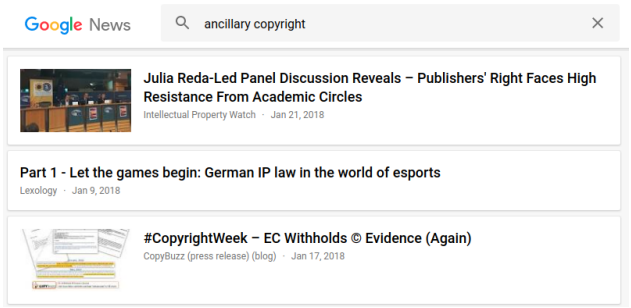

Figure 1: New Google News layout without snippets.

Based on our comprehensive literature survey (Section 2), we are unaware of any evidence that the usability of a search engine is improved by dropping snippets. However, despite recent experiments showing that users may prefer longer snippets over shorter ones [MAM17], not a single experiment has quantified the impact of dropping snippets. Therefore, Google must be given the benefit of the doubt, since extensive A/B tests may have revealed that snippets are unimportant for Google News. Meanwhile, Google recently "reintroduced" featured snippets to the main search engine, where the search result that best answers a question query is highlighted by showing it in a box and above the blue link and the green URL instead of below. Google claims that despite "concerns that they might cause publishers to lose traffic", "it quickly became clear that featured snippets do indeed drive traffic."[5]

Similarly, we are also unaware of any evidence that snippets are useful only if they reuse text from the web page described. This thought gave us a subversive idea: What if a snippet was an *original* explanation of how a web page relates to a query? This would resolve the quandary to some extent since search engines need no longer rely on the intellectual property of others to present their search results, but can resort to technology for snippet synthesis instead. With deep learning-based text generation on the rise, this does not appear impossible, anymore, albeit very difficult.

2 Background and Related Work

Snippet generation is a variant of extractive summarization, where the summaries are biased toward the queries. Extractive summarization and information retrieval have common ancestry, with Luhn, the inventor of term frequency weighting, being one of the earliest contributors [Bax58, Luh58]. Current research on snippet generation for search engines focuses on extractive summarization: Tombros and Sanderson [TS98] ascertained the importance that snippets relate to a user's query, while Brin and Page [BP98] implemented query-biased snippets for the first version of Google.

[5] https://www.blog.google/products/search/reintroduction-googles-featured-snippets

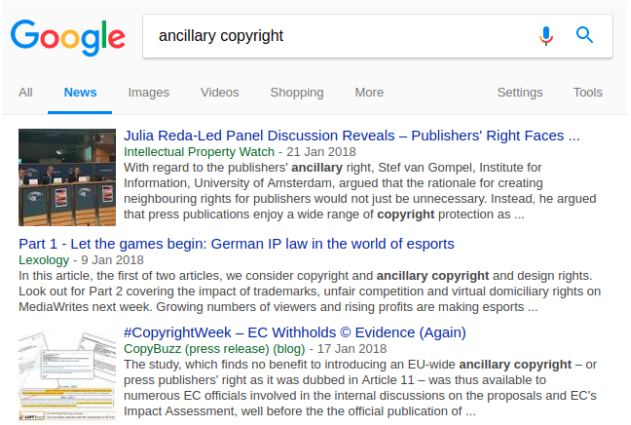

Figure 2: Google News as it used to be, obtained via the "News" facet of the main search engine.

White *et al.* [WRJ02a, WRJ02b] found that snippets should be re-generated based on implicit relevance feedback, selecting different sentences when a user returns to a search results page. To speed up snippet generation, Turpin *et al.* [TTHW07] evaluate software architectures based on compressed data structures and RAM caching. Bando *et al.* [BST10] ask humans to manually create reuse snippets, comparing the results to machine-generated reuse snippets. They observe that humans select the same pieces of text as machines in around 73% of cases. Savenkov *et al.* [SBL11] survey approaches regarding the evaluation of snippet generation, suggesting automated evaluation approaches and A/B testing, which both can only be trained (used) if a search engine with a reasonably large user base is available. Thomaidou *et al.* [TLKV13] consider the special case of snippets generated for ads shown on search results pages to allow users to understand how the ads relate to their queries. Further research has been invested into studying how the length of snippets affects perceived search result quality on desktops [MAM17, KHL08] and mobile devices, where screen space is limited [KTS+17]. Eye-tracking studies have been conducted to determine to what parts of a results page users pay most attention [GJG04, CG07]; unsurprisingly, snippets play a major role. Finally, reuse snippets are also generated in XML retrieval [HLC08] and semantic web search [PWTY08].

The companion task to extractive summarization is *abstractive* summarization, where summaries are synthesized without text reuse. Generating abstractive summaries has been a long-standing task in the natural language generation community [GG17], yet, it has not been applied to snippet generation. In their user study, Bando *et al.* [BST10] come close, using manually written, *original* snippets as a gold standard to evaluate snippets that were generated automatically and manually by extracting text from a web page. It was shown that humans pay attention to the same

Table 1: Survey: How often do you read snippets?

Always	Often	Sometimes	Seldom	Never	$\sum$
1782	2652	1470	87	9	6000
29.7%	44.2%	24.5%	1.4%	0.2%	100%

parts of a document when composing an original snippet compared to when selecting sentences for a snippet. Machines sometimes select different sentences to generate reuse snippets, leaving room for improvement. Recently, neural network models have made great progress toward the task of generating abstractive summaries [CAR16, NZN$^+$16, RCW15, SLM17], which renders snippet synthesis feasible if the lack of large-scale training data can be overcome.

3 Discussion and Future Work

All things considered, the proponents of ancillary copyright have a point: an information economy whose information sources are funded by displaying ads to information consumers cannot withstand information intermediaries that take the information from the sources and share it directly with the consumers for their own benefit. If the "plight" of news publishers does not convince, perhaps that of Wikipedia does: its ongoing decline of editors since 2007 [SCCP09] has been attributed, among other things, to Google's oneboxes [MJH17], which have been introduced around that time. But the opposition has a point, too: information intermediaries offer high-quality services to both sources and consumers of information free of charge; their share of ad revenue is well-deserved. Moreover, major publishers are misusing the intermediaries' platforms to spread significant amounts of clickbait [PKSH16]. Publishers would maybe not mind laws that regulate information systems to only refer users instead of informing them. This, however, would not be in the best interest of the information society, which desperately needs strong(er) retrieval technology.

Given the significant advances in text generation as of recent, we believe that future information systems will not present information as provided by its sources, anymore, but tailor them to a user's information need. Regulating verbatim reuse is hence short-sighted: the true societal challenge ahead is the question whether automatically generated paraphrases are copyright protected, especially when the training data used does not include the to-be-paraphrased subject. We are currently taking the first steps towards a proof-of-concept for non-reuse snippet generation technology to demonstrate its viability. Key to our approach is the crowdsourcing of large-scale training data composed of topics, search results, and original snippets. Out of curiosity, we ask our workers about their snippet reading habits, with (un)surprising results; see Table 1.

References

[Bax58] P. B. Baxendale. Machine-Made Index for Technical Literature - An Experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

[BP98] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117, 1998.

[BST10] L. L. Bando, F. Scholer, and A. Turpin. Constructing Query-biased Summaries: A Comparison of Human and System Generated Snippets. In *Proc. of IICS*, p. 195–204, 2010.

[CAR16] S. Chopra, M. Auli, and A. M. Rush. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proc. of NAACL/HLT*, 2016.

[CG07] E. Cutrell and Z. Guan. What are you Looking for?: An Eye-tracking Study of Information Usage in Web Search. In *Proc. of CHI*, p. 407–416, 2007.

[GG17] M. Gambhir and V. Gupta. Recent Automatic Text Summarization Techniques: A Survey. *Artificial Intelligence Review*, 47(1):1–66, 2017.

[GJG04] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking Analysis of User Behavior in WWW Search. In *Proc. of SIGIR*, p. 478–479, 2004.

[HLC08] Y. Huang, Z. Liu, and Y. Chen. Query biased Snippet Generation in XML Search. In *Proc. of SIGMOD*, p. 315–326, 2008.

[KHL08] M. Kaisser, M.A. Hearst, and J.B. Lowe. Improving Search Results Quality by Customizing Summary Lengths. In *Proc. of ACL*, p. 701–709, 2008.

[KTS$^+$17] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. What Snippet Size is Needed in Mobile Web Search? In *Proc. of CHIIR 2017*, 2017.

[Luh58] H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[MAM17] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proc. of SIGIR*, p. 135–144, 2017.

[MJH17] C. McMahon, I. Johnson, and B. Hecht. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Proc. of ICWSM*, 2017.

[NZN$^+$16] R. Nallapati, B. Zhou, C. Nogueira dos Santos, Ç. Gülçehre, and B. Xiang. Abstractive Text Summarization using Sequence-to-Sequence RNNs and Beyond. In *Proc. CoNLL*, 2016.

[PKSH16] M. Potthast, S. Köpsel, B. Stein, and M. Hagen. Clickbait Detection. In *Proc of ECIR*, 2016.

[PWTY08] T. Penin, H. Wang, T. Tran, and Y. Yu. Snippet Generation for Semantic Web Search Engines. In *Proc. of ASWC*, p. 493–507, 2008.

[RCW15] A.M. Rush, S. Chopra, and J. Weston. A Neural Attention Model for Abstractive Sentence Summarization. In *Proc. of EMNLP*, 2015.

[SBL11] D. Savenkov, P. Braslavski, and M. Lebedev. Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions. In *Proc. of CLEF*, 2011.

[SCCP09] B. Suh, G. Convertino, E.H. Chi, and P. Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Proc. of WikiSym*, 2009.

[SLM17] A. See, P.J. Liu, and C.D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *Proc. of ACL*, 2017.

[TLKV13] S. Thomaidou, I. Lourentzou, P. Katsivelis-Perakis, and M. Vazirgiannis. Automated Snippet Generation for Online Advertising. In *Proc. of CIKM*, 2013.

[TS98] A. Tombros and M. Sanderson. Advantages of Query Biased Summaries in Information Retrieval. In *Proc. of SIGIR*, p. 2–10, 1998.

[TTHW07] A. Turpin, Y. Tsegay, D. Hawking, and H.E. Williams. Fast Generation of Result Snippets in Web Search. In *Proc. of SIGIR*, p. 127–134, 2007.

[WRJ02a] R. White, I. Ruthven, and J.M. Jose. Finding Relevant Documents Using Top Ranking Sentences: An Evaluation of Two Alternative Schemes. In *Proc. of SIGIR*, p. 57–64, 2002.

[WRJ02b] R. White, I. Ruthven, and J.M. Jose. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Proc. of ECIR*, p. 93–109, 2002.