A Comparison of Snippet Generation Techniques

Vikram Sunil Bajaj (vsb259)

December 23, 2018

Abstract

A snippet is a short piece of text that describes a search result. It is a very important part of search engine results, because a snippet may determine whether or not a user clicks on a certain search result. This report discusses a few snippet generation techniques, both query-independent (static) and query-dependent (dynamic). Snippets are then evaluated based on semantic similarity with the snippets generated by Google for the same search results.

1 Introduction

A snippet, in the context of a search engine, is a set of 2-3 lines that describe a search result. Snippets must ideally be readable (complete sentences) and representative of the document's content, either the entire content or the part relevant to the query. A snippet may be **static** i.e. *query-independent* or **dynamic** i.e. *query-dependent*.

Query-independent snippets, as the name suggests, will be the same for any query. This is why they are also called static snippets. Query-dependent snippets, on the other hand, will change based on the query entered by the user. This is why they are also called dynamic.

The following sections discuss query-independent and query-dependent snippets in more detail, as well as some techniques to generate them. The generated snippets are then compared to those generated by Google for the same query search results. The comparison is both manual and empirical. A metric called **Word Mover's Distance** is used to evaluate the semantic similarity of the generated snippets with those generated by Google.

2 Snippet Evaluation

Snippet Evaluation is difficult. There is no established standard to evaluate the *goodness* of a snippet. However, in this report, the generated snippets are compared to those generated by Google for the same query search results. The aim is to estimate the *semantic* similarity. This refers to checking how similar two snippets are, based on not just the words they contain, but the meanings they convey.

Text summarizations are sometimes evaluated using BLEU [1] and/or ROUGE [2]. However, it has been demonstrated [3] that Word Mover's Distance (WMD) [4] is a better metric to compare text summarizations in terms of semantic similarity.

Simply put, BLEU measures *precision*, while comparing a machine generated summary with a human reference summary. It measures how much the words (and/or n-grams) in the machine-generated summaries appeared in the human reference summaries. ROUGE, on the other hand, measures *recall*: how much the words (and/or n-grams) in the human reference summaries appeared in the machine-generated summaries. This means that if there are many words from the machine-generated results appearing in the human references, it will have high BLEU, and if there are many words from the human references appearing in the machine-generated results, it will have high ROUGE.

However, two summaries may be semantically similar even if they do not contain the same words! On the contrary, two summaries that contain the same words need not be semantically similar! Approaches like BLEU and ROUGE fail to address this issue. WMD assesses the *distance* between two documents even of they have no words in common. Therefore, I have used WMD as a metric to evaluate the semantic similarity between the generated snippets and Google's snippets.

Word Mover's Distance is inspired by Earth Mover's Distance [5] and computes *travel costs* based on the word2vec [6] embeddings of the words in the documents being compared. In all the approaches below, the word2vec vectors are first normalized, so that they are of equal length, before the WMD can be computed. WMD uses Euclidean Distance to compute the distance between the normalized vectors.

WMD is used as a metric, assuming that Google's snippets are the gold standard. However, in some cases, a generated snippet could indeed be better than Google's snippet. Therefore, snippets have been displayed side-by-side for manual comparison, along with the corresponding WMDs. Note that all the generated snippets have at most 20 words, across approaches. **Note**: A lower WMD indicates higher semantic similarity.

3 Query-Independent Snippet Generation

As discussed, query-independent snippet generation is static and can be generated in advance i.e. before the query processing phase. It usually deals with generating a summary of the result page. There are two types of summarization techniques:

- **Extractive Summarization**: This deals with extracting sentences from the page text. Sentences are ranked based on their relevance to the document content; this is done using tf-idf, idf.. (for example sentences that contain frequent terms are scored higher, but longer sentences are penalized)
- **Abstractive Summarization**: This deals with *generating* sentences, instead of simply extracting sentences from the page. New sentences are usually generated by extracting and compressing sentences from the document; this is similar to what humans do intuitively, but is more difficult to evaluate.

In this report, extractive text summarization has been implemented, for query-independent snippet generation. Abstractive text summarization is not discussed because it is an expensive task, both in terms of time and computational resources, and may not work well. As a result, it may not be feasible for generating snippets.

3.1 Approach 1: Extracting Part of the Page

In this approach, the snippet for a search result page consists of one of the following:

- the first 20 words of the meta description (from the <meta> tag), if available, OR
- the first 20 words of the first paragraph in the page, that has at least 5 words, OR
- the first 20 words of the page text, if no <p> tags were found

The table below shows the result URLs, Google's snippets, the generated snippets (for the same result URLs) and the corresponding Word Mover's distances, for the query 'new york'.

URL	Google's Snippet	Generated Snippet	WMD
https://en.wikipedia.org/wiki/New_York_City	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 ...	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous ...	0.29
https://www.nycgo.com/	Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.	Find out what to do, where to go, where to stay and what to eat in NYC from the experts ...	0.21
https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html	The company hired a former aide to Mr. Trump's new attorney general, ... the company also relied on Mr. Schumer, the New York senator and ...	Russian meddling, data sharing, hate speech — the social network faced one scandal after another. This is how Mark Zuckerberg ...	1.14
https://www.ny.gov/	The official website of the State of New York. Find information about state government agencies and learn more about our programs and services.	The official website of the State of New York. Find information about state government agencies and learn more about our ...	0.07
https://www.theatlantic.com/ideas/archive/2018/10/new-york-retail-vacancy/572911/	New York's empty storefronts are a dark omen for the future of cities.	New York's empty storefronts are a dark omen for the future of cities. ...	0.0

3.2 Approach 2: Page Summarization

This approach uses the TextRank [7] summarizer from sumy (a Python package) to summarize the page. It then extracts the first 20 words of the summary. This is the snippet. The results for the query 'new york' are shown below.

URL	Google's Snippet	Generated Snippet	WMD
`https://en.wikipedia.org/wiki/New_York_City`	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 ...	The New York Times . The New York Times . The New York Times . The New York Times	0.82
`https://www.nycgo.com/`	Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.	People all around the world know the Bronx as the home of the New York Yankees, the Bronx Zoo and ...	1.01
`https://www.ny.gov/`	The official website of the State of New York. Find information about state government agencies and learn more about our programs and services.	Governor Cuomo unveiled his "2019 Justice Agenda," urging action in the first 100 days of the next legislative session. The ...	1.08
`https://www.timeout.com/newyork`	Your ultimate guide to New York for tourists and locals alike. Discover superb restaurants, amazing bars, great things to do and cool events in NYC.	Things to do The 101 best things to do in New York City Experience the best things to do in ...	0.82

Clearly, Approach 1 seems to have performed better than Approach 2. The snippets generated by Approach 1 are almost the same as those generated by Google. However, those generated by Approach 2 are quite different. This, however, doesn't mean that the snippets generated by Approach 2 are wrong; they're just different.

4 Query-Dependent Snippet Generation

Query-dependent snippet generation is dynamic. The content of the snippet is based on the query terms. Therefore, these snippets cannot be generated in advance and must be generated during the query processing phase.

A naive approach is to select sentences containing query terms and concatenate sub-strings containing a couple of words on either side of the query terms. This, however, will not necessarily create readable sentences/phrases, and is therefore not preferred. Instead, the below approaches have been implemented.

4.1 Approach 3: Extracting Parts of the Page Having Query Terms

First, the page text is extracted. Then, preference is given to sentences that contain all the query terms. After that, sentences that contain at least 1 query term are used. The results for the query 'new york' are shown below.

URL	Google's Snippet	Generated Snippet	WMD
`https://en.wikipedia.org/wiki/New_York_City`	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 ...	City of New YorkNickname(s): Interactive map outlining New York City Location within the State of New York Show map of ...	0.76
`https://www.nycgo.com/`	Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.	People all around the world know the Bronx as the home of the New York Yankees, the Bronx Zoo and ...	1.01
`https://www.nyc.gov/`	The official website of the City of New York. Find information about important alerts, 311 services, news, programs, events, government employment, the office of ...	The official website of the City of New York. Find information about important alerts, 311 services, news, programs, events, government ...	0.16

4.2 Approach 4: Extracting Parts of the Page Having Query Terms and Synonyms

This approach is similar to the previous one but also looks for sentences that contain at least 1 query term synonym. Synonyms of query terms are obtained using WordNet. The results for the query 'new york' are shown below.

URL	Google's Snippet	Generated Snippet	WMD
`https://en.wikipedia.org/wiki/New_York_City`	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 ...	City of New YorkNickname(s): Interactive map outlining New York City Location within the State of New York Show map of ...	0.76
`https://www.ny.gov/`	The official website of the State of New York. Find information about state government agencies and learn more about our programs and services.	The $15 minimum wage goes into effect on December 31, 2018 for workers in New York City employed by businesses ...	1.09
`https://www.timeout.com/newyork`	Your ultimate guide to New York for tourists and locals alike. Discover superb restaurants, amazing bars, great things to do and cool events in NYC.	Announcing the chefs for Time Out Market New York icon-chevron-right Things to do Christmas in New York The most wonderful ...	0.81
`https://www.theatlantic.com/ideas/archive/2018/10/new-york-re\tail-vacancy/572911/`	New York's empty storefronts are a dark omen for the future of cities.	New York's empty storefronts are a dark omen for the future of cities. It's typically simplistic to point at a ...	0.39

The results are the same as the previous approach. However, this depends on the query to a certain extent.

4.3 Approach 5: Query-Based Summarization

This approach gets the page summary using sumy and then extracts sentences from the summary that contain query terms or their synonyms. The results for the query 'new york' are shown below.

URL	Google's Snippet	Generated Snippet	WMD
`https://en.wikipedia.org/wiki/New_York_City`	The City of New York, often called New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2017 ...	The New York Times . The New York Times . The New York Times . The New York Times	0.82
`https://www.nycgo.com/`	Find out what to do, where to go, where to stay and what to eat in NYC from the experts who know it best.	People all around the world know the Bronx as the home of the New York Yankees, the Bronx Zoo and ...	1.01
`https://www.timeout.com/newyork`	Your ultimate guide to New York for tourists and locals alike. Discover superb restaurants, amazing bars, great things to do and cool events in NYC.	Things to do The 101 best things to do in New York City Experience the best things to do in ...	0.82
`https://www.lonelyplanet.com/usa/new-york-city`	Explore New York City holidays and discover the best time and places to visit. — Epicenter of the arts. Dining and shopping capital. Trendsetter. New York City ...	Forget Times Square and the Statue of Liberty – if you want to see the real New York, you need ...	1.01

Having implemented several query-independent and query-dependent snippet generation techniques, it can be seen that the query-independent ones are most similar to those generated by Google. However, this does not mean that the query-dependent ones are not *good*. Upon subjective evaluation, it is clear that there are certain query-dependent snippets that are more informative than Google's snippets.

5 Python Packages Used

- **requests**: to send GET requests to result pages
- **bs4**: BeautifulSoup (to scrape page content)
- **serpscrap**: to retrieve Google results and snippets
- **re**: regex matching
- **gensim**: to load the Google News pre-trained word2vec model for computing WMD
- **pyemd**: used internally while computing WMD
- **sumy**: page summarization using TextRank
- **nltk**: to get synonyms from WordNet

6 Conclusion

This report discussed several query-independent (static) and query-dependent (dynamic) snippet generation strategies/techniques. The generated queries were evaluated using both semantic similarity estimation as well as a manual subjective evaluation, against snippets generated by Google. Word Mover's Distance was used as a metric for semantic similarity estimation.

It was determined that the query-independent snippets were most similar to those generated by Google. However, there were a few query-dependent snippets that were more informative than those generated by Google, although the corresponding Word Mover's Distances indicated low semantic similarity.

A metric that evaluates the goodness of a snippet based on the information-need of the user would be helpful. This is, however, difficult to achieve, since it is highly subjective in nature: it may differ from user to user.

References

[1] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[2] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text Summarization Branches Out (2004).

[3] Kilickaya, Mert, et al. "Re-evaluating automatic metrics for image captioning." arXiv preprint arXiv:1612.07600 (2016).

[4] Kusner, Matt, et al. "From word embeddings to document distances." International Conference on Machine Learning. 2015.

[5] Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. "The earth mover's distance as a metric for image retrieval." International journal of computer vision 40.2 (2000): 99-121.

[6] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[7] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.