Automatic generation of in-text hyperlinks in web publishing

Tomaž Šolc
Zemanta Ltd
tomaz@zemanta.com

May 2008

Abstract

We present a method for automatic generation of in-text explanatory hyperlinks for use in web publishing. A system using this method is currently in production as part of a service for enriching plain-text content. We recognize the importance of link anchors in practical use of such systems, therefore the method is centered around link anchor selection and uses semantic similarity only to resolve ambiguities in the language. We use English Wikipedia as the training set which allows us to capture the current cultural knowledge. Using structured information extracted from Wikipedia we can provide explanatory links to articles in Wikipedia, book and movie databases and other pages on the Internet.

1 Introduction

1.1 Motivation

Readers of news articles on the World Wide Web have become accustomed to using hyperlinks that accompany an article to find new content that may be of interest to them.

Traditionally hyperlinks to related content are created manually, requiring relatively large amount of effort from the content author. Aside from time lost, manual link construction can also be problematic due to inconsistency of human authors [1]. Links that an author may consider as the most relevant for his article may not also be the most convenient links for his readers. This creates a need for a link generation process that tries to take the knowledge and interests of the average reader into account.

1.2 Link classification

There are a number of possible semantic classifications of hyperlinks [2]. In the context of web publishing our experience shows that we can classify most links into three classes based on the semantic relation between the source and destination documents:

- *Explanation links*: Links that provide more in-depth explanation of terms or concepts the reader may not be familiar with.
- *Reference links*: Links that connect the current article to articles from which this article sourced some of its content.
- *Related links*: Links to content that may also be of interest to the reader, but has no direct connection to the current article. For example articles with the same topic or the same author.

We can also classify the links based on the position of its anchor text:

- *In-text links*: such links are provided within the main page content. Single words or short phrases in the body text are used as link anchors.
- *Supplementary links*: links that are provided outside of the main body of text. For example in a side bar, floating body or in the content's footer.

Our experience shows that in-text links, if done properly, are more convenient for users than supplementary links. Reasons for this seem obvious: in-text links do not require the reader to shift her attention from the body of the text and can be followed as soon as an unfamiliar term is encountered

during reading. On the other hand supplementary links require more effort to locate them on the page.

This observation suggests that in-text should be the preferred way of placing explanation and reference links. On the other hand related links are better suited as supplementary links since their role is to guide user to another article after reading through the content.

1.3 Goal

Our goal is to create an automatic in-text hyperlink generation system that performs well enough to be usable in practice for personal and semi-professional web publishing.

We have so far focused on finding in-text explanation links, since that that task appears to be easier than other types of links. Specifically anchors for in-text reference links usually employ longer phrases that are harder to detect and require higher levels of natural language understanding.

A system using the method described in this paper is currently in production as explanatory link generation part of a service that automatically augments plain text content with in-text explanatory hyperlinks, supplementary links to related articles, relevant images and tags[1].

1.4 Similar research

Most of the research into automatic link generation focuses on the task of finding semantic connections within a known set of documents [3] [4]. On the other hand the problem of practical placement of such links within the website (i.e. finding an anchor) remains relatively unknown.

Especially with in-text links the choice of the anchor is of the highest importance, since the user can only deduce its destination from the anchor. Since the link is inserted into the main body of text there can usually be no additional explanation of the link destination. Choice of the anchor is further complicated by the fact that anchor can only be chosen from the limited set of words or phrases that already appear in body text.

In the case of web publishing links also aren't only formed within a well defined set of previously known documents. In practice most links will point to a third-party website, so instead of a symmetric problem of interlinking a set of documents we have to provide links from one document to potentially millions web sites available.

[1] `http://www.zemanta.com/demo`

Entity	Alias	Source
Computer	Computer	Title
	General purpose computer	Anchor
	Machine	Anchor
	Host	Anchor
	Box	Disambiguation
	Rig	Disambiguation
	Computor	Redirect

Table 1: Selection of aliases for entity *computer*

The only project with a similar approach we have found is the automatic link generation part of Kylin [5], which is also based on a data set, extracted from Wikipedia. However our system is able to use any word as an anchor, not only proper nouns and is not limited to links between Wikipedia articles.

2 Link generation

2.1 Overview

At the core of our system is a collection of semantic entities that represent concepts that may in some situation require an in-text explanation link. Each of these entities is associated with a body of text that can serve as a basis for similarity calculation, a list of aliases and a collection of URLs that provide possible destinations for explanatory links.

Aliases are words or phrases that can be used in a text to refer to the concept represented by the entity. Table 1 shows a simple example of entity "computer" and its aliases. It should be noted, that many entities may share a single alias due polysemy and other language ambiguities. Aliases are also associated with metadata, such as their proper capitalization, part of speech and named entity status.

During the construction of explanatory in-text links, we take entity aliases as candidate anchors for links. Using a multistring matching algorithm we can generate a list of all possible in-text links for a body of text, with each link supported by one or more possible anchors at some position in the text. From this list of candidates we then select those links that are relevant for this particular text, disambiguating any links that share the same

anchors.

This selection step is done by assigning a number of numeric features to each candidate link (lf_i) and to each anchor (af_i). Based on these features we then calculate an aggregate confidence value for each link.

$$AC_i = f_a(af_{1i}, af_{2i}, af_{3i}, af_{4i}) \quad (1)$$

$$LC_j = f_l(lf_{1i}, lf_{2j}, AC_n \dots AC_m) \quad (2)$$

Here AC_i is confidence of ith anchor, LC_j is confidence of jth link and $AC_n \dots AC_m$ are anchor confidences of anchors supporting jth link.

Candidate links are sorted by their confidence values and the best N are selected. In the case of multiple candidate anchors for one selected link, the final anchor is again chosen based on its anchor confidence.

We also perform an anchor conflict resolution step - if selected links have any overlapping pairs of anchors, the link with a lower confidence is discarded.

The number of selected links N is chosen on the basis of article length. We have chosen a simple linear relation of N with the number of tokens in the input text with a maximum of 10.

2.2 Training data

We have chosen Wikipedia as our basic training set for our algorithm. Wikipedia is a free, collaboratively edited encyclopedia. As of writing this article the English version contains around 2.300.000 articles, covering a multitude of topics. Because of the open nature of its content (most pages can be edited by anyone visiting its web page), there is a notable bias in its content towards topics that are currently present in the popular culture. This feature has often been a source of criticism. However since the text we are analyzing with this data set also has a bias towards these topics this feature is actually a benefit for us.

lf_1	TF/IDF similarity between article and entity description
lf_2	Normalized count of different possible anchors found for this entity
af_1	Capitalization correctness
af_2	Length of the anchor
af_3	Comparison between named entity extraction and selected anchor
af_4	Comparison between part of speech and selected anchor

Table 2: Features used for calculating link and anchor confidence

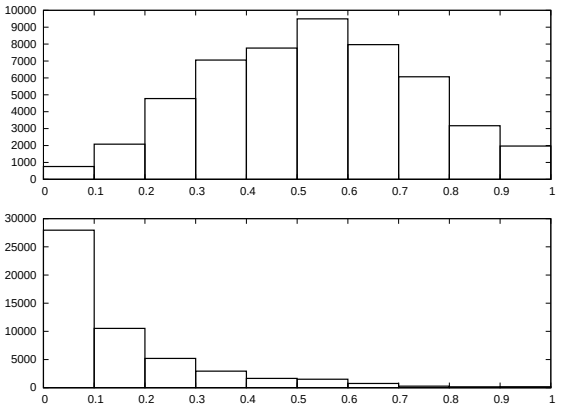

Figure 1: Precision (top) and recall (bottom)

We have used Wikiprep preprocessor [6] to process data dumps provided by WikiMedia foundation.

For our purposes we have created a one-to-one mapping between a subset of English Wikipedia articles and semantic entities so that each Wikipedia article provided body text for one entity (we have discarded category, disambiguation and similar pages that do not provide descriptions of a single semantic entity).

Since Wikipedia articles are themselves composed of hypertext, we have also used this data for creating entity aliases. Entity aliases were created from anchors of existing links in Wikipedia, article titles and redirects and terms and phrases extracted from Wikipedia's disambiguation pages.

This process has yielded approximately 14.000.000 aliases. Due to the size of the training set there was no need to use lemmatisation on the input text, since most of the aliases are present in all their common inflections. Table 1 also shows that this way of gathering aliases provides our system with a great deal of cultural knowledge. "rig", "box", "host" and "machine" are all common ways of referring to "computer" in texts published on the Internet. Aliases also include common typographical and spelling errors such as "computor", which means that our system is also resistant to some degree to such errors in the input text.

In addition to internal links to other Wikipedia pages, each Wikipedia page also provides a number of external hyperlinks to pages connected to its topic. These can be used to enrich the set of pages we link to. For most entities the Wikipedia page itself is a fitting destination of an explanation link our algorithm inserts into text (Wikipedia is one of the most linked to pages on the Internet). For some types of entities however we use these external URLs instead. For example entities that correspond to companies will link to company official web site. Another example is entities that correspond to books or movies. In that case a page in a book or movie database provides a better destination.

3 Results

We have tested the performance of our system by running it on a sample of 100.000 English Wikipedia articles. To calculate precision and recall we have compared automatically generated links to other articles in Wikipedia with those created manually by article authors.

For the purposes of the first test, a link was considered correct if the original article had the same internal link with the same string as the anchor (but possibly at another position in text).

Using this criterion we achieved an average precision of 50% (figure 1).

If we relaxed the condition to allow that the generated anchor is only a part of the manually created one or vice versa, the average precision rose to 63%.

When disregarding anchor selection and only checking for correct link source and destination, the average precision was 74%.

9% of links were wrongly disambiguated (i. e. same anchor was chosen as the manually created link, but with different destination page).

Results also show that poor recall is achieved for majority of articles. This is caused by the limit of maximum 10 in-text links per article we have imposed on the algorithm - we have optimized the system for publishing where the desired density of in-text links is much lower than in Wikipedia's encyclopedic content. Therefore we have given more attention to increasing precision than recall.

While these results do give a general overview of method's performance, it should be noted that proper assessment would require a test data set that is different from the training set. Testing the system against a set of news articles with manually constructed hyperlinks is planned for the future.

4 Conclusion

Our system gives adequate precision in practical use, however user interaction is still needed to confirm the generated links before insertion in the text.

The practical nature of this project has prevented us from using more processing intensive methods for finding semantic relation. This may be attempted in the future as more processing power becomes available.

Further research is needed for the mechanism of selecting the links once confidences have been assigned. Linear dependence of the number of generated links and the number of tokens in the input text has proved to be problematic, since different texts contain different densities of concepts that need explanation.

Linking to biographical pages also proved to be problematic, since this type of explanatory links are hard to disambiguate. One particular problem is how to recognize that a text contains a mention of a person that does not have a biographical page in the Wikipedia (and so doesn't have an entity in our system), but shares the same name with some other person that does have a biographical page.

References

[1] S. J. Green: *Automatically generating hypertext by computing semantic similarity.* University of Toronto, 1997.

[2] J. Allan: *Automatic hypertext link typing.* University of Massachusetts, 1996.

[3] S. J. Green: *Building hypertext links in newspaper articles using semantic similarity.* University of Toronto, 1999.

[4] M. Agosti, et al.: *On the use of information retrieval techniques for the automatic construction of hypertext.* Universita di Padova, 1997.

[5] F. Wu, D. S. Weld: *Autonomously Semantifying Wikipedia.* University of Washington, 2007.

[6] E. Gabrilovich: *WikiPrep.* 2007. http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/