Using the Wikipedia Link Structure to Correct the Wikipedia Link Structure

Benjamin Mark Pateman
University of Kent
England
bmp7@kent.ac.uk

Colin Johnson
University of Kent
England
C.G.Johnson@kent.ac.uk

Abstract

One of the valuable features of any collaboratively constructed semantic resource (CSR) is its ability to – as a system – continuously correct itself. Wikipedia is an excellent example of such a process, with vandalism and misinformation being removed or reverted in astonishing time by a coalition of human editors and machine bots. However, some errors are harder to spot than others, a problem which can lead to persistent unchecked errors, particularly on more obscure, less viewed article pages. In this paper we discuss the problems of incorrect link targets in Wikipedia, and propose a method of automatically highlighting and correcting them using only the semantic information found in this encyclopaedia's link structure.

1 Introduction

Wikipedia, despite initial scepticism, is an incredibly robust semantic resource. Armed with a shared set of standards, legions of volunteers make positive changes to the pages of this vast encyclopaedia every day. Some of these editors may be casual – perhaps noticing an error in a page they were reading and being motivated to correct it – while others actively seek to improve the quality of a wide variety of pages that interest them. Facilitated by a relatively minimalist set of editing mechanics and incentives, Wikipedia has reached a state in which it is, for the most part, a reliable and stable encyclopaedia. Just enough regulation to prevent widespread vandalism or inaccuracy (including, on occasion, the temporary locking of particularly controversial pages), and enough editing freedom to maintain accuracy and relevance.

There are a number of potential approaches to minimizing misinformation and vandalism, falling into two broad categories: adding human incentives, and creating Wiki-crawling bots. There already exists a wide variety of natural and Wiki-based incentives (Kuznetsov, 2006) that have been crucial to the encyclopaedia's success. By implementing additional incentives, it may be possible to, for example, increase editor coverage of less-viewed articles. There are many avenues to explore regarding this, from additional community features such as a reputation system (Adler and de Alfaro, 2007), to ideas building upon recent work relating to games with a purpose (von Ahn, 2006), providing a form of entertainment that simultaneously aids page maintenance.

Wikipedia also benefits from a wide variety of bots and user-assistance tools. Some make the lives of dedicated editors easier (such as WikiCleaner[1]), providing an interface that facilitates the detection and correction of errors. Others carry out repetitive but important tasks, such as ClueBot[2], an anti-vandalism bot that reverts various acts of vandalism with surprising speed. Similar bots have been of great use in not only maintaining existing pages but also in adding new content (such as RamBot[3], a bot responsible for creating approximately 30,000 U.S city articles).

In recent years, researchers have taken an increasing interest in harnessing the semantic data contained in Wikipedia (Medelyan et al., 2009). To this end, the encyclopaedia now serves as not only a quick-lookup source for millions of people across the world, but also as an important semantic resource for a wide range of information retrieval, natural language processing and ontology building applications. With all this utility, it is increasingly beneficial for Wikipedia to be as accurate and reliable as possible.

In this paper, we will discuss an algorithm that aims to use Wikipedia's inherent link structure to detect and correct errors within that very same

[1] https://launchpad.net/wikicleaner
[2] http://en.wikipedia.org/wiki/User:ClueBot
[3] http://en.wikipedia.org/wiki/User:Rambot

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 10–18,

structure. In Section 2 we will explore the nature and causes of this error, outlining the motivations for our algorithm. Section 3 discusses the inspirations for our approach, as well as our reasons for choosing it. We will then describe its method in detail, before evaluating its effectiveness and analysing its strengths and weaknesses.

2 A Reliable Encyclopaedia

"It's the blind leading the blind – infinite monkeys providing infinite information for infinite readers, perpetuating the cycle of misinformation and ignorance" (Keen, 2007). There has been much debate over the value of Wikipedia as a reliable encyclopaedia. Fallis (2008) talks at length about its epistemic consequences, acknowledging these criticisms but ultimately reaching a positive conclusion. In particular, he emphasizes the merits of Wikipedia in comparison with other easily accessible knowledge sources: If Wikipedia did not exist, people would turn to a selection of alternatives for quick-lookups, the collection of which are likely to be much less consistent, less verifiable and less correctable.

The fallacies of Wikipedia come from two sources: disinformation (an attempt to deceive or mislead) and misinformation (an honest mistake made by an editor). These can exist both in the textual content of an article, as well as the structural form of the encyclopaedia as a whole (e.g. the link structure or category hierarchy). The consequences can be measured in terms of the lifespan of such errors: a fairly harmless issue would be one that can be noticed and corrected easily, while those that are harder to detect and correct must be considered more troublesome.

For this reason, to be more potent on less frequently visited pages, as mentioned in Section 1. However, (Fallis, 2008) argues that "because they do not get a lot of readers, the potential epistemic cost of errors in these entries is correspondingly lower as well", suggesting that a balance is struck between misinformation and page traffic that stays somewhat consistent across all traffic levels. While inaccuracies may linger for longer on these less visited pages, it follows that fewer people are at risk of assuming false beliefs as a result.

An interesting pitfall of Wikipedia pointed out by Fallis (2008) comes as a result of the nature of its correctability. As readers of any piece of written information, certain factors can make us less trustworthy of its content; for example, grammatical or spelling mistakes, as well as blatant falsehoods. However, these are the first things to be corrected by Wikipedia editors, leaving what appears to be – on the surface – a credible article, but potentially one that embodies subtle misinformation that was not so quickly rectified.

2.1 Ambiguous Disambiguations

It is therefore important that methods of detecting and resolving the not-so-obvious inaccuracies are developed. One such not-so-obvious error can occur in Wikipedia's link structure. This problem stems from the polysemous nature of language (that is, that one word can map to multiple different meanings). In Wikipedia, different meanings of a word are typically identified by adding additional information in the relevant page's name. For example, the article "*Pluto (Disney)*" distinguishes itself from the article "*Pluto*" to avoid confusion between the *Disney* character and the dwarf planet. Adding extra information in brackets after the article name itself is Wikipedia's standard for explicitly disambiguating a word. Note that the article on the dwarf planet *Pluto* has no explicit disambiguation, because it is seen as the primary topic for this word. In other cases, no primary topic is assumed, and the default page for the word will instead lead directly to the disambiguation page (for example, see the Wikipedia page on "*Example*").

This system, while effective, is susceptible to human error when links are added or modified. The format for a link in WikiText is: "`[[PageName | AnchorText]]`" (the anchor text being optional). It is not hard to imagine, therefore, how a slightly careless editor might attempt to link to the article on *Pluto* (the *Disney* character) by typing "`[[Pluto]]`", assuming that this will link to the correct article, and not something completely different.

Is "*Jaguar*", generally the name of a fast feline, more likely to make you think of cars? "*Python*" is a genus of snake, but also a programming language to those involved in software development. Apple, a common fruit, but to a lot of people will be heavily associated with a well-known multinational corporation. These examples suggest that when a word takes on a new meaning, this new meaning – as long as it remains relevant – can be-

come more recognizable than the original one (as yet another example, consider how your reaction to the word "*Avatar*" fluctuated in meaning as James Cameron's film went by). One particular potential problem is that someone editing an article will be focused on the context of that particular article, and will therefore be likely to not consider the polysemous nature of a word that they are using. For example, someone editing the article on the Apple *iPad* will have the company name Apple prominently in their mind, and therefore may momentarily forget about the existence of a particular kind of small round fruit.

The effects of these blunders can vary greatly depending on the word in question. For example, just about anyone who – expecting to be directed to a page on a *Disney* character – instead finds themselves at a page about a well-known dwarf planet in our Solar System, is going to know that there is an error in the source article. In this example, then, the error would be fixed very quickly indeed – faster still if the source page was popular (such as the article on *Disney* itself). However, there are cases where linking to the wrong sense of a polysemous word may not be as obvious an error for a lot of users. Someone following a link to "*Jagúar*" (the band) is less likely to notice a mistake if they're taken to the incorrect page of "*Jaguar (band)*" (a different band) than if they're taken to the incorrect page "*Jaguar*" (the feline). We argue that the extent of this problem depends on the difficulty of distinguishing between two different meanings of the same word. This difficulty is based upon two factors: the reader's level of background knowledge about the expected article, and the semantic similarity between it and the incorrect article being linked to. If the reader has absolutely no knowledge concerning the subject in question, they cannot be certain that they are viewing the correct page without further investigation. Furthermore, a reader with some relevant knowledge may still be unaware that they have been taken to the wrong page if the incorrectly linked-to page is semantically very similar to the page they were expecting. If these are common responses to a particular pair of polysemous articles, then it follows that a link error concerning them is likely to persist for longer without being corrected.

3 The Semantic Significance of Wikipedia's Link Structure

Wikipedia consists of, for the most part, unstructured text. Originally constructed with only the human user in mind, its design makes machine interpretations of its content difficult at best. However, the potential use of Wikipedia in a wide range of computational tasks has driven a strong research effort into ways of enriching and structuring its information to make it more suitable for these purposes. For example, DBpedia[4] takes data from Wikipedia and structures it into a consistent ontology, allowing all its information to be harnessed for various powerful applications, and is facilitating efforts towards realizing a semantic web (Bizer et al., 2009).

At the same time, research has also been carried out in ways of making use of the existing structure of Wikipedia for various natural language processing applications. For example, Shonhofen (2006) proposed using the hierarchical category structure of Wikipedia to categorize text documents. Another example of a system which makes use of word-sense disambiguation in the context of Wikipedia is the *Wikify!* system (Mihalcea and Csomai, 2007), which takes a piece of raw text and adds links to Wikipedia articles for significant terms. One of the biggest challenges for the authors of that system was linking to polysemous terms within the raw text. A combination of methods was used to determine the best disambiguation: overlap between concepts in the neighbourhood of the term and dictionary definitions of the various possible link targets, combined with a machine learning approach based on linguistic features.

In this paper we are concerned with another method of using Wikipedia without prior modifications: exploiting the nature of its network of links. This approach was pioneered by Milne and Witten (2007; 2008a; 2008b), responsible for developing the Wikipedia Link-Based Measure, an original measure of semantic relatedness that uses the unmodified network of links existing within Wikipedia.

Indeed, the link structure is one of the few elements of Wikipedia that can be easily interpreted by a machine without any restructuring. It contains within it informal – often vague – relationships between concepts. Whereas, ideally, we would like to

[4] http://dbpedia.org/

be dealing with labelled relationships, being able to directly analyse collections of untyped relationships is still very useful. Importantly, however, we must not concern ourselves with the significance of a single link (relationship), due to its class being unknown. In an article there may be links that are more significant – semantically speaking – than others, but this information cannot be retrieved directly. For example, the article on a famous singer might have a link to the village in which she grew up, but this is arguably – in most contexts – less semantically significant than the link to her first album, or the genre that describes her music.

Instead, then, we would like to look at collections of links, as these loosely summarize semantic information and de-emphasize the importance of knowing what type of relationship each link, individually, might express. Every single page on Wikipedia can be seen as a collection of links in this way; ignoring the raw, unstructured text within an article, we are still able to determine a great deal about its meaning just by looking at the underlying link structure. In doing this, comparing the similarity of two articles is as simple as comparing the outgoing links that each has. The more outgoing links that are common between the two articles, the more similar we can gauge them to be.

Looking at the links pointing to an article also provides us with additional cheap information. Of particular interest is deriving an estimated "commonness" of a concept by counting the number of links pointing in to it. The Wikipedia Link-Based Measure uses this information to weight each link, giving additional strength to links that have a lower probability of occurring. This accounts for the fact that two articles are less likely to share uncommon links; if they do, then this link overlap accounts for a higher degree of similarity. Conversely, two articles sharing a very common link (such as a page on a country or capital city) should not be considered very similar on that fact alone.

The motivations behind taking this approach for our link checking algorithm come largely from the inexpensive nature of this measure. While a large amount of potential information is ignored – such as the content of an article itself – the computational cost is an order of magnitude lower, and minimal preprocessing is required. With the English Wikipedia consisting of several million pages, and the search for incorrect links being essentially blind, processing speed is an important factor in providing useful page coverage.

4 Detecting Incorrect Links

The detection of incorrectly targeted links in Wikipedia is a trial of semantics; by estimating how similar in meaning a linked page is to the theme of an article, we can determine whether there might be an alternative page that would be more suitable. In finding significantly more suitable alternatives to these semantically unrelated links, we are able to hypothesise that the original link was incorrect. In the following subsections, we will describe the details of this algorithm.

4.1 Preparing the Database

Snapshots of Wikipedia can be downloaded from its database dump page[5], and then loaded into a local database. While this database is used by the algorithm, the practicality of such an application demands that live Wikipedia pages be used as the input. Checking a week old snapshot of Wikipedia for incorrect links will be less effective, as a number of them may well have been already fixed on the website itself. For this reason, the algorithm accepts a URL input of the page to be analysed, and will extract its current links directly.

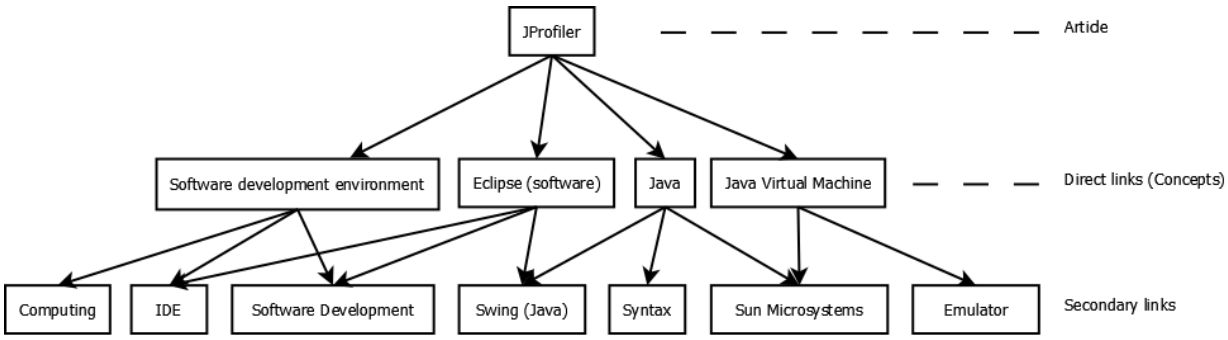
4.2 Determining the Theme of an Article

The first step is to compute the semantic theme of the original article in question. This is done using an algorithm loosely based on that of Milne and Witten (2008a), which was discussed in section 3. To begin with, the original article is arranged as a list of linked pages (pages that it links directly to). Each of these derived pages is considered as a semantic "concept".

We represent each concept as a further list of its outgoing page links, creating a wide tree structure of depth 2, with the original article at the root (see Figure 1). The theme of this article is determined by propagating link information up the tree, essentially consolidating the individual themes of each of its concepts. As new links are discovered, they are assigned a commonness weighting (see section 3), and multiple encounters with the same link are tallied. For each link, this information (the commonness rating and link frequency) is used to sculpt the

[5] http://dumps.wikimedia.org/backup-index.html

Figure 1: A simplified link structure diagram for the article on "*JProfiler*".



overall theme of the article.

4.3 Semantic Contribution

We use the phrase "semantic contribution" to describe how much meaning a particular concept "contributes" to the theme of the article in question. This is based on the nature of each of its links and how frequently they occur amongst the rest of the article's concepts. We therefore quantify the semantic contribution of a given concept by using the formula:

$$S_c = \sum_{l=1}^{n} \begin{cases} \log(f_l) w_l & \text{if } f_l >= 2 \\ 0 & \text{if } f_l < 2 \end{cases}$$

In other words, for each link l with a frequency (f, the number of times this link appears across all concepts) of 2 or more, its semantic contribution is a product of its frequency and its weight (w), as defined by:

$$w = \frac{1}{\log(i_l + 1)}$$

Where i_l is the total number of incoming links (Wikipedia-wide) pointing to the same target as link l. The total semantic contribution of a concept is the summation of all of the contributions of its outgoing links. By quantifying each concept in this manner, we can immediately see which concepts contribute a lot, and which contribute very little, to the theme of an article.

4.4 Extracting Dissimilar Links

With an aggregated theme established for an article, it is a simple task to flag up those concepts that have a low semantic contribution. Due to how semantic information was propagated up the tree (see the previous section), each concept represents some subset of the article's theme. Qualitatively speaking, this essentially equates to looking at how much of its theme overlaps with the most accentuated aspects of the article's theme. The dominant features of an article's theme will come from those links that are uncommon and frequently occurring, so any concept that consists of a good number of these links will be have a high semantic contribution.

By scoring each concept in terms of its contribution to the article theme, we are able to examine those concepts that scored particularly low. The value to use as a threshold for flagging potential errors is somewhat arbitrary, but in our experiments we have found best results using a simple variable threshold:

$$\text{Threshold} = \frac{\text{average contribution}}{2}$$

Any concepts with a semantic contribution below this value are considered as candidate errors, although it's important to note that, in many cases, a perfectly valid link can have a low contribution. For example, a link from a famous film director to a country he once filmed in. In these cases, however, we expect that it is unlikely for a more relevant alternative to be found.

4.5 Finding Possible Alternatives

With one or more potentially incorrect links found, the algorithm must now search for alternative targets that are more suitable. This method is built on the assumption that the link is in error due to pointing towards the wrong disambiguation, accounting for the typical scenario of an editor linking to the wrong sense of a polysemous word.

An editor who has accidentally pointed to the article "*Pluto*" rather than "*Pluto (Disney)*" has not made any spelling errors. As we discussed in Section 2.1, the error is typically a result of a presumption being made on the most typical meaning of

the target article. With this in mind, an error of this nature is likely to be resolved by looking at other articles that share the same name. There are a number of ways to do this, such as simply searching the database for all articles containing the word "*Pluto*". However, we chose instead to locate the relevant disambiguation page, if it exists (in this example, "*Pluto (disambiguation)*"). For the type of error we are targeting, this disambiguation page can be expected to contain the correct, intended page as one of its outgoing links.

4.6 Choosing the Best Alternative

With a list of possible alternatives for a particular weakly related concept, we then go about calculating their potential semantic contribution to the original article (using the same formula as was seen in section 4.4. To continue the example, the semantic contribution of "*Terry Pluto*" is unlikely to be at all high when considering the original article on *Disney*. The same goes for other possible alternatives, such as "*Pluto (newspaper)*" or "*Pluto Airlines*". However, the concept "*Pluto (Disney)*" contributes considerably more than the original link, and this becomes evidence to suggest it as a likely correction.

For each plausible alternative, a score is assigned based on the increased semantic contribution it provides over the original link. By doing this, the suggestions can be ordered from best to worst, expressing a degree of confidence in each option.

5 Evaluation

We evaluated the effectiveness of this algorithm by testing it on a snapshot of Wikipedia from November 2009. By using old Wikipedia pages we can, in most cases, easily validate our results against the now-corrected pages of live Wikipedia. However, finding examples of incorrectly linked articles is no simple task. Indeed, much of the justification for the algorithm this paper describes stems from the fact that finding these incorrect links is not easy, and actively searching for them is a somewhat tedious task. While we would like to leave our script crawling across Wikipedia detecting incorrect links by itself, in order to evaluate its performance we need to evaluate how well it performs on a set of pages that are known to contain broken links. It is impossible to generate such a set automatically, as by their nature these broken links are concerned with the meaning of the text on the pages.

We gauge the performance of our algorithm by looking at how many of the "best" suggestions (those with the highest calculated semantic contribution) given for a particular link are, in fact, correct.

5.1 Gathering Test Data

We found that a satisfactory method for finding incorrect links was to examine the incoming links pointing to a particularly ambiguous page. However, pages can have hundreds or thousands of incoming links, so we need to choose ones that are likely to be linked to in error, using ideas discussed in section 2.1. For example, if we look at the long list of links pointing towards the article "*Jaguar*", we will mostly see articles relating to the animal: geographical locations, ecological information or pages concerning biology, for example. If, among these pages, we notice an out of place page – relating, perhaps, to cars, racing or business – we have reason to believe this article was supposed to be linking to something different (most likely, in this case, "*Jaguar Cars*"). After basic investigation we can confirm this and add it to our collection of pages for evaluation. While still not fast by any means, this method is considerably more effective than randomly meandering around the pages of Wikipedia in search of link errors. For this evaluation, we used the first 50 error-containing pages that were encountered using this method.

Another potentially effective method would be to download two chronologically separate snapshots of the Wikipedia database (for example, one taken a week before the other). We could then compare the incoming links to a particular article across both snapshots: If there are more incoming links in the newer snapshot than the old, then we can attempt to find them in the older snapshot and check their outgoing links. For example, the new snapshot might have a link from the article "*Jim Clark*" to "*Jaguar Cars*" that does not exist in the old snapshot. Upon checking the old snapshot's version of the page on "*Jim Clark*", we see it has a link to "*Jaguar*" and have immediately found a suitable error. This enables us to quickly find links that have been repaired in the time between the two snapshots, providing a fast, systematic method of gathering test data.

Nonetheless, finding a substantial set of examples of incorrectly linked pages is a significant challenge for work in this area. It is an important task, however, as without such a set it is impossible to determine a number of important features of a proposed correction algorithm. Firstly, without such a set it is impossible to determine which wrongly allocated links have been ignored by the algorithm, which is an important measure of the algorithm's success. Secondly, determining whether the algorithm has suggested the correct link requires that these correct links have been specified by a human user. As a result, the development of a substantial database of examples is an important priority for the development of this area of work.

5.2 Discussion

Overall, the results (given in Table 1) show that the algorithm performs well on this test set, with the best suggestion being the correct choice in 76.1% of cases.

As expected, the algorithm works best on larger articles with a well-established theme. For example, the articles on "*Austin Rover Group*" and "*CyberVision*" were riddled with links to incorrect pages, but with a total of 194 and 189 outgoing links respectively, there was sufficient information to confidently and accurately find the most suitable corrections, despite the number of errors. Conversely, "*Video motion analysis*", with only 7 outgoing links, fails to form a strong enough theme to even be able to highlight potential errors.

One might argue that the accurate result for the article on "*Synapse (disambiguation)*" is somewhat of an anomaly. Being a disambiguation page, there is inherently no common theme; typically, each link will point to a completely different area of semantic space. Correctly repairing the link to "*Java*" comes as somewhat of a coincidence, therefore, and it should be noted that disambiguation pages are not suited to this algorithm. Conversely, due to the nature of disambiguation pages, we might assume that users editing them are – in general – more careful about the target of their links, minimizing the occurrence of these sorts of errors.

There is a unique limitation with the algorithm that these results clearly highlight, however. An example of this lies in the results from programming-themed pages dealing with the link to "*Java*": There are a handful of recurring concepts being suggested, such as "*Java (programming language)*", "*Java (software platform)*" or "*Java Virtual Machine*". These suggestions are often accompanied by very similar values of semantic contribution, simply because they are all very semantically related to one another. As a result, if the theme of an article is related to one, it will be typically be related to them all. Which is the correct choice, if all are semantically valid? The one that fits best with the context of the sentence in which it is found.

This reveals an important limitation of this algorithm, in that the position of links within the text – and the surrounding text itself – is completely unknown to it. Dealing only with what is essentially a "bag of links", there is no information to discern which article (from a selection of strongly related articles) would be most appropriate for that particular link to point to. Indeed, in these isolated cases we observed the algorithms accuracy drop to 47%, although it should be noted that in almost all cases the correct link was suggested, just not as the best choice.

6 Conclusion

The results of our evaluation not only display the effectiveness of this algorithm at detecting and correcting typical link errors, but also clearly mark its limitations when dealing with multiple semantically similar suggestions. When considering the impact of these limitations, however, we must not forget that the algorithm was still able to recognize an invalid link, and was still able to offer the correct solution (often as the best choice). The impacts, then, are just on the consistency of the best choice being correct in these situations. However, the aim of this work was to build an algorithm that can be of significant assistance to a human editor's efficiency, and not to replace the editor. With that in mind, the output of the algorithm provides enough information to enable the editor to promptly pick the most appropriate suggestion, based on their own judgment.

While carrying out the evaluation on these 6 month old Wikipedia pages, we checked the results against the live pages of Wikipedia. A surprisingly large number (as many as 40%) of errors found had yet to be corrected half a year later, which, ultimately, is highly indicative of the potential benefits of this utility in repairing the errors that nobody knew existed.

Table 1: Counts of the correct link being given as the best suggestion.

Page Name	Best Correct	Page Name	Best Correct
Acropolis Rally	2/2	JProfiler	0/1
Austin Rover Group	6/6	KJots	0/1
Barabanki district	2/2	Lady G	0/1
Batch file	0/1	List of rapid application development tools	3/3
Belong (band)	1/1	Video motion analysis	0/1
Comparison of audio synthesis environments	3/4	Logo (programming language)	1/3
Comparison of network monitoring systems	2/3	Maria Jotuni	0/1
Computer-assisted translation	0/1	Mickey's delayed date	1/1
Convention over configuration	1/1	Neil Barret (Fashion Designer)	1/1
CyberVision	18/21	Ninja Gaiden (Nintendo Entertainment System)	2/3
Daimler 2.5 & 4.5 litre	1/2	Planetary mass	1/1
Dance music	3/3	Population-based incremental learning	1/1
Deiopea	1/1	Streaming Text Oriented Messaging Protocol	1/2
David Permut	3/3	Spiritual Warfare (video game)	1/2
Demon (video game)	1/1	Sonic Heroes	1/1
Disney dollar	1/1	Soulseek Records	2/2
DJ Hyper	1/1	Synapse (disambiguation)	1/1
DJ Qbert	1/2	Tellurium (software)	2/2
Eliseo Salazar	1/2	Testwell CTC++	1/1
Fixed point combinator	0/1	The Flesh Eaters (band)	3/3
Gravity Crash	1/1	Trans-Am Series	3/4
Hyphenation algorithm	2	Ultima IV: Quest of the Avatar	1/1
IBM Lotus Notes	1/2	Uma Thurman	4/6
Jaguar XFR	2/2	Unlabel	1/1
Jim Clark	0/1	Virtual World	1/2
		Total:	86/113

7 Further Work

In continuing this work, there are a number of avenues to explore. Fundamentally, there is room to fine tune various aspects of the algorithm, such as the threshold value used to determine candidate errors, or the relationship between a link's frequency and its commonness. In doing so we might include additional variables, in particular investigating how the size of an article affects the algorithm, or the distribution of a central theme amongst its concepts.

Additionally, there is work to be done on constructing a practical application from this; adding, for example, an accessible GUI as well as direct Wikipedia integration to allow for users to easily commit corrected links to the Wikipedia server itself. This could lead to a further evaluation step in which we analyse the effectiveness of these corrections after the system has been running "in the wild" for a number of months. In order to use this system to correct the live Wikipedia it would be important to have an up-to-date local copy of Wikipedia in order to rapidly access the up-to-date link structure.

As mentioned earlier, an important challenge for the accurate evaluation of systems of this kind would be the development of a substantial, annotated database of examples of this kind of broken link. Clearly, it is difficult for a single development team to curate such a database, as the discovery process is time consuming. One approach to this would be through some form of crowdsourcing effort to gather a large number of examples. This could be as simple as encouraging readers of Wikipedia to report such corrections, for example by using a specific keyword in the revision notes made on that change. A more sophisticated approach could be to draw on the concept of *games with a purpose* (von Ahn, 2006), as exemplified by the *Google Image Labeler*[6] which uses a two-player game to find new tags for images. A game could be created based on the notion of presenting the user with a choice of links for a particular Wikipedia page, and rewarding them when they agree with another user on a target that is not currently pointed at by that link.

One further useful measure would be to devise a baseline algorithm to compare against. One possibility for this baseline would be to select the most heavily referenced choice from the list of candidates. This is similar to the approach used in data mining, where classifiers are compared against the naive classifier that classifies every instance as the most frequent item in the training set.

Finally, taking the reverse approach to the algorithm and looking primarily at incoming links – following the intuition behind our method of selecting test data (see section 5.1) – may prove very useful in locating articles that potentially contain incorrect links, allowing the algorithm to accurately and efficiently seek out pages to repair without having to crawl blindly across the entire encyclopaedia.

[6] http://images.google.com/imagelabeler/

References

Adler, B. Thomas and Luca de Alfaro. 2007. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA. ACM.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September.

Fallis, Don. 2008. Toward an epistemology of wikipedia. *Journal of the American Society for Information Science and Technology*, 59(10):1662–1674.

Keen, Andrew. 2007. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Broadway Business, June.

Kuznetsov, Stacey. 2006. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2), June.

Medelyan, Olena, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from wikipedia. May.

Mihalcea, Rada and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

Milne, David and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence*.

Milne, David and Ian H. Witten. 2008b. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA. ACM.

Milne, David. 2007. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*.

Schonhofen, Peter. 2006. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA. IEEE Computer Society.

von Ahn, L. 2006. Games with a purpose. *Computer*, 39(6):92–94.