Dec 11th, 12:00 AM

The Currency of Wiki Articles – A Language Model-based Approach

Bernd Heinrich
University of Regensburg, bernd.heinrich@wiwi.uni-regensburg.de

Maximilian Felix Huber
University of Regensburg, maximilian1.huber@stud.uni-regensburg.de

Thomas Krapf
University of Regensburg, thomas.krapf@wiwi.uni-regensburg.de

Alexander Schiller
University of Regensburg, alexander.schiller@wiwi.uni-regensburg.de

The Currency of Wiki Articles – A Language Model-based Approach

Completed Research Paper

Bernd Heinrich
University of Regensburg
Universitätsstr. 31, 93053 Regensburg
Bernd.Heinrich@ur.de

Maximilian Huber
University of Regensburg
Universitätsstr. 31, 93053 Regensburg
Maximilian1.Huber@stud.uni-regensburg.de

Thomas Krapf
University of Regensburg
Universitätsstr. 31, 93053 Regensburg
Thomas.Krapf@ur.de

Alexander Schiller
University of Regensburg
Universitätsstr. 31, 93053 Regensburg
Alexander.Schiller@ur.de

Abstract

Wikis are ubiquitous in organisational and private use and provide a wealth of textual data. Maintaining the currency of this textual data is important and difficult, requiring large manual efforts. Previous approaches from literature provide valuable contributions for assessing the currency of structured data or whole wiki articles but are unsuitable for textual wiki data like single sentences. Thus, we propose a novel approach supporting the assessment and improvement of the currency of textual wiki data in an automated manner. Grounded on a theoretical model, our approach makes use of data retrieved from recently published news articles and a language model to determine the currency of fact-based wiki sentences and suggest possible updates. Our evaluation conducted on 543 sentences from six wiki domains shows that the approach yields promising results with accuracies over 80% and thus is well-suited to support assessment and improvement of the currency of textual wiki data.

Keywords: Data quality, currency, wikis, textual data, unstructured data, language model

Introduction

Wikis provide an easy and efficient way to share, store, reuse, adapt, mobilise, and aggregate information (Beck et al. 2015; Yates et al. 2009). Their popularity, availability, and scope have increased over time, to the point where they are now widely applied across all areas and industries (Bhatti et al. 2018). Beyond organisational application, wikis have also become commonplace in private use (Alqahtani 2017). For instance, Wikipedia can not only be seen as the most comprehensive information repository in human history (Dang and Ignat 2017), but it also is one of the world's most popular websites, with close to 2 billion page visits per month (Wikimedia 2023). The English Wikipedia alone contains more than 6 million articles providing data about a variety of topics such as politics, science, culture, and sports as well as a large number of biographies (Wikipedia 2023).

Such a wealth of data, however, also comes with a drawback: It makes maintaining high data quality (DQ) in wikis, especially in a manual manner, very difficult and costly. This issue is further exacerbated by the fact that large parts of wiki articles consist of natural language text, which DQ maintenance methods have to cope with. DQ can be defined as "the measure of the agreement between the data views presented by an information system and that same data in the real world" (Orr 1998, p. 67; cf. also Heinrich et al. 2009). It is a multidimensional construct (Cichy and Rass 2019; Wang and Strong 1996) comprising several DQ

dimensions such as currency, completeness, accuracy, and consistency (Nelson et al. 2005). In accordance with DQ literature (Heinrich and Klier 2015; Nelson et al. 2005; Zak and Even 2017), we refer to currency in this paper, expressing whether a data view presented by an information system, which was originally accurately captured, still is an accurate representation at the time of assessment. It is related to, but different from other time-related dimensions such as timeliness, which in DQ literature has mainly been defined as being directly dependent on the age of a data value and the maximum length of time the value of the considered attribute remains up-to-date (Ballou et al. 1998; Heinrich et al. 2018). Timeliness in this sense is not applicable to wikis, as there is no fixed and known maximum length of time for which certain data in a wiki can remain up-to-date (for a detailed discussion cf. Heinrich and Klier 2015, pp. 83-84). In contrast, currency is of particular importance for wikis as outdated data makes the wiki less helpful, less credible and leads to detrimental effects such as misinformation and less efficient use (Bhatti et al. 2018; He and Yang 2016; Shah et al. 2015). Moreover, there is a strong correlation between the currency of data contained within a wiki and the frequency of its use, as well as the satisfaction of its users (Bhatti et al. 2018). Thus, we focus on the currency of wikis in this paper. In the context of wikis, currency expresses whether a wiki article (and especially the textual wiki data) still describes the current state of the corresponding entity in the real world at the time of assessment (Klier et al. 2021; Lewoniewski 2019).

As already pointed out, wiki articles usually mainly consist of (unstructured) textual data, sometimes paired with data in structured form (e.g., information boxes) or other unstructured data (e.g., pictures or short videos). While for structured data, many approaches to assess and improve currency have been proposed (e.g., Heinrich and Klier 2015; Zak and Even 2017), there are only a few initial contributions regarding the currency of textual data (e.g., Batini and Scannapieco 2016; Hao et al. 2020). Yet, textual data is the main component of wiki articles and frequently becomes outdated due to events in the real world (Klier et al. 2021). Due to the sheer number of articles and the velocity of required updates, keeping textual wiki data current is highly challenging. Indeed, it has been noted that even Wikipedia, a highly popular wiki with a very large number of voluntary authors, has difficulties keeping articles current (Hatcher-Gallop et al. 2009). Consequently, it is crucial to assess and improve the currency of textual wiki data with automated support (Seibert et al. 2011). Thus, the research question we aim to address in this paper is as follows:

How can the assessment and improvement of the currency of textual wiki data be supported in an automated manner?

The remainder of the paper is organised as follows. In the next section, we discuss the problem context and present a running example. Subsequently, we provide an overview of related prior work. We then develop a novel language model-based approach to assess and improve the currency of textual wiki data. Thereafter, we instantiate our approach using a dataset from Wikipedia and evaluate and discuss its performance. Finally, we conclude, reflect on limitations, and provide an outlook on future research.

Problem Context and Running Example

Textual wiki data usually contains a large amount of factual data which may become outdated over time due to changes in the real world. To illustrate this idea, we consider a sentence about the political position of the prime minister (PM) of the United Kingdom currently held by Rishi Sunak, and use it as a running example throughout this paper (any fact-based sentences from other Wikipedia domains can be used analogously). This is typical data which may be provided by Wikipedia or other wikis containing general information. Sunak's predecessor Liz Truss took office in early September 2022 (point in time t_1 in Figure 1), and this data may have been stored correctly in a wiki shortly after (point in time t_2 in Figure 1). At this time, the data provided by the wiki is current. However, in late October 2022, Sunak succeeded Truss as PM (point in time t_3 in Figure 1). This is a change in the real world, which renders the data previously stored in the wiki outdated. The data provided by the wiki remains outdated until it is correctly updated (point in time t_4 in Figure 1), which restores its currency. The same sequence takes place for all previous changes of office (e.g., May succeeding Cameron, Johnson succeeding May, etc.). Ideally, only a short amount of time elapses between the real-world change and the corresponding wiki update (e.g., between t_3 and t_4). However, there are cases where articles remain outdated for weeks before the respective data is (manually) updated.

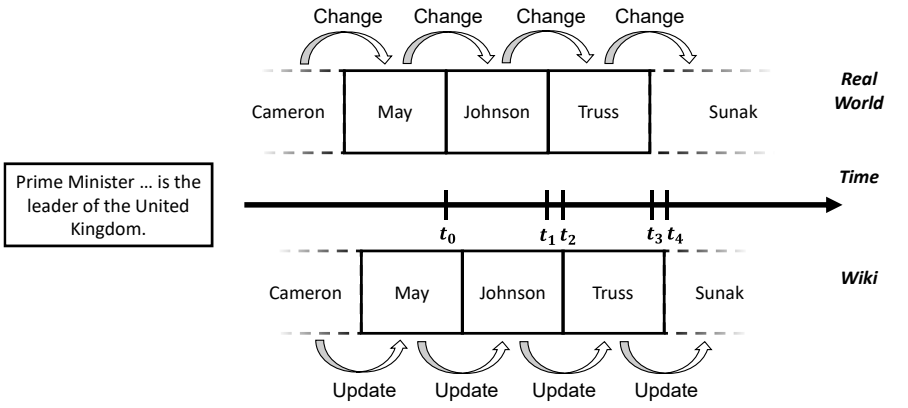

Figure 1. Illustration of Textual Wiki Data becoming Outdated

Related Work

The literature offers a large body of work for the assessment and improvement of currency for structured data (e.g., Heinrich and Klier 2015; Zak and Even 2017) as well as some initial contributions on unstructured data (e.g., Batini and Scannapieco 2016; Hao et al. 2020). Moreover, there are works discussing the assessment and improvement of DQ with a focus on wiki articles (e.g., Klier et al. 2021; Wang and Li 2020). This section is thus structured as follows. First, we briefly summarise existing work on currency in structured data and whether this work can be transferred to unstructured data. We then provide an overview of contributions on (textual) unstructured data. Subsequently, we review works focusing on the DQ of wiki articles in general, and works dealing with the currency of wiki articles in particular.

Currency of Structured Data

One of the earliest and most influential contributions to the assessment of time-related DQ dimensions of structured data was made by Ballou et al. (1998). They use parameters such as attribute value's age at the time of assessment, the (fixed and given) maximum lifespan of the attribute, and a sensitivity parameter to adjust the assessment based on the context. However, not all attributes have a predetermined lifespan, which led other authors to bring forward probability-based metrics (Heinrich et al. 2009; Heinrich and Klier 2011, 2015; Wechsler and Even 2012). Heinrich et al. (2009) propose a general procedure for developing probability-based metrics for the currency of structured data. Heinrich and Klier (2011) and Wechsler and Even (2012) both suggest metrics for the assessment of currency assuming an exponential distribution of attribute value lifespan. Heinrich and Klier (2015) relax this assumption, introduce additional metadata and provide a metric based on conditional probabilities, resulting in an interval-scaled and interpretable metric value. Taking up ideas from Heinrich and Hristova (2014) and Wechsler and Even (2012), Zak and Even (2017) suggest a continuous-time Markov chain model for detecting and handling currency declines. These approaches form important contributions and can be applied in an automated manner to assess the currency of structured data. Yet, being developed for structured data, they crucially rely on values of given, defined attributes (e.g., as contained in a relational database). Such attributes do not exist in the unstructured texts of wiki articles, rendering the approaches unsuitable for the direct application to this kind of data. To alleviate this issue, Batini et al. (2011) suggest to pre-process textual data and then apply an approach for structured data. However, pre-processing textual data still does not yield standardised and homogenous attributes across large bodies of text data (e.g., wiki articles) as required for the application of these approaches. Moreover, even if attributes were somehow derived from the textual data, the attributes are then themselves erroneous, as they are based on the textual data with imperfect DQ. Thus, these approaches cannot be applied to assess the currency of (unstructured) textual data.

Currency of Unstructured Data

The currency of unstructured data is commonly assessed based on the most recent update time of the considered data (cf., e.g., Batini and Scannapieco 2016; Firmani et al. 2016; Shah et al. 2015). The most

recent update time intuitively appears valuable in assessing currency. Additionally, it is usually automatically saved as metadata and is thus commonly available. Nevertheless, relying on the time elapsed since the last update is not sufficient to assess currency by far. For instance, a wiki article about the Russian war in Ukraine may already be outdated even though it was updated just days or hours ago, while another wiki article about the football world cup in 1990 may still be current despite its last update having occurred years ago. Hao et al. (2020) suggest an initial way to mitigate this issue by additionally considering average update frequencies. While these frequencies can be valuable in assessing currency, they have a drawback in that they can only be based on past data (i.e., previous updates). They cannot capture a change in pace of suddenly occurring events regarding a certain topic. Moreover, using the most recent update time, the average update frequency or other metadata does not offer suggestions for the improvement of outdated textual data, as the actual semantical content of the text is not considered at all.

DQ of Wiki Articles

The literature contains a large body of work on the DQ of wiki articles in general. These approaches generally strive to assess the DQ of whole articles (i.e., no indication of the DQ of, e.g., individual sentences, is provided). Several works aim to classify wiki articles into different quality categories, based on, for example, the length of the article or the reputation of its authors (cf., e.g., Blumenstock 2008; Shen et al. 2017; Wang and Li 2020; Wöhner et al. 2015; Zhang et al. 2020). Wikipedia employs a scoring system called ORES which assigns quality scores to single edits and full articles (MediaWiki 2023). These approaches have the ability to assess the quality of a wiki article based on its edit history and other factors. However, a shared issue among all these works is that they perceive DQ as a unidimensional construct, rather than a multidimensional one. Consequently, various aspects of DQ are compressed into a single rating, making it impossible to pinpoint specific quality problems (like outdated sentences). Therefore, capturing the currency of textual wiki data is out of the scope of these approaches.

Currency of Wiki Articles

To the best of our knowledge, despite the topic's undeniable relevance, only a handful of approaches dealing specifically with currency in wiki articles have been proposed. Stvilia et al. (2005) assess the currency of a wiki article based on the time it was last updated. Thus, this approach shares the drawback of the works for unstructured data presented above in that the time since the last update is not a sufficient indicator of currency. Moreover, such an approach does not offer suggestions for improvement of the data. Tran and Cao (2013) aim to find outdated data contained in the information boxes of Wikipedia articles by searching the Web for related data based on pattern recognition. This is an intriguing approach, but it is restricted to structured data contained in information boxes (which are not present in every wiki). In particular, it is not applicable to the textual content of wiki articles for the reasons outlined above. Lewoniewski (2017) proposes to improve the quality of Wikipedia articles by enriching the content of information boxes with data from information boxes of other language versions of the same article, taking into account DQ and popularity metrics. This approach may also be useful to address currency issues. However, it relies on multiple language versions of the same article being available. Moreover, it is limited to information boxes, and not applicable to the textual content of the article. Lewoniewski (2019) proposes to assess the currency of information boxes by examining the recent update frequencies, which is subject to the limitations discussed above. Finally, Klier et al. (2021) suggest an event-driven approach for the assessment of currency of wiki articles. They aim to detect events which affect the currency of wiki articles by monitoring their pageviews, with a sudden spike in pageviews indicating a currency-related event. This approach supports the automated detection of potentially outdated articles, but it does not indicate which part of the article has actually become outdated. Moreover, it does not offer suggestions for the improvement of currency.

Summary

In summary, while there are some important contributions for assessing and improving the currency of structured and unstructured data, they are not suited for textual wiki data. This is because they either rely on structured data attributes, which are not available in the context of textual wiki data, or only consider simple metadata such as most recent update time, but not the textual content itself. Further, valuable work has been conducted on the quality of wiki articles, but most of these works provide just a general quality assessment and only a few studies have discussed currency of wiki articles specifically. These proposed

approaches have drawbacks that limit their usefulness, such as only being applicable to certain structured elements of an article. They do not consider the textual content of the wiki articles itself and thus cannot improve its currency. Therefore, to address this research gap, we propose an approach for the assessment and improvement of the currency of textual wiki data.

Language Model-based Approach for the Currency of Wiki Articles

In this section, we first introduce a theoretical model and our approach. We then discuss the assessment of the currency of a wiki sentence based on our approach.

Theoretical Model and Approach

In general, changes occurring over time in the real world determine the currency of textual wiki data. As usually not all these real-world changes are known, uncertainty arises. To account for these uncertainties, we model the temporal changes in the real world as a stochastic process. Therefore, let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{X_t, t \in T\}$ a stochastic process that is defined on $(\Omega, \mathcal{F}, P)$. In our setting, the parameter set T reflects time and contains all considered points in time. For each point in time $t \in T$, the function $X_t: \omega \to X_t(\omega), \Omega \to S$ is a random variable (which is discrete or continuous) taking values in the data space S. In the case of certainty (i.e., all real-world changes regarding the data are known), $X_t(\omega)$ attains the single value $s \in S$ with probability 1. In reality however, certainty is not always given and $X_t(\omega)$ follows a (non-trivial) probability distribution.

In particular, X_{t_0} denotes the random variable describing real-world data at $t_0 = 0$, the time of creation of the data, where it attains one value $x_{t_0} \in S$ (under certainty). The real-world data then may change over time depending on $\{X_t, t \in T\}$, and at a certain point in time $t_a > t_0$ is stored in a wiki as the (certain) value $x_{t_a} \in S$. From then on, the real-world data may continue to change over time depending on $\{X_t, t \in T\}$, and at any later point in time $t_b > t_a$, the true value $x_{t_b} \in S$ may differ from the value $x_{t_a} \in S$ provided by the wiki. This stochastic process can also be used to reflect updates of the data in the wiki: If at a certain point in time t_c the data in the wiki is updated as the value $x_{t_c} \in S$, from then on, at any point in time $t_d > t_c$, the true value $x_{t_d} \in S$ may be different from the value $x_{t_c} \in S$ due to, once again, a change in the real world. For instance, in our running example, the stochastic process $\{X_t, t \in T\}$ begins at some point in time t_0, depending on the temporal horizon of interest (e.g., the start of Johnson's term). At a point in time t (between t_0 and t_1), this data is still unchanged (x_t) and stored in the wiki. At the point in time t_1, Truss becomes PM, and the true value $x_{t_1} \in S$ (Truss) is different from the value $x_t \in S$ (Johnson) provided by the wiki: The data provided by the wiki is outdated. It remains so until the point in time t_2, when the wiki is updated to reflect $x_{t_1} \in S$ (Truss), but then again becomes outdated when Sunak becomes PM.

In our model, at any point in time $t_z > t_a$ (the time at which the data was initially stored in the wiki), the data may be outdated. Today, manual effort would constantly be required to search and obtain the correct real-world value $x_{t_z} \in S$ under certainty. Using x_{t_z}, the data stored in the wiki can then either be current or has to be updated. However, our idea is to automatically assess the distribution of the random variable X_{t_z}, which may then be used to support the validation or facilitate the update of the data stored in the wiki. In particular, if the distribution of X_{t_z} shows that the data stored in the wiki is potentially outdated (e.g., its probability falls below a threshold), the need for an update can be suggested, with probable current real-world values also suggested by X_{t_z}. Generally, the distribution of the random variable X_{t_z} could be assessed by the term

$$f_{X_{t_z}|\{X_t, t \in T, t < t_z\}}(x_{t_z}|\{x_t, t \in T, t < t_z\}) \quad (1)$$

where $f_{X_{t_z}|\{X_t, t \in T, t < t_z\}}$ is a function incorporating information about changes of real-world data over time (e.g., how often the PM of a country changes), $\{X_t, t \in T\}$ is the stochastic process modelling changes in the real world as introduced above, and $x_t \in S$ for all t. However, using formula (1) is rarely feasible in realistic situations, as this assessment requires the knowledge of all x_t (including for points in time $t < t_a$ where no data was stored in the wiki yet, and including for points in time very briefly before t_z). In fact, the only points in time at which x_t is known are the points in time at which the data was initially stored in the wiki,

or the points in time at which it was updated previously. Denoting the set of all such points in time as $T_U \subset T$, one can then assess the distribution of X_{t_z} by the following term:

$$f_{X_{t_z}|\{X_t,t\in T_U\}}(x_{t_z}|\{x_t, t \in T_U\}) \quad (2)$$

Using this formula generally allows to support the validation or facilitate the update of the data stored in the wiki. However, it may not produce good results, in particular if the data is not updated frequently or the last update occurred a long time ago. Indeed, it may take a long time to detect changes that occurred in the real world and support an update of the data in the wiki. We thus propose to additionally use data y_t provided by recent external sources (e.g., news articles; details provided in the next section) at points in time $t \in T_E \subset T$, with T_E being defined as the subset of T, for which additional data are available. This data may serve as a proxy for x_t for points in time t close to t_z. The use of such data is crucial to detect real-world changes and to facilitate updates. The assessment of the distribution of X_{t_z} can then be conducted based on the following term:

$$f_{X_{t_z}|\{X_t,t\in T_U\},\{Y_t,t\in T_E\}}\left(x_{t_z}\middle|\{x_t, t \in T_U\}, \{y_t, t \in T_E\}\right) \quad (3)$$

To summarise, the approach in (3) allows taking into account

- known data stored in the wiki at the points in time $t \in T_U$
- data provided by external sources at the points in time $t \in T_E$
- a function incorporating information about changes of real-world data over time

to validate and potentially update data stored in a wiki at the point in time t_z.

With respect to a practical application, using the approach (3) to assess the distribution of X_{t_z} seems to require substantial manual effort for extracting the external data y_t and for conducting the assessment itself. However, language models (e.g., Brown et al. 2020; Devlin et al. 2019) can serve to address these problems and deliver a prediction for the distribution of X_{t_z} (cf. next section).

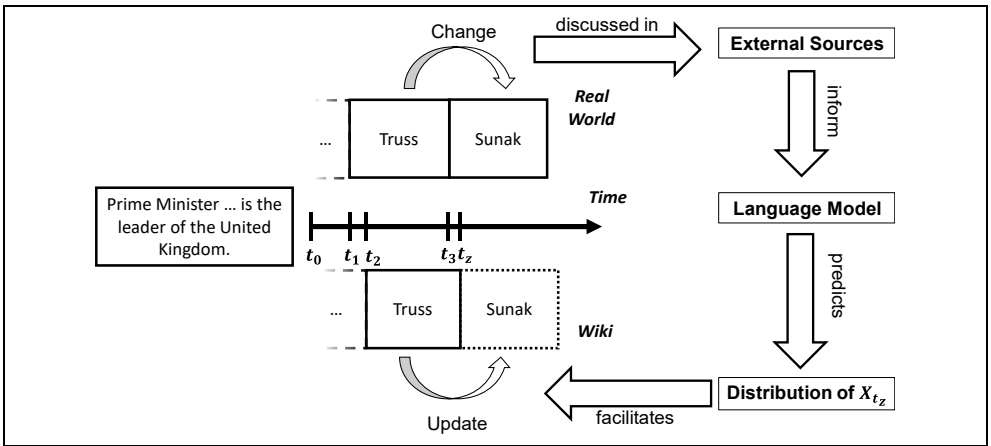

Figure 2. Overview of the Approach

Figure 2 illustrates our approach using the example of the latest change of the PM. The data that Truss is PM was stored in the wiki at the point in time t_2. At the point in time t_z (e.g., $t_z = 26/10/2022$; one day after the beginning of Sunak's tenure), the wiki still provides this data. Yet, Sunak has already taken office one day before t_z, at the point in time $t_3 = 25/10/2022$. The data provided by the wiki at t_z is outdated. However, Sunak's appointment as PM has been reported by external sources, for instance media outlets in their news articles, at times t_e with $t_3 \leq t_e \leq t_z$ (e.g., Ott 2022; BBC 2022). These news articles can be collected and processed in an automated manner. They support the prediction of the distribution of the random variable X_{t_z} by a language model according to formula (3). The language model's prediction of X_{t_z} suggests in our example that the PM is Sunak. In this way, the update of the wiki at the point in time t_z is facilitated. As the result of the update, Sunak is correctly provided as the PM by the wiki (dotted line in Figure 2).

Assessing the Currency of a Wiki Sentence

In this subsection, we elaborate on the assessment of the currency of a wiki sentence (called *focus sentence* in the following) based on the theoretical model and the approach introduced above. In particular, a focus sentence contains at least one *currency-related statement* which can either be current or outdated regarding the (real-world) fact it refers to. For example, the focus sentence "*Prime Minister Sunak is the leader of the United Kingdom*" contains, for example, the currency-related statements that **Sunak** is the current PM (and not, e.g., Truss), and that he is still alive ("**is** the leader"). In this sense, multiple different currency-related statements can be examined when assessing the currency of a focus sentence. Importantly, this example further illustrates that the currency of a currency-related statement crucially depends on (typically) one corresponding term ("is"/"was", "Sunak"/"Truss"/"Johnson"/...), which we call the *currency-related term*. In our running example, we mainly consider the currency-related statement about the current UK PM for which the name of the PM is the currency-related term.

To assess and improve a currency-related statement of a focus sentence, we propose to traverse four steps: First, a) keywords are extracted from the focus sentence. Ideally, these keywords express the meaning of the focus sentence. Naturally, the currency-related term (in our example, the name of the PM) is not used for the keyword extraction. Then, b) news articles from sources which contain the keywords are identified. Next, c) the news articles are divided into smaller paragraphs. Using semantic comparisons, the paragraphs potentially containing the sought update-relevant data are determined. Finally, d) the identified paragraphs are transferred to a language model and are used as input together with the focus sentence. If the paragraphs do indeed contain update-relevant data, the language model recognises the semantic relationships between the paragraphs and the focus sentence resulting in an up-to-date prediction for the currency-related term. This prediction corresponds to the result of our approach in (3) introduced above. The four steps are shown in Figure 3 in the case of our running example of the PM and are described in more detail below.

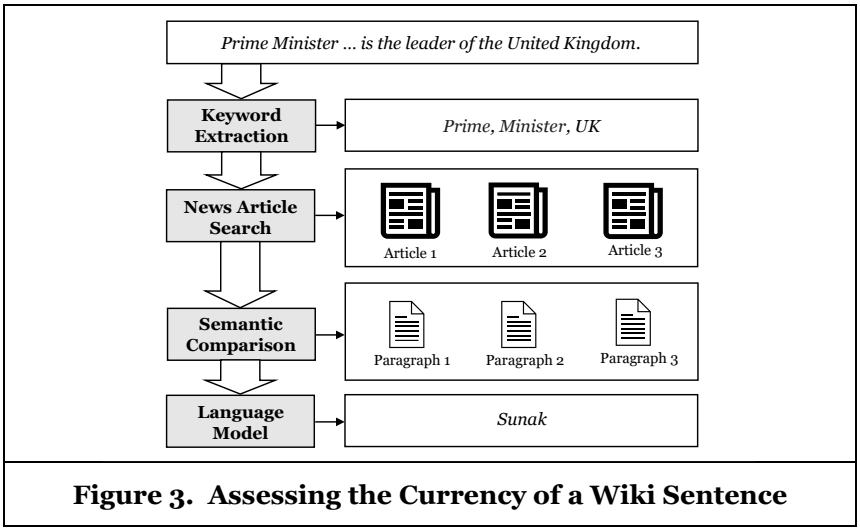

Figure 3. Assessing the Currency of a Wiki Sentence

Ad a): The starting point is the focus sentence, based on which keywords are extracted in the first step. For a valid result of the search for news articles, the keywords must contain the most important entities and terms mentioned in the focus sentence. For the keyword extraction, we propose to use the language model GPT-3.5 from OpenAI. This works well in a few-shot learning setting (Brown et al. 2020) in which a few exemplary sentences together with desirable keywords (e.g., "Djokovic currently holds ... Australian Open titles" as sentence and {"Djokovic", "Australian", "Open", "titles"} as keywords) are provided. On this basis, GPT-3.5 can be instructed to return a number of keywords for a given focus sentence. In our running example, the keywords "Prime", "Minister" and "UK" are extracted.

Ad b): In the second step, news articles containing all extracted keywords are searched. For our example, the result of this search should thus be articles that mention the name of the PM. Various news sources can be used as a basis for this search, in particular, renowned news outlets such as the *New York Times* or *The Guardian*. Importantly, the choice of adequate sources is crucial for our approach as it determines the set

of retrievable articles. For instance, using the *New York Times* as the only news source might not be helpful when assessing the currency of focus sentences from the domain of second division German football. Therefore, in the case of focus sentences from a specific domain, a news source which is known to cover this specific domain should be employed. On the other hand, if the focus sentences stem from various domains, a broad mixture of different news outlets is preferable. We explicitly focus on news articles as opposed to making use of a language model-infused search engine such as Bing or Bing Chat for three reasons. First, news articles are usually reliable whereas the results of a general web search may not be. Second, we can thus focus on obtaining recently published data (e.g., by specifying a certain search period). Third, these search engines themselves often use Wikipedia as a source, making them unsuitable for assessing and improving the currency of Wikipedia articles.

To further ensure that only news articles containing update-relevant data are retrieved, a search period can be specified, limiting the maximum age of eligible articles. With the specification of a longer search period, the number of potentially found articles increases, however their currency might decrease since already outdated data could also be contained. This results in a trade-off between the number of articles found and their currency. In the case of our example, a search period which is chosen too high might also yield articles which still refer to Truss as PM.

Ad c): We proceed by dividing all articles found into smaller paragraphs. A paragraph is defined as an excerpt consisting of a certain number of consecutive sentences. Moreover, the division of an article into paragraphs can either be disjoint or allow paragraphs to overlap. Then, semantical text embeddings can be used to identify the paragraphs that are most similar to the focus sentence. More precisely, the text embedding converts each paragraph into a numerical vector of a certain dimension. After also embedding the focus sentence into the same vector space, vector similarities between each paragraph embedding and the focus sentence embedding can be computed. Our idea is that the paragraphs most similar to the focus sentence are also most likely to contain update-relevant data. In our running example, such a paragraph should consist of, for instance, three sentences of which at least one ideally mentions the current PM.

Ad d): Finally, the identified similar paragraphs and the focus sentence are passed to a language model. Based on these paragraphs, the language model predicts an answer for the currency-related term which is decisive for the currency assessment of the focus sentence. Similar as in step a), a few-shot learning setting can be established based on which the language model is instructed to make a current prediction based on the additional data provided by the news articles. This prediction corresponds to the result of approach (3) introduced above. If we instruct the language model to make a suggestion for the current version of the focus sentence, we would ideally obtain “Sunak” for the current PM.

Evaluation

To demonstrate the performance of our approach in assessing and improving the currency of textual wiki data, we evaluate it on a set of fact-based sentences from Wikipedia containing currency-related statements. To this end, we have to show that our approach indeed is able to indicate whether a given currency-related statement is current, and – especially in the case of an outdated statement – whether it can provide a suggestion about the actually current statement, thus improving its currency. In this section, we first outline the methodology for our evaluation and describe how our approach was operationalised in detail. We further discuss the dataset used, and finally present our results.

Methodology

To evaluate our approach, we applied it to fact-based sentences including one or more potentially outdated currency-related statements. We masked currency-related terms and then let our approach predict the current state of the masked currency-related term (cf. Figure 4). For many language models such as BERT, filling in a masked word based on the implicit knowledge of the model already is an innate task (Devlin et al. 2019). This (masking and prediction) procedure can be applied to each currency-related term of a considered wiki sentence separately. By doing so, it can not only be assessed whether the whole focus sentence is current (i.e., in case at least one currency-related term is outdated), but it can also be determined which currency-related terms are outdated. For all outdated currency-related terms, particular suggestions for improvement are provided.

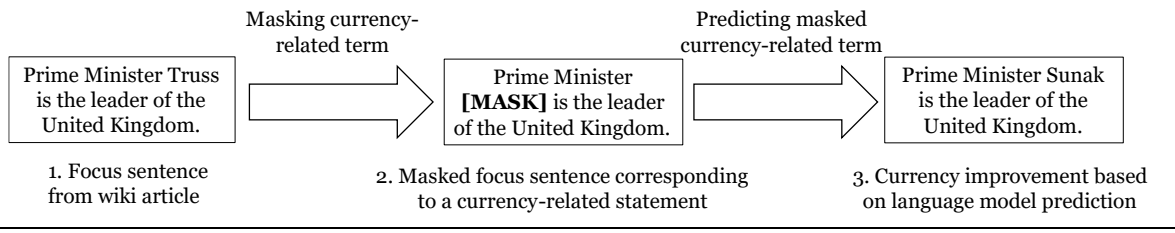

Figure 4. Currency Assessment and Improvement of a Focus Sentence

We evaluated and analysed different variants of our language model-based approach, which mainly differ in the language model used and the additional external data. Hence, we analysed the effectiveness of our approach along two dimensions: We first considered the two well-known language models BERT (Devlin et al. 2019) and GPT-3.5 (more precisely, the version gpt-3.5-turbo (OpenAI 2023a)) without any additional data. Since BERT was (inter alia) mainly pretrained on the English Wikipedia corpus in 2018 (Devlin et al. 2019), we did not expect BERT to have implicit knowledge about real-world changes which happened in 2019 or later. GPT-3.5, as a more recent and potentially more powerful language model, complements BERT as second baseline variant. As GPT-3.5 is mainly trained on data (including Wikipedia) from up to September 2021 (Brown et al. 2020; OpenAI 2023a), this variant can be expected to provide currency assessments and improvements for facts that have not changed in the real world since this point in time. By employing these baseline variants, we can separately analyse the performance of the sole language models as a foundation for currency assessment and improvement. This gives us the opportunity to isolate their respective contribution. More precisely, we strive to gain a deeper understanding of their differences by examining their predictions in terms of implicit knowledge and overall quality of predictions.

Second, the impact of additional data from news articles is evaluated in further variants. In particular, one of our goals in the evaluation is to analyse the influence of different sets of news articles. To this end, we consider multiple variants using news sources which differ in quantity and characteristics of articles they comprise. Based on the performance of their variant, we can not only analyse their general suitability to support our approach, but also gain insights into the topics they cover, respectively. From this point on, we focus on GPT-3.5 as language model because in a preliminary test, its results were much more promising than BERT's. The first variant incorporating additional data from an external news source, *GPT + Guardian*, employs GPT-3.5 and whole online articles from the well-known British news outlet *The Guardian*[1]. Analogously, *GPT + Aylien* is an alternative variant utilising additional data from the *Aylien News API*[2] by extracting excerpts of news articles from a multitude of different news outlets. Hence, the main difference between the *GPT + Aylien* and the *GPT + Guardian* variants lies in the pool of accessible news sources. In the case of *GPT + Aylien* this pool comprises a much higher number of different outlets, while in the case of the *GPT + Guardian* variant, it consists of the online articles of one single high-quality outlet which presumably only covers a certain share of real-world changes. As a fifth and final variant, we also consider the combination *GPT + News* which comprises both, the news articles from *The Guardian* and the *Aylien News API*. By doing so, we can examine whether the combination of multiple news sources positively affects the performance of our approach. The variants are summarised in Table 1.

Variant	**Description**
BERT	This variant uses the standard pre-trained BERT base model together with its inherent mask prediction task, which is part of BERT's pretraining procedure.
GPT	This variant uses the GPT-3.5 language model from OpenAI. The task for the language model is defined to respond with the current term for the [MASK].
GPT + Guardian	This variant uses the GPT-3.5 language model together with external data in the form of selected news articles from *The Guardian*.
GPT + Aylien	This variant uses the GPT-3.5 language model together with external data in the form of selected news articles provided by the *Aylien News API*.

[1] https://open-platform.theguardian.com/

[2] https://aylien.com/product/news-api

GPT + News	This variant uses the GPT-3.5 language model together with external data in the form of selected news articles provided by *The Guardian* and the *Aylien News API*.

Table 1. Summary of the Variants of our Approach

Operationalisation of the Approach

In the following, we describe the operationalisation of our approach for the variants incorporating news articles in more detail.

a) Keyword Extraction: Keywords are extracted from each sentence using GPT-3.5 and two different prompts employing a few-shot learning setting (Brown et al. 2020) as described above. The two prompts (called Keywords_specific and Keywords_general) extract different sets of keywords, where Keywords_specific contains more keywords (which mandatorily must be contained in an article to be selected) than Keywords_general. Using different sets of keywords for the search increases the chances of finding suitable articles. They extracted on average 54 (Keywords_specific) and 89 (Keywords_general) articles, leading to a set of on average 115 articles without duplicates together. The full prompts are publicly available and provided online[3].

b) Article Search: For the news article search, the sources *The Guardian* and the *Aylien News API* are used. The search is limited to articles published from January 15th, 2023, to April 15 th, 2023. Each set of keywords is used for one search. *The Guardian* returns full news articles that contain all of the keywords. The *Aylien News API* returns only the summary of news articles whose titles contain all keywords. The results of both APIs are ordered by relevance to the search term.

c) Paragraph Extraction: Each result returned by the news sources is pre-processed by removing characters and strings not relevant to the content (e.g., HTML tags and layout information). Next, each article and summary is divided into smaller paragraphs consisting of three consecutive sentences. Each sentence is the start of one paragraph, meaning sentences 1, 2, and 3 form the first paragraph, sentences 2, 3, and 4 form the next paragraph, etc. Paragraphs that do not contain any of the keywords are removed, as it is very likely that they do not contain update-relevant data. Next, a text embedding of each paragraph and the masked sentence is calculated, which captures the semantic content of the text. To this end, we use the 1536-dimensional contextual text embedding *text-embedding-ada-002* from OpenAI (Greene et al. 2022) to convert text into comparable vectors of equal dimension. Then, the cosine similarity between the embedding of the masked sentence and the embedding of each paragraph is calculated and the paragraphs are ordered from highest to lowest similarity. If a sentence from a news source is contained in multiple paragraphs, only the paragraph with the highest cosine similarity is kept.

d) Selection of Paragraphs for the Language Model: The four most similar paragraphs that exceed a cosine similarity of an empirically determined threshold (0.78) are selected and passed to the language model as additional input. Using a prompt, the model is instructed to make a current prediction based on the paragraphs provided. This is done in a few-shot learning setting as described above. If no paragraph is selected, no additional data is passed to the language model and the prediction is obtained in the same manner as in the baseline variant.

Dataset

As there is no established dataset for the assessment and improvement of the currency of wiki sentences, we created a novel dataset comprising fact-based sentences from Wikipedia. To this end, we focused on sentences referring to facts that can become outdated at all, such as the current incumbent of a political position or the current club a certain football player is playing for, such that the sentences include at least one currency-related term. As Wikipedia was part of the training data for BERT and GPT 3.5, we also emphasised facts that have changed recently (2018 or later). In this way, it could be avoided that large parts of our dataset are included in the training datasets of the language models. We extracted such facts based on Wikipedia articles and collected 543 sentences referring to those facts. To evaluate the performance of our approach across different domains, these sentences were collected across multiple domains of Wikipedia, namely football (164 sentences), (general) sports (34 sentences), politics (144 sentences),

[3] https://docs.google.com/spreadsheets/d/188-iVFDopohL8UVBnSyYT59IQmiOMLu_v4Lk3RTi9e8/edit?usp=sharing

economics (98 sentences), science (62 sentences) and culture (41 sentences). Within these domains, sentences about people and facts with different backgrounds and varying degrees of public awareness were selected to further ensure a broad evaluation basis. In a pre-processing step, dispensable tokens such as phonetics and superfluous terms such as the second name of a person were removed from the Wikipedia sentences. On this basis, we identified the currency-related terms. We then masked one of the currency-related terms according to a fixed systematic approach determined beforehand (e.g., for a number of political positions, we always masked the incumbent). Finally, we labelled the sentence with the current term, which is consistent with the fact holding true in the real world, as gold token. The final labelling was conducted on April 21st, 2023 and each label was verified with an online search. For example, the gold token of the sentence "Prime Minister [MASK] is the leader of the United Kingdom" would be "Sunak", because he became PM in October 2022 and has been incumbent until April 21st, 2023. To provide full transparency, the dataset and all evaluations are publicly accessible[3].

Results

Even though it would be possible to evaluate multiple predictions (e.g., the top 3 predictions) of each variant, we deliberately focus on the top 1 prediction to make the evaluation very challenging. Providing only one prediction is also most helpful to support updates, as it allows for an easier decision (keep previously stored data or use the prediction). Thus, we generated the top 1 prediction for each currency-related term and for each variant of our approach as described in the subsection "Methodology". Two researchers manually checked every prediction for equality with the previously defined gold token and labelled the prediction as correct (thus current) in this case. Non-equal predictions potentially synonymous with the gold token were discussed by the researchers until consent was reached. The accuracy (percentage of semantically correct and current predictions) of each variant for all domains is presented in Table 2.

Variant Dataset	BERT	GPT	GPT + Guardian	GPT + Aylien	GPT + News
Overall	0.335	0.615	0.713	0.816	0.821
Football	0.311	0.506	0.628	0.848	0.847
Sports (general)	0.294	0.647	0.676	0.764	0.705
Economy	0.296	0.612	0.653	0.684	0.673
Science	0.339	0.694	0.710	0.742	0.758
Culture	0.146	0.610	0.805	0.805	0.854
Politics	0.453	0.701	0.833	0.917	0.938

Table 2. Performance (Accuracy) of the Variants

Analysis of Predictions of GPT + News

For the strongest performing variant *GPT + News*, we additionally conducted a deeper analysis of the predictions. To this end, each prediction was examined manually. The results are presented in Table 3.

Correct predictions	**446**
(I): based on paragraphs	358
(II): based solely on implicit knowledge of the language model	88
Incorrect predictions	**97**
(III): no paragraphs are found	33
(IV): paragraphs are found, but do not contain update-relevant data	41
(V): paragraphs containing update-relevant data are found, but the data is not recognised	18
(VI): paragraphs contain retrospectives written in the present tense	5

Table 3. Detailed Evaluation of Variant *GPT + News*

Overall, 446 out of 543 predictions of the variant *GPT + News* are correct. In (I) 358 out of these 446 cases, paragraphs are extracted from external sources, and the correct predictions are made taking these paragraphs into account. In (II) 88 out of the 446 cases, no paragraphs are found or the paragraphs found do not contain update-relevant data, but still a correct prediction is made.

97 predictions of the variant *GPT + News* are incorrect. In (III) 33 of these cases, no paragraphs are found because no articles that contain the extracted keywords are available or none of the paragraphs have a similarity score above the threshold. Thus, the prediction is made based on the implicit knowledge of the model, which in this case is incorrect. In (IV) 41 cases, paragraphs are found, but they do not contain the update-relevant data for the currency-related term. As a result, the model gives an outdated or incorrect prediction. In (V) 18 cases, paragraphs containing update-relevant data are extracted from external sources, but the model does not recognise the update-relevant data and makes a wrong prediction. In nine of these cases, the update-relevant data is contained directly in the paragraphs. In the nine other cases, the update-relevant data is contained indirectly in the paragraphs and can be understood from the context (e.g., that James Taiclet was (and is not anymore) the CEO of American Tower can be understood even though it is not directly mentioned in the paragraph: "… American Tower Corp Says CEO Thomas Bartlett's 2022 Total Compensation …"). Finally (VI), in five cases, the paragraphs contain retrospectives written in the present tense, and thus the model gives an outdated prediction based on the retrospective.

Discussion

In the evaluation section, we considered five different variants of our approach and evaluated their performance. In this section, we discuss these results and compare the results of the different variants. We then explicate our choice of parameters for the different variants and discuss the robustness of the results.

Discussion of Results

As a first baseline variant, we examined the language model BERT without any additional data, using only BERT's innate mask-fill task to predict the masked currency-related term of a focus sentence. As shown in Table 2, BERT achieves the lowest accuracy across all domains (0.335 overall). A deeper analysis of BERT's predictions reveals two main reasons for its rather poor performance in currency assessment and improvement. First, in many cases BERT does not capture the semantical content and the currency relevance of the focus sentence. Rather, it tends to make predictions based on known syntactic structures and patterns of adjacent words which both are adopted from the training corpus. This is in line with discussions in literature (Kassner and Schütze 2020; Poerner et al. 2020). For example, for the masked sentence "The host of the most recent Summer Olympics was the city [MASK]", BERT predicted "hall". The expression "city hall" might be reasonable in some other contexts, but in our case, disregarding the given context of the host city of the last Olympic games leads to a nonsensical answer. Second, as already pointed out in the previous section, BERT was mainly pretrained on the English Wikipedia corpus in 2018. Hence, BERT's implicit knowledge about the real world is outdated regarding facts which underwent real-world changes since this point in time. For example, this phenomenon becomes evident in sentences about persons who died recently (e.g., the American actor Lance Solomon Reddick, who died in March 2023), or politicians who no longer hold a political position they indeed once held (e.g., former German chancellor Angela Merkel, who held this position until December 2021). Here, BERT gives outdated predictions.

GPT as a second baseline variant for our approach achieved significantly higher accuracy than BERT across all domains (0.615 overall). Based on a detailed analysis of GPT's predictions, we found that this substantial improvement in accuracy is mainly due to the alleviation of BERT's major weaknesses discussed above. First, GPT's predictions have a higher overall quality in terms of awareness for the context and currency of the focus sentence. More precisely, GPT is consistently able to provide predictions that semantically match the actual content of the focus sentence, while not resorting to known syntactical patterns adopted from the training data. Importantly, this enables GPT to perform valid currency assessment and improvement, because an actual currency-related prediction is made based on the content of the sentence, and not based on known patterns from other contextual different sentences. Second, as already pointed out, GPT's implicit knowledge is more up to date because of its more recent training data. Therefore, unlike BERT, GPT has the potential to correctly assess the currency of statements which underwent changes in the real world between 2018 and 2021. For these two reasons, GPT provides the current prediction "Tokyo" for the masked focus sentence "The host of the most recent Summer Olympics was the city [MASK]" without any additional input (the most recent Summer Olympics took place in Tokyo in July and August 2021). Overall, we found that GPT provides high-quality predictions for the currency assessment and improvement of textual wiki data and thus poses a suitable language model for our approach.

Thus, the three further variants are based on GPT with each variant utilising a different basis of sources of news articles. *GPT + Guardian* produces 387 correct predictions, resulting in an overall accuracy of 0.713, strongly improving over the baseline GPT variant. *GPT + Aylien* performs even better, with 56 more correct predictions and an overall accuracy of 0.816. These outperformances against the baseline demonstrate that the incorporation of news articles is an effective strategy to assess the currency of textual wiki data. For example, while the baseline variants predict that Queen Elizabeth still is Queen of the UK based on their outdated implicit knowledge, the variants which incorporate news identify news articles which report the passing of Queen Elizabeth or refer to her as former Queen of the UK. They thus correctly suggest that Queen Elizabeth no longer is Queen of the UK.

The better performance by *GPT + Aylien* compared to *GPT + Guardian* is mainly due to two factors: First, the *Aylien News API* provides news articles for 43 more sentences than *The Guardian*. Second, the average similarity score of the paragraphs found by the *Aylien News API* is higher, indicating that the articles found are more relevant. This is supported by the fact that the predictions of *GPT + Guardian* are wrong twice as often compared to *GPT + Aylien* even when paragraphs are found. The reason for this is most likely that the *Aylien News API* retrieves articles from a large number of outlets, and as a result, *GPT + Aylien* has a much broader base of articles from which to select the most relevant paragraphs. Indeed, topics such as local sports are more likely to be covered by an outlet contained in the *Aylien News API* as compared to *The Guardian*. Also, *The Guardian* provides full articles, while the *Aylien News API* only retrieves article summaries. These summaries contain the update-relevant data in a more compact form, which is beneficial for our approach as it only uses paragraphs of a maximum length of three sentences.

GPT + News which combines both sources of news articles, shows the best performance with an overall accuracy of 0.821. Our analysis of the predictions of *GPT + News* shows that, for most sentences in the dataset, this variant finds relevant paragraphs articles and makes a correct prediction taking these paragraphs into account. However, compared to *GPT + Aylien*, only a small improvement can be observed. The *Aylien News API* retrieves the summaries of articles from many different news outlets, including articles from *The Guardian*. While the full articles from *The Guardian* provide paragraphs leading to an additional correct prediction in a few cases, occasionally, it can also be observed that using only one of the news sources leads to a correct prediction, but adding the other source interferes with it and an incorrect prediction results. Thus, adding *The Guardian* on top of the *Aylien News API* only leads to a minor improvement in overall performance.

The differences in performance between the five variants can be seen across all six domains (cf. Table 2). The exception is that, in a few domains, a slight decrease in accuracy from *GPT + Aylien* to *GPT + News* can be observed because of the reason described above. Across all variants politics always performs best, with *GPT + Aylien* and *GPT + News* achieving an accuracy above 0.9. The reason for this is that politics is a very well-covered topic in many news outlets. Thus, a multitude of articles are available from which the most relevant paragraphs can be extracted, leading to a high accuracy of predictions. In summary, all three variants, *GPT + Guardian*, *GPT + Aylien* and *GPT + News*, substantially outperform the baseline variants (cf. Table 2), showing the effectiveness of our approach. In particular, *GPT + Aylien* and *GPT + News* exhibit a very strong performance with an accuracy above 0.8. Thus, these variants are suitable for supporting the assessment and improvement of the currency of textual wiki data.

Choice of Parameters and Robustness

Our approach involves several parameters. In the following, we make transparent our choice of parameters for the different variants of our approach and discuss the robustness of our results.

The first parameter in our approach is the threshold which each paragraph passed to the language model must satisfy to ensure it has a minimum level of relevance to the focus sentence. Based on empirical tests, a value of 0.78 was selected as the value for this parameter. Slightly higher or lower values could also be chosen without strongly affecting the results, as the similarity score of a paragraph to the focus sentence is only an approximation of its relevance. The second parameter is the maximum number of paragraphs which are passed to the language model. For this parameter, the value four was selected. Based on our observations, usually, the first paragraph contains the most update-relevant data. However, sometimes update-relevant data is only contained in the third or fourth paragraph, despite having lower similarity scores. This is the reason for us to always pass four paragraphs to the model, but more (or less, e.g., three) paragraphs could also be passed without strongly affecting the results. The third parameter is number of

sentences included in each paragraph. Here, the value three was selected. Based on our observations, in many cases, one single sentence contains all update-relevant data. However, in some cases, all three sentences contain update-relevant data. Therefore, to not disregard any update-relevant data, we decide to include three sentences in one paragraph. Again, slightly altering the value of this parameter would not substantially change the results. The fourth parameter of our approach is the search period, which determines the maximum age of eligible articles. As described above, the trade-off between the number of articles and their currency must be considered when specifying this parameter. Using a longer search period results in more data, but some of this data may already be outdated, making it irrelevant or even detrimental. We chose a search period of three months because it is long enough to find data even for less popular topics, but not so long that outdated articles are retrieved. As with the other parameters, smaller changes would not significantly affect the results.

In general, the main factor in the choice of the described parameters is the trade-off between the amount of data and its relevance: By increasing the threshold, decreasing the number of paragraphs, decreasing the number of sentences per paragraph, or decreasing the length of the search period, less but potentially more relevant data is passed to the model, which means focusing on relevance. By doing the opposite, more but potentially less relevant data is passed to the model, focusing on the amount of data. This is the focus of our choice of parameters. We instructed the model to recognise if the paragraphs do not contain any update-relevant data, and to make a prediction based on implicit knowledge in this case. This worked well for our model, as irrelevant data only very rarely led to an incorrect prediction (cf. Section "Analysis of Predictions of *GPT + News*"). Indeed, if the model can be instructed in a way that irrelevant data never leads to a wrong prediction, a strong focus on the amount of data is favourable. On the other hand, the easier the model can be distracted by irrelevant data, the more the choice should focus on relevance.

We further analysed the different variants of our approach and their performance for currency improvement with respect to just recently changed sentences to infer whether the knowledge of the language models (attained during their training period) plays a significant role for our results. More precisely, we considered only sentences (of the overall 543 sentences) for which the last change of the gold token occurred after a certain cutoff date and let the cutoff date vary. The results of these analyses are presented in Table 4.

Variant Cutoff date	BERT	GPT	GPT + Guardian	GPT + Aylien	GPT + News	Number of sentences
01/01/18	0.251	0.456	0.600	0.758	0.763	355
01/01/19	0.231	0.436	0.585	0.751	0.757	337
01/01/20	0.214	0.382	0.553	0.730	0.737	304
01/01/21	0.219	0.310	0.498	0.709	0.716	261
01/01/22	0.235	0.164	0.399	0.634	0.667	183
01/01/23	0.156	0.190	0.483	0.621	0.672	58

Table 4. Performance (Accuracy) of the Variants regarding Different Cutoff Dates

The declining performance of the baseline variants *BERT* and *GPT* with more recent cutoff dates clearly indicates that their training data mainly lie a few years in the past. Importantly, the results show that incorporating news articles is beneficial for each cutoff date, with the most benefit realized for more recent cutoff dates (i.e., more recently changed facts). For instance, while the baseline variant *GPT* only achieved an accuracy of 0.190 for a cutoff date of 01/01/2023, the variant *GPT + News* achieved an accuracy of 0.672. This means that *GPT + News* could be used to improve the currency of 3.54 times as many sentences, stressing the importance of incorporating news articles for currency improvement, and thus, our approach.

Further, we analysed the relation between the number of news articles found for a focus sentence and the performance of our approach. This would give us an idea whether our approach would benefit from including more (diverse) sources besides *The Guardian* and the *Aylien News API* leading to a higher number of found news articles. For that reason, we categorized each sentence in the dataset into one of the four categories "no news coverage" (0 news articles found in the sources for *GPT + News*), "limited coverage" (1-49 articles), "moderate coverage" (50-200 articles), and "extensive coverage" (>200 articles). The results of the variants are presented in Table 5. The results clearly show that higher coverage leads to improved performance for the variant *GPT + News* (naturally not for the baseline variants, which do not incorporate news), with its accuracy steadily increasing from 0.671 (no coverage) to 0.908 (extensive

coverage). This further highlights the benefits of incorporating additional data for currency improvement and especially of including more and diverse sources (in the future).

Variant Coverage	BERT	GPT	GPT + Guardian	GPT + Aylien	GPT + News	Number of sentences
No news coverage	0.303	0.632	0.671	0.671	0.671	76
Limited coverage	0.335	0.608	0.634	0.773	0.773	194
Moderate coverage	0.322	0.678	0.760	0.909	0.884	121
Extensive coverage	0.361	0.566	0.796	0.868	0.908	152

Table 5. Performance (Accuracy) of the Variants regarding News Coverage

Finally, we also examined which results our approach yields when applied with the new language model GPT-4 (more precisely, the version GPT-4-0613 (OpenAI 2023b)). To this end, we employed two additional variants of our approach, *GPT-4* and *GPT-4 + News*, which were constructed and applied analogously to the corresponding variants with GPT-3.5. *GPT-4* yielded an accuracy of 0.624, marginally outperforming the variant with GPT-3.5 (accuracy of 0.615). This can be explained by the fact that GPT-4 is a more powerful language model, but like GPT-3.5, is mainly trained on data from up to September 2021 (OpenAI 2023b). *GPT-4 + News* was the best performing variant overall with an accuracy of 0.840, as it was slightly better in using the additional data from news articles than the variant with GPT-3.5 (accuracy of 0.821). This indicates that our approach has the potential to provide even stronger results with possibly more powerful language models in the future.

Conclusion, Limitations and Future Work

Wikis are ubiquitous in organisational and private use and provide a wealth of textual data. Maintaining the currency of this textual data is both important and difficult, requiring large manual efforts. Previous approaches from literature provide valuable contributions for assessing the currency of structured data or wiki articles as a whole but are unsuitable for textual wiki data. Thus, we propose both a theoretical model and a novel approach which support the assessment and improvement of the currency of textual wiki data in an automated manner. The theoretical model comprehensively describes currency changes in wikis. Grounded on this theoretical model, our approach makes use of data retrieved from recently published news articles and a language model to determine the currency of fact-based wiki sentences and suggest possible updates. Our evaluation conducted on 543 sentences from six domains shows that the approach yields promising results with accuracies over 80% (focussing only on the top 1 prediction and a 3-month window for selecting news articles, which both make the evaluation very challenging). Thus, it is well-suited to support the assessment and improvement of the currency of textual wiki data. It can be employed to continually assess the currency of data which may become outdated, or to conduct specific assessments (e.g., when the last update of a currency-related term was a year ago, or for all football-related data when the football transfer period ends).

Nevertheless, the work at hand has some limitations which could be a starting point for future research. To begin with, the presented operationalization of the approach does not yet take full advantage of the entire theoretical model and the general approach. Future work could incorporate known data stored in the wiki in the past, and it could explicitly consider information about changes of real-world data over time (e.g., leading to individual "decline rates" of textual data). This way, the performance could be further improved. In addition, we applied our approach with fixed parameter settings in the evaluation. The approach could be refined to adjust the parameters depending on the domain or data (e.g., using a shorter search period for data which changes highly frequently). Furthermore, our approach relies on obtaining data from recently published news articles to facilitate current predictions, and thus, updates. In the future, it could be extended to additionally consider further trustworthy sources (e.g., websites of universities and corporations). Moreover, our approach needs to be applied frequently to fact-based sentences in a wiki to assess their currency. An intriguing idea which could be explored is to apply our approach in "reverse order" by first extracting current data from news articles as soon as they are published, and then searching and updating respective textual data in the wiki. Finally, we focused on public wikis in this paper. It would be highly interesting to extend our approach to organisational knowledge bases, such as enterprise wikis or texts in a document management system. Here, updates could be facilitated by other, more recent documents in the same organisation.

References

Alqahtani, F. H. 2017. "Users' Perspectives on Using Wikis for Managing Knowledge: Benefits and Challenges," *Journal of Organizational and End User Computing* (29:3), pp. 1-23.

Ballou, D. P., Wang, R. Y., Pazer, H. L., and Tayi, G. K. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4), pp. 462-484.

Batini, C., Barone, D., Cabitza, F., and Grega, S. 2011. "A Data Quality Methodology for Heterogeneous Data," *International Journal of Database Management Systems* (3:1), pp. 60-79.

Batini, C., and Scannapieco, M. 2016. *Data and Information Quality*, Cham: Springer.

BBC. 2022. "Rishi Sunak: World Leaders Welcome Next UK Prime Minister," available at https://www.bbc.com/news/uk-63378673, accessed on September 1st, 2023.

Beck, R., Rai, A., Fischbach, K., and Keil, M. 2015. "Untangling Knowledge Creation and Knowledge Integration in Enterprise Wikis," *Journal of Business Economics* (85:4), pp. 389-420.

Bhatti, Z. A., Baile, S., and Yasin, H. M. 2018. "Assessing Enterprise Wiki Success from the Perspective of End-Users: An Empirical Approach," *Behaviour & Information Technology* (37:12), pp. 1177-1193.

Blumenstock, J. E. 2008. "Size Matters: Word Count as a Measure of Quality on Wikipedia," in *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China, pp. 1095-1096.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. 2020. "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, pp. 1877-1901.

Cichy, C., and Rass, S. 2019. "An Overview of Data Quality Frameworks," *IEEE Access* (7), pp. 24634-24648.

Dang, Q.-V., and Ignat, C.-L. 2017. "An End-to-End Learning Solution for Assessing the Quality of Wikipedia Articles," in *Proceedings of the 13th International Symposium on Open Collaboration,* Berlin, Germany, pp. 1-10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, Minnesota, pp. 4171-4186.

Firmani, D., Mecella, M., Scannapieco, M., and Batini, C. 2016. "On the Meaningfulness of "Big Data Quality"," *Data Science and Engineering* (1:1), pp. 6-20.

Greene, R., Sanders, T., Weng, L., and Neelakantan. 2022. "New and improved embedding model," available at https://openai.com/blog/new-and-improved-embedding-model, accessed on April 30th, 2023.

Hao, S., Chai, C., Li, G., Tang, N., Wang, N., and Yu, X. 2020. "Outdated Fact Detection in Knowledge Bases," in *Proceedings of the IEEE 36th International Conference*, pp. 1890-1893.

Hatcher-Gallop, R., Fazal, Z., and Oluseyi, M. 2009. "Quest for Excellence in a Wiki-Based World," in *Proceedings of the IEEE International Professional Communication Conference*, pp. 1-8.

He, W., and Yang, L. 2016. "Using Wikis in Team Collaboration: A Media Capability Perspective," *Information & Management* (53:7), pp. 846-856.

Heinrich, B., and Hristova, D. 2014. "A Fuzzy Metric for Currency in the Context of Big Data," in *Proceedings of the 22nd European Conference on Information Systems,* Tel Aviv, Israel.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., and Szubartowicz, M. 2018. "Requirements for Data Quality Metrics," *Journal of Data and Information Quality* (9:2), pp. 1-32.

Heinrich, B., and Klier, M. 2011. "Assessing Data Currency—A Probabilistic Approach," *Journal of Information Science* (37:1), pp. 86-100.

Heinrich, B., and Klier, M. 2015. "Metric-Based Data Quality Assessment — Developing and Evaluating a Probability-Based Currency Metric," *Decision Support Systems* (72), pp. 82-96.

Heinrich, B., Klier, M., and Kaiser, M. 2009. "A Procedure to Develop Metrics for Currency and its Application in CRM," *Journal of Data and Information Quality* (1:1), pp. 1-28.

Kassner, N., and Schütze, H. 2020. "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811-7818.

Klier, M., Moestue, L., Obermeier, A., and Widmann, T. 2021. "Event-Driven Assessment of Currency of Wiki Articles: A Novel Probability-Based Metric," in *Proceedings of the 42nd International Conference on Information Systems,* Austin, TX.

Lewoniewski, W. 2017. "Enrichment of Information in Multilingual Wikipedia based on Quality Analysis," in *Business Information Systems Workshops. BIS 2017. Lecture Notes in Business Information Processing*, W. Abramowicz (ed.), Cham: Springer, pp. 216-227.

Lewoniewski, W. 2019. "Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia," in *Business Information Systems Workshops. BIS 2018. Lecture Notes in Business Information Processing*, W. Abramowicz and A. Paschke (eds.), Cham: Springer, pp. 619-633.

MediaWiki. 2023. "ORES," available at https://www.mediawiki.org/wiki/ORES, accessed on April 30th, 2023.

Nelson, R., Todd, P., and Wixom, B. 2005. "Antecedents of Information and System Quality: An Empirical Examination within the Context of Data Warehousing," *Journal of Management Information Systems* (21:4), pp. 199–235.

OpenAI. 2023a. "GPT-3.5," available at https://platform.openai.com/docs/models/gpt-3-5, accessed on April 30th, 2023.

OpenAI. 2023b. "GPT-4," available at https://platform.openai.com/docs/models/gpt-4, accessed on September 1st, 2023.

Orr, K. 1998. "Data Quality and Systems Theory," *Communications of the ACM* (41:2), pp. 66-71.

Ott, H. 2022. "Rishi Sunak Appointed Prime Minister of the United Kingdom," available at https://www.cbsnews.com/news/rishi-sunak-appointed-prime-minister-of-the-united-kingdom/, accessed on September 1st, 2023.

Poerner, N., Waltinger, U., and Schütze, H. 2020. "E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT," in *Findings of the Association for Computational Linguistics*, pp. 803-818.

Seibert, M., Preuss, S., and Rauer, M. 2011. *Enterprise Wikis: Die erfolgreiche Einführung und Nutzung von Wikis in Unternehmen*, Wiesbaden: Gabler Verlag.

Shah, A. A., Ravana, S. D., Hamid, S., and Ismail, M. A. 2015. "Web Credibility Assessment: Affecting Factors and Assessment Techniques," *Information Research* (20:1), pp. 1-28.

Shen, A., Qi, J., and Baldwin, T. 2017. "A Hybrid Model for Quality Assessment of Wikipedia Articles," in *Proceedings of the Australasian Language Technology Association Workshop*, pp. 43-52.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. 2005. "Assessing Information Quality of a Community-Based Encyclopedia," in *Proceedings of the International Conference on Information Quality,* Cambridge, MA, pp. 442-454.

Tran, T., and Cao, T. H. 2013. "Automatic Detection of Outdated Information in Wikipedia Infoboxes," *Research in Computing Science* (70:1), pp. 211-222.

Wang, P., and Li, X. 2020. "Assessing the Quality of Information on Wikipedia: A Deep-Learning Approach," *Journal of the Association for Information Science and Technology* (71:1), pp. 16-28.

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5-33.

Wechsler, A., and Even, A. 2012. "Using a Markov-Chain Model for Assessing Accuracy Degradation and Developing Data Maintenance Policies," in *Proceedings of the 10th International Conference on Design Science Research in Information Systems*, K. D. Joshi and Y. Yoo (eds.).

Wikimedia. 2023. "Wikimedia Statistics," available at https://stats.wikimedia.org/#/all-wikipedia-projects, accessed on April 30th, 2023.

Wikipedia. 2023. "Wikipedia: Size of Wikipedia," available at https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia#:~:text=As%20of%2015%20March%202023,of%20all%20pages%20on%20Wikipedia., accessed on April 30th, 2023.

Wöhner, T., Köhler, S., and Peters, R. 2015. "Good Authors= Good Articles?-How Wikis Work," in *Proceedings of the Wirtschaftsinformatik,* Osnabrück, Germany, pp. 872-886.

Yates, D., Wagner, C., and Majchrzak, A. 2009. "Factors Affecting Shapers of Organizational Wikis," *Journal of the American Society for Information Science and Technology* (61:3), 543-554.

Zak, Y., and Even, A. 2017. "Development and Evaluation of a Continuous-Time Markov Chain Model for Detecting and Handling Data Currency Declines," *Decision Support Systems* (103), pp. 82-93.

Zhang, H., Ren, Y., and Kraut, R. E. 2020. "Mining and Predicting Temporal Patterns in the Quality Evolution of Wikipedia Articles," in *Proceedings of the 54th Hawaii International Conference on System Sciences,* Lahaina, HW, pp. 1-10.