Literature Review - WiNLG

Matthew Craig
University of Cape Town
South Africa
crgmat002@myuct.ac.za

1 Abstract

The Wikidata project has been successful in maintaining a vast, language-independent collection of knowledge. This data, alongside natural language generation techniques, can be leveraged to develop an abstract, multilingual Wikipedia. Such an undertaking would promote open collaboration and reduce the issues facing low-resource languages. Existing research has predominantly included on propositions, specifications, and proof-of-concept implementations. Future work must prioritise the development of a working, user-oriented system. The review explores and contextualises research relating to:

(1) The development of an abstract, multilingual Wikipedia.
(2) Established and innovative natural language generation processes.

2 Introduction

This literature review acts as the preliminary investigative work that will inform part of the WiNLG project. WiNLG itself is a component of the broader Abstract Wikipedia project. The Abstract Wikipedia project aims to create a multilingual Wikipedia, whereby content is generated from language-independent representations [16].

The particular sub-project, to which the literature review pertains, involves the development of a tool for managing templates used within the natural language generation (NLG) pipeline.

On account of this, the chosen research papers are divided amongst the following two domains:

(1) Abstract Wikipedia
(2) Natural Language Generation

Papers discussed within each domain are organised chronologically to demonstrate a narrative evolution within both fields. For each paper, a discussion of the following is included:

- Key insights and findings
- Criticisms and shortcomings
- The paper's context within the broader literature
- The relevance it holds to the WiNLG sub-project

This report aims to develop an integral understanding of the current research such that further developments can occur in an informed and directed manner.

2.1 Project Background

2.1.1 Abstract Wikipedia

The goal of Abstract Wikipedia is to leverage Wikidata and Wikifunctions to generate articles abstractly from any individual language [17].

Wikidata is an open collaborative knowledge graph in which entities (items) and their relationships are given unique (language-independent) identifiers [15]. The content in Wikidata can thus be accessed or edited in any supported language.

Wikifunctions (previously Wikilambda) is a collaborative library of functions that perform calculations and answer questions. Wikifunctions can interface with the data in Wikidata [14]. The functions will be used to transform data into its desired form as well as in the NLG pipeline.

Most of the papers covered in the "Abstract Wikipedia" section of this report are direct contributions to the Abstract Wikipedia and Wikidata projects. A few additional papers are included that cover approaches that have attempted to achieve a similar goal.

2.1.2 WiNLG - Template Management Tool

The project "Composing Wikipedia Articles from Wikidata" (WiNLG) consists of multiple sub-projects. This literature review focuses on papers that relate to the sub-project involving the development of a tool for managing templates.

The templates form a key part of the NLG pipeline [**?**]. The realisation of a template should result in the production of a Wikipedia article for a given language. A set of templates must exist for each constructor. A constructor represents abstract, language-agnostic content that is selected from Wikidata. The template management tool will be necessary to allow for the easy creation of templates, which will be used to render the content specified in the constructors. The relationship between constructors and templates will be discussed when reviewing Vrandečić (2014) [15] and Arrieta et al. (2024) [4].

Due to its relevance to the sub-project, a substantial portion of this report has been dedicated to NLG (specifically templatic NLG) research.

3 Abstract Wikipedia

3.1 2009 - Automatically Generating Wikipedia Articles

This section discusses the article "Automatically Generating Wikipedia Articles: A Structure-Aware Approach" (2009) by Christina Sauper and Regina Barzilay [10].

3.1.1 Summary

The paper aims to use a structured machine-learning approach to automatically generate overviews for Wikipedia articles. Sauper and Barzilay (1997) [10] wishes to gather information on a given topic from various internet sources.

The gathered information is summarised by the system in a structured format that adheres to predefined templates. The templates themselves are automatically generated from human-authored documents within a given domain. An example of such a template is

given for the domain "diseases" [10]. The disease template would outline the following sections: Title, Diagnosis, Causes, Symptoms, Treatment. The system would collect excerpts for each section from related internet sources. Finally, text is generated that is both locally relevant and globally coherent.

3.1.2 Criticism

The approach uses an algorithm to rank the relevance of collected excerpts. One of the key metrics used to determine the relevance of an excerpt is search engine ranking. This means that it is probable that incorrect, outdated, or biased information may be selected. There is no mechanism for determining the validity of chosen data. It appears that the paper intends is to showcase the authors' machine learning model, rather than for practical application within Wikipedia.

3.1.3 Context and Relevance

This work [10] does not form part of the Abstract Wikipedia project, which was detailed a decade later [13]. However, the paper does mark the first attempt to systematically generate Wikipedia articles. [10] is particularly relevant in that it highlights the shortcomings of using sources and selection techniques that are not verified or robust. Methods employed by Vrandečić and Krötzsch (2014) [15] as well as Arrieta et al. (2024) [4] avoid these issues.

3.2 2014 - Wikidata: A Free Collaborative Knowledgebase

This section discusses the article "Wikidata: a free collaborative knowledgebase" (2014) by Denny Vrandečić, Markus Krötzsch [15].

3.2.1 Summary

Wikidata is a project by the Wikimedia Foundation. Wikidata's goal is to take the open and collaborative model of Wikipedia and apply it to centralised, structured data. The stated intent of the project is to provide a source of information that is not tied to any specific language [15]. The content represented in Wikidata includes entities, relationships between these entities, as well as various structured properties. The result is an open knowledge graph that can be edited, analysed, and utilised by anyone who wishes to.

The paper was written in 2014, two years after the launch of Wikidata. The paper details Wikidata's rapid growth in users and applications that has occurred within those two years. Wikidata facilitates programmatic interactions with the data by exposing it for use in external applications. Later, in 2015, Wikidata would benefit from more powerful and expressive querying through a SPARQL endpoint [4].

3.2.2 Criticism

Vrandečić and Krötzsch [15] make a strong case for the utility of Wikidata. There is, however, a contradiction between the stated goals and the implications of Wikidata. It is stated that Wikipedia is committed to "a world in which every single human being can freely share in the sum of all knowledge" [15]. Wikidata requires a non-negligible amount of technical experience to understand and interact with the system. This means that the information represented would be biased towards people with technical skills, going against Wikipedia's "every single human" mission. Future work will need to prioritise the development of user-friendly interfaces to further democratise Wikidata.

3.2.3 Context and Relevance

The article emphasises the multilingual design of Wikidata [15]. It is evident that the goals of Wikidata align with those of Abstract Wikipedia [12]. Wikidata is an integral component in the Abstract Wikipedia project. Abstract Wikipedia aims Wikidata as a source of knowledge from which multilingual articles will be generated [17]. One of the authors, Denny Vrandečić, is a key player in the Abstract Wikipedia project and has contributed to many of the related works [12] [13] [17] [14].

3.3 2018 - Toward an Abstract Wikipedia

This section discusses the article "Toward an Abstract Wikipedia" (2018) by Denny Vrandečić [12].

3.3.1 Summary

This paper justifies and explains a means by which Wikipedia articles could exist abstractly, separate from any particular language [12]. It is explained that Wikipedia articles need to be generated from a single source of information to allow people of all languages to read and contribute to Wikipedia. This would eliminate the current issues that plague languages with fewer contributors. In particular, the lack of content and outdated information would be mitigated. The produced articles are intended to prompt readers to contribute further.

A draft architecture for such a system is also included. The system is divided into three components: constructors, content, and renderers. The constructors abstractly capture and relate content fetched from Wikidata. A separate renderer exists for each language and is used to realise the constructor in natural language.

Vrandečić (2018) [12] also addresses alternative strategies such as machine translation. The paper positions a strong defence of the chosen approach.

3.3.2 Criticism

The detailed architecture and implementations are extremely rough and ill-defined. However, the author did not intend for this to represent the final design. The proposition also has similar issues to that of Vrandečić and Krötzsch (2014) [15]. The project aims to "create a novel system" whereby "everyone" can freely collaborate [15], however, the technical complexity of the implementation prevents laypeople from contributing. The WiNLG project is attempting to amend this by producing tools that allow the public to engage with the project.

3.3.3 Context and Relevance

Vrandečić (2018) [12] is a preliminary detailing of what would become known as Abstract Wikipedia. The justification is of particular relevance as it defines goals that must be achieved by WiNLG project contributions. The justifications are developed further in Vrandečić (2019) [13]. The draft architecture discussed would later be expanded upon by Denny Vrandečić (2020) [17].

3.4 2019 - Collaborating on the Sum of all Knowledge Across Languages

This section discusses the article "Collaborating on the Sum of all Knowledge Across Languages" (2019) by Denny Vrandečić [13]. The length of this section is intentionally reduced due to the substantial overlap with Vrandečić (2018) [12]

3.4.1 Summary

The paper features a proposal for an Abstract Wikipedia. Emphasis is placed on the proposal's benefits and justifications, rather than technical specifications. Vrandečić (2019) [13] notes the high volume and expressiveness of knowledge contained within Wikidata. The paper argues that leveraging this data to create a multilingual Wikipedia would prove more effective than relying on organic growth.

3.4.2 Criticism

Vrandečić (2019) [13] includes a rebuttal of the common criticism pointed towards the concept. It is often claimed that such a system would reduce language diversity. Vrandečić counters by noting the past prevalence of fascist content on Croatian Wikipedia. While technically more diverse, Vrandečić argues, the quality and acceptability of the content were reduced due to a smaller pool of contributors.

3.4.3 Context and Relevance

The paper builds upon the propositions of Vrandečić (2018) [12] and develops a substantive foundation for Abstract Wikipedia. Similarly to Vrandečić (2018) [12], the paper cements the intentions that will guide WiNLG.

3.5 2020 - Architecture for a Multilingual Wikipedia

This section discusses the article "Architecture for a Multilingual Wikipedia" (2020) by Denny Vrandečić [17].

3.5.1 Summary

Vrandečić [17] formalises many of the overarching design decisions for Abstract Wikipedia's architecture. The proposed architecture involves the following:

(1) Selection of content from Wikidata
(2) Arrangement and abstract represent of content via constructors
(3) Language-specific renderers for generating articles

The constructors first identified in [12] are explored in greater depth. Vrandečić [17] portrays constructors as a means of building complex abstract representations of content from Wikidata. Constructors are demonstrated to be capable of substantially more expressive and varied representations than would be possible with Wikidata alone. The constructors make particular use of the unambiguous identifiers present in Wikidata, allowing the presented items to exist in a language-independent capacity.

The paper also proposes Wikilambda, a collaborative library of functions. Wikilambda's functions would be able to be utilised by the constructors to produce more complex and insightful information. The functions would be capable of performing calculations, contextualising relationships, and answering questions by operating on Wikidata's vast library of content. Wikilambda would contain language-specific renderer functions that generate articles from the constructors. Wikilambda was later renamed to Wikifunctions [14].

3.5.2 Criticism

The system proposed in Vrandečić (2020) [17] indicates a well-directed vision. The implementation details, however, are left largely unsubstantiated. It is not made clear enough that the given examples for constructors and renderers are not representative of any existing implementation or specification. The "renderers" and the means by which they are expected to convert constructors into natural language ignores the complexities of templatic NLG [9]. Despite these shortcomings, the future specification of Arrieta et al. (2024) [4] demonstrates that, holistically, the architecture in Vrandečić [17] is extremely promising.

3.5.3 Context and Relevance

Vrandečić [17] proposes, demonstrates, and justifies various architecture decisions that greatly influence the direction of Abstract Wikipedia. In the paper, the role and utility of constructors is highlighted. WiNLG's template tool intends to assist in the creation of templates for the realisation (NLG) of these constructors. The renderers discussed within the paper are related to WiNLG's templates, however, the renderers do not accurately portray the intricacies inherent to the NLG pipeline. Vrandečić [17] also introduces Wikifunctions (Wikilambda) which will prove to be a pivotal component in Abstract Wikipedia [4].

3.6 2021 - A Study of the Quality of Wikidata

This section discusses the article "A study of the quality of Wikidata" (2021) by Kartik Shenoy, Filip Ilievski, Daniel Garijo b, Daniel Schwabe, and Pedro Szekely [11]

3.6.1 Summary

Shenoy et al. (2021) [11] assesses the prevalence of low-quality statements in Wikidata. The analysis is performed with Knowledge Graph Toolkit. The following 3 indicators are used to classify statements as low quality:

(1) Community consensus: Statements permanently deleted.
(2) Deprecation: Statements explicitly marked as deprecated.
(3) Constraint violations: Statements Wikidata's property constraints.

The findings that emerge from Shenoy et al. (2021) [11] indicate that there are considerable quality issues with Wikidata, however, this is improving with time. There are millions of constraint violations and redundant entries; although, there is a substantial de-duplication effort in the community. Astronomy-related data had a disproportionately high concentration of deprecated entries.

3.6.2 Criticism

The chosen indicators for classifying low-quality statements are not fully capable of making qualitative judgements about the data they measure. For example, deprecation does not necessarily indicate poor quality. Rather, it is possible that deprecation may indicate an

effort to decrease the prevalence of outdated or incorrect information.

3.6.3 Context and Relevance

Shenoy et al. (2021) [11] is not a component of Abstract Wikipedia. However, the paper is insightful in that it reveals potential shortcomings with the approach outlined by Vrandečić (2018) [12]. If the quality of knowledge stored in Wikidata was notably poor, it could jeopardise the outcomes targeted by Abstract Wikipedia. Low quality data would leave the constructors [16] incapable of producing content that is consistent with Wikipedia's standards.

3.7 2022 - Using Natural Language Generation to Bootstrap Missing Wikipedia Articles

This section discusses the article "Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective" (2022) by Lucie-Aimée Kaffee, Pavlos Vougiouklis ,and Elena Simperl [6]

3.7.1 Summary

Kaffee et al. (2022) [6] details the development and evaluation of a tool named ArticlePlaceholder. ArticlePlaceholder is a tool that leverages NLG techniques to generate single-sentence summaries based on Wikidata. The summaries are generated for articles that do not exist in a given language. The summaries are intended to act as a placeholder for the non-existent article by matching the topic to a Wikidata entry. The generated text acts as a starting point that is intended to "bootstrap" organic contributions from the community. The content is rendered in natural language such that it can be interpreted by casual users. This stands in contrast to the status quo, whereby Wikidata entries are rendered as structured data that is not immediately comprehensible to most readers.

I substantial quantity of Kaffee et al. (2022) [6] is dedicated to a mixed-method evaluation of the merits of the developed system. The following techniques are used to determine ArticlePlaceholder's effectiveness:

(1) Automatic evaluation using standard NLG metrics
(2) Multi-language interviews with Wikipedia editors
(3) Comparison with machine translation baselines

3.7.2 Criticism

The approach taken by Kaffee et al. (2022) [6] is successful in achieving it's stated goals. ArticlePlaceholder is, however, substantially less ambitious than Abstract Wikipedia. Unlike Abstract Wikipedia, ArticlePlaceholder is incapable of developing complex language-independent representations of information. The results of ArticlePlaceholder are, as its namesake implies, only capable of acting as temporary placeholders.

3.7.3 Context and Relevance

While ArticlePlaceholder [6] is not a contribution to Abstract Wikipedia, many parallels between the projects can be drawn. Both leverage Wikidata and templatic NLG to produce language-agnostic Wikipedia articles. The intentions are, however, quite divergent in scope. AbstractWikipedia aims to develop a more expressive and robust means of representing content abstractly.

Kaffee et al. (2022) [6] will prove extremely helpful in guiding the development of the WiNLG's tools. The techniques for evaluating the output of Kaffee et al. (2022) will be highly influential for the template tool's development process.

3.8 2023 - Using Wikidata Lexemes and Items to Generate Text From Abstract Representation

This section discusses the article "Using Wikidata Lexemes and Items to Generate Text from Abstract Representations" (2023) by Mahir Morshed [8].

3.8.1 Summary

Morshed (2023) [8] outlines the specification and development of an NLG system for Abstract Wikipedia titled Ninai/Udiron,. Nina/Udirion consists of two core components:

(1) Nina: Searches Wikidata for items and lexemes for concepts relating to a constructor. It processes the results into syntax trees.
(2) Udiron: Manipulates the syntax trees and converts them into natural language.

Nina/Udiron utilises dependency grammars to represent syntax in a universal way that aligns with Abstract Wikipedia's goals.

3.8.2 Criticism

Nina/Udirion [8] is largely a prototype system. The implementation does not interface with Wikifunctions due to Nina/Udirion being developed before Wikifunction's release. Nina/Udirion proves to be an effective proof-of-concept that successfully proves that complex content can be generated from Wikidata. Nina/Udirion is likely to encourage further interest in Abstract Wikipedia's development.

3.8.3 Context and Relevance

Nina/Udirion [8] differs from the approach that will be used in the WiNLG project. Nina/Udirion does not make use of templates in its NLG pipeline. Instead, Nina/Udirion makes use of a syntax-tree representations for both the constructor and rendering steps. This conflicts with the constructor methodology employed in Arrieta et al. (2024). Despite these differences, Nina/Udirion acts as an exemplary achievement in the space. Nina/Udirion demonstrates that Abstract Wikipedia's objectives are achievable and desirable.

3.9 2024 - CoSMo: A multilingual modular language for Content Selection Modelling

This section discusses the article "CoSMo: A multilingual modular language for Content Selection Modelling" (2023) by Kutz Arrieta, Pablo R. Fillottrani, C. Maria Keet [4].

3.9.1 Summary

Arrieta et al. (2024) [4] species a language for modular content selection named CoSMo. CoSMo is intended for use in the content selection step within Abstract Wikipedia, however, its content selection modelling is non-specific. Thus, CoSMo could be utilised for any system that would benefit from modular information modelling.

The CoSMo language is designed to meet the following criteria:

- Embedded Multilinguality
- A fully modular approach
- Supports both classes and instances
- Declarations can include functions to perform computations on selected content

In the context of Abstract Wikipedia, the modules represented by CoSMo would be equivalent to the constructors described in Vrandečić [17]. CoSMo enables declarative statements about the content that is selected from Wikidata. Declarations can also specify functions from Wikifunctions so that they may compute the data selected from Wikidata.

CoSMo is intended to be platform-independent. This means that the realisation step of the NLG pipeline can occur with whichever method the community converges on. Similarly, CoSMo is applicable regardless of the chosen storage format for Wikidata. CoSMo serves to standardise and formalise Abstract Wikipedia's previously ill-defined constructors.

3.9.2 Criticism

CoSMo's [4] adherence to platform-independence comes with a few drawbacks. CoSMo's non-specificity results in reduced clarity in the necessary implementation process. It is unclear how to content produced by the constructors will interact with subsequent steps in the NLG pipeline.

3.9.3 Context and Relevance

CoSMo [4] marks a distinct progression in the Abstract Wikipedia project. Previously, the constructors were the most illusive component of Abstract Wikipedia. CoSMo is a defined specification for constructors and represents a clear stepping stone in the project. A substantial component WiNLG project will involve the realisation of the vision described in Arrieta et al. (2024) [4].

4 Natural Language Generation

4.1 1997 - Building Applied Natural Language Generation Systems

This section discusses the article "Building Applied Natural Language Generation Systems" (1997) by Ehud Reiter and Robert Dale [9].

4.1.1 Summary

Reiter & Dale (1997) [9] is a comprehensive report on the use-cases, architecture, and implementation of NLG systems as they existed in 1997. Many terms introduced in the paper are still relevant to the field of NLG. NLG is represented as a pipeline whereby modular stages further process information until natural language is produced.

The report details a series of tasks that need to be carried out within an NLG pipeline. Content Determination involves the filtering, summarisation, and selection of content that will be communicated. Discourse Planning involves the ordering and structuring of messages. Messages are categorised and grouped as sentences in the Sentence Aggregation task. Lexicalization is the conversion of messages into words and phrases. Finally, Linguistic Realisation is the application of grammar rules. Each task is followed in a linear fashion (pipeline) to produce natural language as an end result.

Templatic NLG is introduced as an alternative NLG system that avoids Syntactic Realisation. Rather, a template-based approach makes use of pre-defined structures. These structures have empty slots that are filled by selected content.

4.1.2 Criticism

The paper [9] was written nearly three decades ago. On account of this, many of the techniques presented may not align with currently recognised techniques and advancements.

4.1.3 Context and Relevance

Reiter & Dale (1997) [9] represents a foundational work in the field of natural language generation. The paper is frequently cited and has influenced other papers discussed in this literature review [1] [2].

Abstract Wikipedia is set to include an implementation of natural language generation. Reiter & Dale (1997) [9] offers a comprehensive view of the NLG process that will be a valuable aid when contributing to the project. WiNLG's template management tool will benefit greatly from the paper's analysis of template-based approaches.

4.2 2002 - NLTK: The Natural Language Toolkit

This section discusses the article "NLTK: The Natural Language Toolkit" (2002) by Edward Loper and Steven Bird [5].

4.2.1 Summary

Loper & Bird (2002) [5] introduce NLTK, a Python library for processing and generating natural language. Since the publication of this paper, NLTK has become an industry-standard tool. The paper discusses the implementation details and merits of the library.

4.2.2 Criticism

NLTK [5] has been continually updated since the publication of this paper. Due to this, the state of the library depicted in the paper is severely out of date when compared with the modern NLTK.

4.2.3 Context and Relevance

It is likely that NLTK [5] will be utilised in the realisation of templates within the WiNLG process. For this reason, a familiarisation with the library's workings is paramount. AbstractPlaceholder [6] utilised NLTK to generate summarised Wikipedia articles.

4.3 2005 - Real versus Template-Based Natural Language Generation

This section discusses the article "Real versus Template-Based Natural Language Generation: A False Opposition?" (2005) by Kees van Deemter, Emiel Krahmer, and Mariët Theune [1].

4.3.1 Summary

Krahmer et al. (2005) [1] attempts to challenge the reigning consensus surrounding template-based NLG approaches. The paper explains that the previous works such as [9] present templatic NLG as a simplistic and inferior approach. It is demonstrated that developments have led to a merger between traditional and templatic NLG. The combination of techniques is shown to be capable of leveraging a multitude of benefits.

4.3.2 Context and Relevance

Krahmer et al. (2005) [1] acts largely as a rebuttal. The paper argues that the portrayal of templatic NLG given by Reiter & Dale (1997) [9] is outdated and inaccurate.

The paper stands as a justification for the use of templatic NLG within the WiNLG project. Wikidata's structured data [15] is well-aligned to capitalise on the intentional nature of templates.

4.4 2013 - Generating Natural Language from Linked Data

This section discusses the article "Generating Natural Language from Linked Data: Unsupervised Template Extraction" (2013) by Daniel Duma and Ewan Klein [2].

4.4.1 Summary

Duma and Ewan (2013) [2] describe an NLG system designed to realising information contained within RDF triple-stores. The system uses RDF triple stores as the input data source from which templates are automatically extracted. The templates are then converted into natural language. Each triple yields a single sentence that describes the relationship.

An RDF triplestore is a database that stores entries in the subject-predicate-object format. The database can be conceptualised as a directed graph, where:

- The subjects and objects are nodes
- The predicates are labelled, directed edges that connect the subject nodes to the object nodes.

Coincidentally, the paper generates Wikipedia stubs as a proof-of-concept. This is, however, unrelated to any of the work done in the Abstract Wikipedia project [17].

4.4.2 Criticism

The end product of Duma and Ewan (2013) [2], single-sentence natural language descriptions of RDF triple is very simplistic. Future attempts should include graph traversal in the template-extraction processes. This would mean that complex relationships and emergent properties could be described.

4.4.3 Context and Relevance

Wikidata currently stores its knowledge graph in an RDF triplestore whereby each item and property is uniquely identified [15]. Duma and Ewan (2013) [2] utilises entity identifiers in an RDF triplestore to automatically extract templates. Perhaps, this approach could be implemented in WiNLG. Extracting templates from the content specified in a constructor [4] could initiate the template creation process. This may reduce preliminary work for contributors by employing the "bootstrapping" technique outlined in AbstractPlaceholder [6].

4.5 2021 - ToCT: A Task Ontology to Manage Complex Templates

This section discusses the article "ToCT: A Task Ontology to Manage Complex Templates" (2021) by Zola Mahlaza, C. Maria Keet [7].

4.5.1 Summary

Mahlaza and Keet (2021) [7] formalise a systematic specification for templates used in the NLG pipeline. The result is ToCT: A shared conceptualisation (ontology) for template specification across Languages.

Mahlaza and Keet (2021) [7] identify that, in the status quo, template specifications are repeatedly developed in an ad-hoc manner. To remedy this, the paper introduces a formalised task ontology in OWL: a language for authoring web ontologies. To ensure that the resulting ontology is valid, a series of competency questions and SPARQL queries are included. The effectiveness is demonstrated through example templates in English and isiZulu.

4.5.2 Criticism

The paper [7] correctly identifies that the lack of an existing standard is hampering emerging template specifications. It may, however, be possible that no such ontology exists due to the wildly varying implementations of templates. Perhaps, template designs are too dependent on input formats and processes steps in the NLG system. Although, it must be conceded that the existence of a shared conceptualisation is undoubtedly preferable to the status quo. Regardless of a system's exact adherence to the ontology, a great deal of guidance can be gleaned from exposure to such a formalisation.

4.5.3 Context and Relevance

ToCT [7] aims to formalise the specification of NLG templates. The merits of a templatic approach were argued for by Krahmer et al. (2005) [1]

WiNLG's template management tool would benefit greatly from adhering to ToCT [7]. The multilingual templates utilised within the project intend to realise complex natural language. Following a ToCT's standardised approach would ensure consistency and adaptability across various languages. The clear direction outline in ToCT [7] may also facilitate more powerful features within the template management tool.

4.6 2023 - Using Dependency Grammars in Guiding Templatic Natural Language Generation

This section discusses the article "Using Dependency Grammars in Guiding Templatic Natural Language Generation" (2023) by Ariel Gutman, Anton Ivanov, and Jessica Saba Ramírez [3].

4.6.1 Summary

Gutman et al. (2023) [3] proposes a templatic NLG system with integrated dependency grammars. Dependency grammar annotations are added to the templates to ensure linguistic accuracy and grammatical validity of the output. The templates in the proposition are capable of containing dynamic and static elements.

The Universal Dependency framework would be used to implement the system's dependency grammar implementations. As a result, the grammatical specifications are mostly shared between languages.

4.6.2 Context and Relevance

The system proposed by Gutman et al. (2023) [3] is reminiscent of the combined approach described in Krahmer et al. (2005) [1]. Dependency Grammars are being utilised to mitigate some of the downsides inherent to a templatic NLG system.

An implementation centred around dependency grammars was followed by Nina/Udirion [8]. Perhaps, following the proposal of Gutman et al. (2023), a future version of Abstract Wikipedia could leverage both template and dependency grammar oriented NLG.

5 Conclusion

This literature review has involved the analysis of existing research relating to Abstract Wikipedia and natural language generation.

Exploration of articles relating to Abstraction Wikipedia has revealed that there is a wealth of existing content on the subject. However, the existing research has predominantly involved proposition, discussion, and experimentation. The WiNLG project intends to use the existing research to move the Abstract Wikipedia forward. WiNLG will involve the development and implementation of tools that capitalise and improve upon current findings.

The NLG articles have successfully contextualised the historical and contemporary state of natural language generation. There are a variety of valid techniques, however, it is evident that a template based approach will work best with Wikidata's structured approach. The varied approaches demonstrated throughout the research will prove highly beneficial in the implementation of the NLG pipeline.

References

[1] Kees van Deemter, Mariët Theune, and Emiel Krahmer. 2005. Real versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics* 31, 1 (03 2005), 15–24. https://doi.org/10.1162/0891201053630291 arXiv:https://direct.mit.edu/coli/article-pdf/31/1/15/1798168/0891201053630291.pdf

[2] Daniel Duma and Ewan Klein. 2013. Generating Natural Language from Linked Data: Unsupervised template extraction. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, Alexander Koller and Katrin Erk (Eds.). Association for Computational Linguistics, Potsdam, Germany, 83–94. https://aclanthology.org/W13-0108

[3] Ariel Gutman, Anton Ivanov, and Jessica Saba Ramírez. 2022. Using Dependency Grammars in guiding templatic Natural Language Generation. https://research.google/pubs/using-dependency-grammars-in-guiding-templatic-natural-language-generation/

[4] Pablo R. Fillottrani Kutz Arrieta and C. Maria Keet. 2024. CoSMo: A multilingual modular language for Content Selection Modelling. *ACM/SIGAPP Symposium on Applied Computing (SAC '24)* 39 (2024). https://doi.org/10.1145/3605098.3635889

[5] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. (2002). https://doi.org/10.48550/arXiv.cs/0205028 arXiv:cs/0205028 [cs.CL]

[6] Kaffee Lucie-Aiméea, Vougiouklis Pavlosb, and Simperl Elenac. 2022. Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective. *Semantic Web* 13 (2022). https://doi.org/10.3233/SW-210431

[7] Zola Mahlaza and C. Maria Keet. 2021. ToCT: A Task Ontology to Manage Complex Templates. In *Joint Ontology Workshops*. https://api.semanticscholar.org/CorpusID:240005311

[8] Mahir Morshed. 2023. Using Wikidata Lexemes and Items to Generate Text from Abstract Representations. *Semantic Web* (2023). https://www.semantic-web-journal.net/content/using-wikidata-lexemes-and-items-generate-text-abstract-representations-0

[9] EHUD REITER and ROBERT DALE. 1997. Building applied natural language generation systems. *Natural Language Engineering* 3, 1 (1997), 57–87. https://doi.org/10.1017/S1351324997001502

[10] Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: a structure-aware approach. (2009), 208–216. https://doi.org/10.3115/1687878.1687909

[11] Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of Wikidata. *Journal of Web Semantics* 72 (2022), 100679. https://doi.org/10.1016/j.websem.2021.100679

[12] Denny Vrandečić. 2018. Toward an Abstract Wikipedia. *International Workshop on Description Logics* 31 (2018). http://ceur-ws.org/Vol-2211/#paper-03

[13] Denny Vrandecic. 2020. Collaborating on the Sum of All Knowledge Across Languages. (10 2020). https://doi.org/10.7551/mitpress/12366.003.0016 arXiv:https://direct.mit.edu/book/chapter-pdf/2247832/9780262360593_c001200.pdf

[14] Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Commun. ACM* 64, 4 (mar 2021), 38–41. https://doi.org/10.1145/3425778

[15] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (sep 2014), 78–85. https://doi.org/10.1145/2629489

[16] Denny Vrandečić. 2020. Architecture for a multilingual Wikipedia. (2020). https://doi.org/10.48550/arXiv.2004.04733 arXiv:2004.04733 [cs.CY]

[17] Denny Vrandečić. 2020. Architecture for a multilingual Wikipedia. https://doi.org//10.48550/arXiv.2004.04733 arXiv:2004.04733 [cs.CY]