UNIVERSITY OF LONDON

Goldsmiths
UNIVERSITY OF LONDON

MSc Data Science

DSM500 Final Project Report

Lisa Mitford

Automatic knowledge graph construction from news articles

Table of contents

Figures

Tables

This project started life as a literature review and project proposal for DSM060 Data Science Research Topics and has been substantially developed upon during the course of the DSM500 Final Project module. The final report includes parts that are either informed by work done previously, or includes sections from work done previously.

I am grateful for the support and encouragement of the following people during this project:

Foaad Haddod, Kosma von Maltitz, Sarah Rauchas, Noureddin Sadawi,
Bruce Bassett, Barry Benjamin, Denise Rack Louw, Ian Burgess-Simpson,
Tobie Vermeulen, Adriaan Basson, Kyle Cowan, Adele Hamilton, Ewald Zietsman

And in loving memory of *Geoffrey Louw* and *Othello Mitford*:
you started this journey with me, and I know you would
have been thrilled to see me in sight of the finish line!

Abstract

Knowledge graphs (KGs) can be highly beneficial to the news industry, particularly for information retrieval applications like search, summarization and question answering – with the potential to improve both accuracy and explainability of results. However, the practicalities of automating KG construction from news articles remain underexplored in the current literature. This project proposes a novel approach – including the use of human-in-the-loop (HITL) techniques – to address this gap; and demonstrates its feasibility by constructing a KG from 2081 articles on the topic of the *Zondo Commission* sourced from News24, a South African news site. Core components of the proposed model, based on design science methodology, include: 1) use of open-source components as the foundation; 2) creation of a partial gold-standard dataset using HITL annotation to enable evaluation during the development phase; 3) use of subject-matter experts for the final evaluation on a sample of the KG.

While it was found that open-source models can be used as a starting point for such a project, they were insufficient on their own. Named entity recognition is a mature task and achieved excellent F1 scores > 0.9 in all cases. However, the best baseline F1 score for coreference resolution was just 0.71688 – rather lower than expected. For relation extraction, the best baseline achieved was an F1 score of 0.53658 – very poor compared to reports in the literature. Furthermore only one of the entity linking models tested proved viable. Significant custom development using natural language processing techniques was required to transform the outputs from the open-source models into usable inputs for the KG algorithm. Final results indicate that further development of rule- and pattern-based semi-supervised methods as well as vector similarity methods, would be necessary to produce a KG of sufficient quality.

Use of an HITL partial gold-standard dataset as well as evaluation by subject-matter experts were both found to add value to the project. These methods ensured that decisions at each stage were based on quantitative assessment. They also highlighted promising directions for further exploration of the design search space during each iteration of the build-and-evaluate cycle. In addition, involving the editorial team surfaced requirements that could enhance the KG's utility for both readers and journalists.

This project contributes to clarifying the challenges in automating KG construction from news articles and provides a preliminary roadmap for future efforts in this area.

Keywords: knowledge graph; automatic knowledge graph construction; information extraction; journalistic knowledge platform

1 Introduction

The term *knowledge graph* (KG) came into focus in 2012 with Google's announcement of the Google Knowledge Graph "that understands real-world entities and their relationships to one another: things, not strings" (Singhal). KGs use the graph data model to express information about a domain of interest, where nodes represent entities and edges represent relations between those entities.

Since 2012 KGs have found many applications in industry. In the news industry specifically, several interesting use cases have been explored – including semantic annotation, event detection, relation extraction, and even fake news detection (Opdahl, et al., 2022). The focus of this project is on automating the extraction of entities and their relations from news articles into a KG suitable for enhancing reader-facing information retrieval (IR) processes.

1.1 Problem definition

While much has been written on the topic of KGs in the news industry, only 11% of the papers in a recent survey covered relation extraction, and many of these focused on niche areas or used public datasets (Opdahl, et al., 2022). As a result relatively little is known about the real challenges likely to be faced when building this type of KG from news articles in general, or about potential best practices that could be applied to the task.

News articles are rich sources of information but extracting that information "with high precision and recall for the purposes of creating or enriching a knowledge graph is a non-trivial challenge" (Hogan, et al., 2021, 6.2). Furthermore, newsrooms with a modest budget will want to use open-source tools as far as possible – and while doing so should reduce development time and cost, it may also impose constraints on the design. Opdahl, et al. conclude: "Pilot projects that match high-benefit tasks with low-risk technologies and tools are therefore essential to successfully introduce semantic KGs in newsrooms." (2022)

This project investigated the extent to which, by using a selection of the currently available low-risk technologies, automatic KG construction from news articles can be achieved. The methods outlined suggest a roadmap which can act as a starting point for media houses wanting to automate KG construction using their own articles as inputs.

To ringfence the scope of the project, articles about the *Zondo Commission* produced by news site News24 were used. Officially known as the *Judicial Commission of Inquiry into Allegations of State Capture*, the purpose of the commission was "to investigate allegations of state capture, corruption and fraud in the Public Sector including organs of state" in South Africa (President of the Republic of South Africa, 2018). It is a good

test case for automatic KG construction as it covers a broad range of entities and relationships in government, business, politics and the judiciary, reported on over a period of 4 years.

1.2 Motivation

With the advent of large language models (LLMs), consumers increasingly expect to engage with online news sites in highly personalised and specific ways. News outlets like *Forbes*, the *Financial Times*, and *Bild* have all recently introduced generative AI tools for interacting with their content Q&A style – and other media houses are either following suit, or investigating the possibility. However, the trustworthiness and explainability of such systems is crucial, and to get the best results one "must be able to integrate sub-symbolic deep learning, symbolic knowledge representation and logical reasoning" (Ocaña & Opdahl, 2023) – which is where KGs come in.

KGs allow for the representation of knowledge in both human- and machine-readable format, which in turn facilitates reasoning and inference capabilities (Hogan, et al., 2021, 1). This can be particularly useful when it comes to answering so-called multi-hop questions, where the real question can only be answered by first answering one or more intermediate questions (Peng, et al., 2023). Consider the scenario in *Figure 1*: the best retrieval outcome is where one first obtains synonyms associated with the entities *before* using more traditional vector-based search techniques to retrieve the relevant content with which to answer the question:

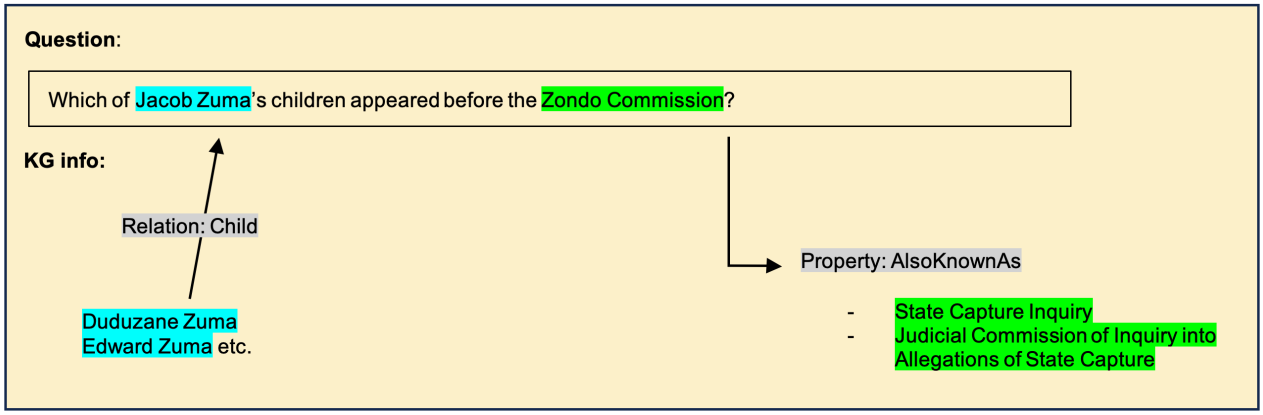

Figure 1: Multi-hop question scenario where best results first obtain synonymous terms

Public knowledge graphs such as Wikidata or DBPedia only cover a subset of entities reported on by media. For this reason news outlets wanting to incorporate KG technology into semantic search or retrieval augmented generation (RAG) processes will need to construct KGs based on their own content. And given the volumes that are typically involved, such KGs will need to be automatically constructed.

How best to proceed with such an initiative remains unclear. As Opdahl, et. al point out, there is "a double challenge for newsrooms that want to use KGs for news: It is usually not known how robust the proposed

techniques and tools are in practice, and it is usually not known how well they fit actual industrial needs" (2022). This project investigates a subset of those proposed tools and techniques in detail, in order to quantify their effectiveness individually, and to evaluate their contribution to the overall process.

1.3 Key concepts

Figure 2 illustrates how a KG can be used to encode knowledge.

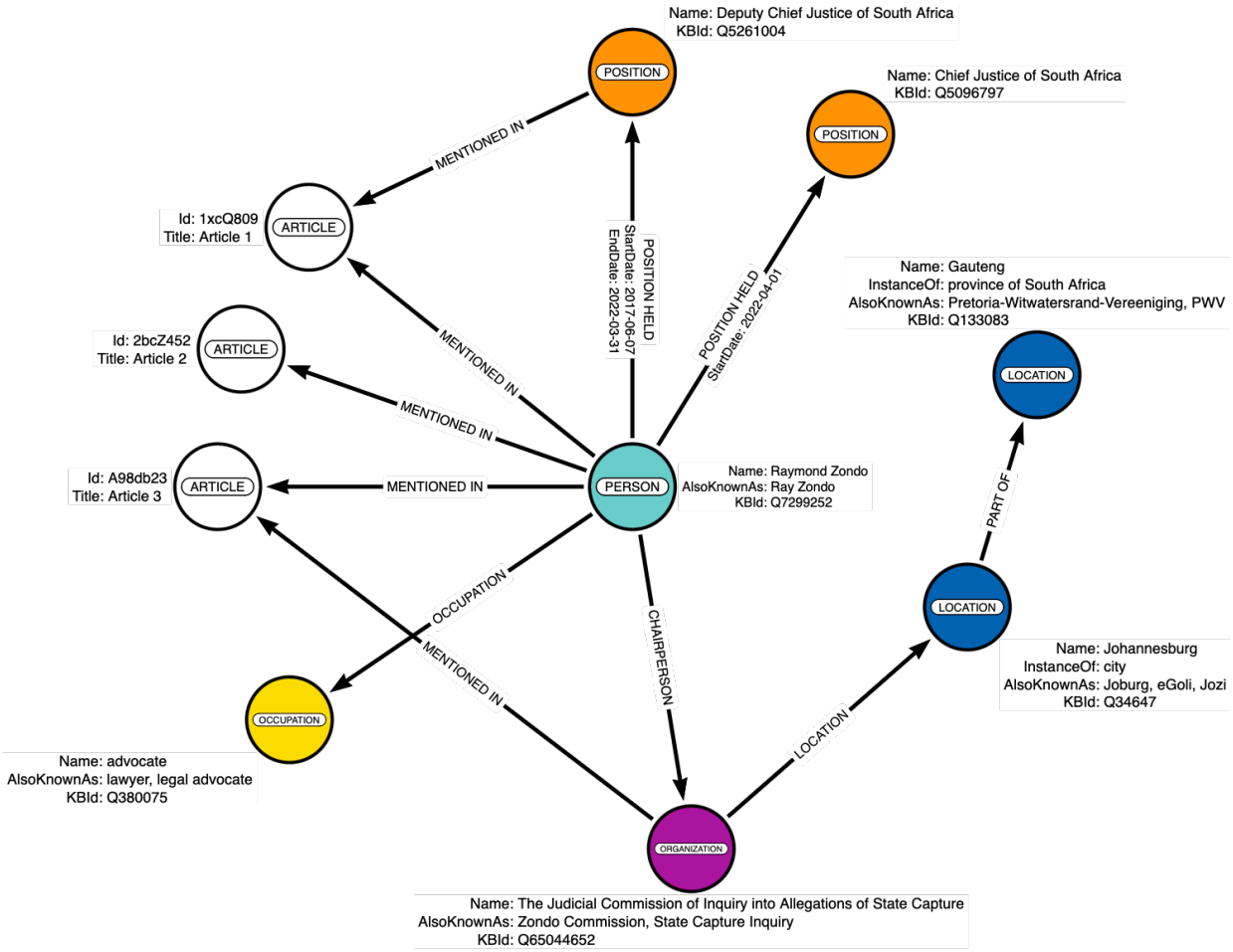

Figure 2: Envisaged sample output of a news KG with Wikidata KBIds where available

The **nodes** represent 'things' including physical objects, entities, or even abstract concepts. Each node is defined by its **type**, also known as a **label**, for example *POSITION*. The links (or **edges**) between nodes represent the relationship between them, for example *POSITION HELD*. Edges are not only labelled but also **directed**. Node names, labels, and directed edges represent the minimum information required to build the KG. However, both nodes and edges may include additional **properties** like *AlsoKnownAs*. And where possible

nodes may be **linked** to public KGs like Wikidata (see *KBId*) which enables further **enrichment** from the properties maintained in the public KG such as *StartDate* and *EndDate*.

Information extraction (IE) is the process by which unstructured text is transformed into structured information as depicted above. For example a paragraph like "*Chairperson Raymond Zondo presided over the first hearing of the State Capture Inquiry today. Appointed as Deputy Chief Justice in South Africa in 2017, he is the ideal leader for this investigation, which promises to be highly contentious.*" should yield the nodes and edges highlighted in yellow in *Figure 2*.

This project used the **design science** methodology in which "knowledge and understanding of a problem domain and its solution are achieved in the building and application of the designed artifact" (Hevner, et al., 2004). By exploring the design search space, successive iterations of the build-and-evaluate loop clarify which methods show promise and which do not, and surface important considerations when undertaking such a project.

2 Literature review

Many different models and methods have been proposed to achieve automatic KG construction. A recent survey by Tamašauskaitė & Groth identifies as many as 414 possible unique tasks described across 57 papers. Core themes recur across these tasks, however – in particular **ontology** definition, **information extraction**, and **knowledge processing** (2023).

An **ontology** is the "organizing principle" that underpins a KG (Barrasa & Webber, 2023, p. 23): it formalizes the representation of knowledge at a high level and encodes rules about how information is stored and what forms it can take. If we consider the example in *Figure 2* in section *1.3*, an ontology might define that where a relation like *POSITION HELD* exists the holder of the position is a *PERSON* (which is an instance of the class *HUMAN*). Unlike traditional databases, graph databases do not enforce a schema (Barrasa & Webber, 2023, p. 18). However, for information retrieval purposes, query expansion (addition of synonymous terms to a search string) and deductive reasoning (extrapolation of new relations inferred by existing ones) are two important applications that can be enabled by a robust ontology (Hogan, et al., 2021, 4.2).

There are several large-scale open ontologies like the Dublin Core Metadata Initiative (DCMI) and schema.org (Barrasa & Webber, 2023, p. 46), and public KGs like Wikidata also maintain ontologies which can be leveraged (Wikidata, 2024).

It is worth mentioning the BBC has opted to develop and maintain their own ontologies (BBC, 2023), however, ontology engineering is an intricate and nuanced undertaking which requires a strong justification. Using an existing ontology is preferred if the chosen domain fits an available standard (Barrasa & Webber, 2023, p. 46). Several news-specific ontologies are publicly available in addition to the BBC's, such as the NEWS ontology (Fernández, et al., 2010) and SNaP (PA Media, n.d.). These include useful properties that reflect the way journalists think, for example *notablyAssociatedWith* from SNaP (Wilton, et al., 2012). However, extracting relations like these would either require building a custom relation extraction model, or encoding rule-based extraction techniques on top of existing models, which could add complexity to an already complex task.

More than one ontology can be overlaid onto a KG structure (Barrasa & Webber, 2023, p. 42) which does provide flexibility to fulfil different requirements post-construction.

The **information extraction** (IE) step is most frequently proposed as a pipeline that makes use of natural language processing (NLP) techniques to extract structured data from unstructured text (Hogan, et al., 2021, 6.2). The exact composition of the pipeline may vary but the two core tasks that most sources agree on are named entity recognition (NER) and relation extraction (RE).

NER identifies possible nodes for inclusion in the KG by tagging entities and their types, for example *Raymond Zondo >> PERSON*. Weikum, et al. suggest starting with more "premium" data sources for node identification such as Wikipedia or WordNet (2021). However, for the situation that most news outlets face – where the scope of coverage is unquantified and many of the entities mentioned will not exist in public knowledges bases – it makes more sense to start by identifying entities from the corpus itself, particularly since deep learning NER models yield good results on this task (Li, et al., 2022).

RE identifies possible edges or relations between entities, where the "simplest case is that of extracting binary relations in a closed setting" (Hogan, et al., 2021, 6.2.4), such as *Raymond Zondo >> position held >> Deputy Chief Justice*. Such closed models will of course only extract the relations they were originally trained on, and therefore the choice of model is a fundamental constraint on the finished product: if the text contains relations that fall outside the scope of that model they will be overlooked. On the other hand a closed system results in a standardized ontology which, as already mentioned, can improve retrieval and reasoning capabilities (Mahouachi & Suchanek, 2020).

Pattern- and rule-based methods are discussed at length by Weikum, et al. They claim even simple manual patterns can perform quite well, but for wider recall suggest "seed-based pattern learning and statement extraction" techniques (2021). The drawback here is that such methods require considerably more development time compared to using pre-trained models.

Using LLMs to extract triples has been tried but so far yields rather inconsistent results and has proved highly dependent on prompt engineering. In addition it remains unclear whether triples come from memorized training data or the actual text provided (Pan, et al., 2023).

Some models combine tasks. A promising F1 score of ≈0.91 was obtained on the New York Times (NYT) dataset with an encoder-decoder variant called Grapher (which even includes the KG generation step), but the model does not scale well (Melnyk, et al., 2022). REBEL-large combines NER and RE, yielding triples that identify both entities and the relations between them (Cabot & Navigli, 2021). A shortcoming of this model is that the entity type is not output. Ensembling the results of different pipeline components as suggested by Al-Moslmi & Ocaña is one approach that could overcome this issue, and in general could yield an improved overall result (2020).

The outputs of NER and RE are likely to be noisy and inconsistent, hence the **knowledge processing** step where the aim is to improve the quality of the KG and standardize wherever possible (Tamašauskaitė & Groth, 2023).

Canonicalization, or uniquely identifying entities is an important consideration here. In most texts synonymous terms (like *African National Congress* and *ANC*, or *Jacob Zuma* and *Zuma*) will be used interchangeably; and

ambiguous terms like *East London* (which may refer to a region of the city in the United Kingdom or a town in South Africa) will occur. A KG where synonymous terms are not normalized is still useful – but clearly it is preferable they are recognized as such (Weikum, et al., 2021).

Entity linking (EL) is frequently proposed as one means for disambiguating entities. EL takes in a mention of an entity in context, finds candidate nodes in a public knowledge graph such as Wikidata or DBPedia, and returns the node with the highest probability of being a match, given that context (Hogan, et al., 2021, 6.2.3). This linking has the added benefit of enabling enrichment of the KG from public data sources. However, as already mentioned, many entities will not be found in public KGs.

Coreference resolution (CR) is another tool proposed for aiding disambiguation. CR groups mentions of entities into clusters in order to "disambiguate the input text and replace pronouns and acronyms with its associated entity mention" (Jaradeh, et al., 2023). It has the added benefit of potentially bolstering the results obtainable from RE. Consider the sentence "*He is chairperson of the Zondo Commission*": being able to replace *He* with *Raymond Zondo* may identify a new relation which was not previously discovered.

Statistical methods may further enhance the quality of knowledge extracted – the simplest option being to consider relation frequencies when deciding whether to include a data point in the KG or not (Weikum, et al., 2021). But of course this method would only work well on larger corpora with many references to the same entities or relations – and runs the risk of excluding valid data points with few mentions.

Beyond these core tasks, **architectural guidelines** on knowledge engineering in general, and for the news industry in particular are informative. Allen & Ilievski stress the importance of defining and understanding the use case, and testing against that use case, to ensure the final product design and implementation is fit-for-purpose (2024). Ocaña & Opdahl describe requirements for a complete journalistic platform. Very relevant to the KG component of that platform is the question of provenance, or knowing where information is sourced from (2023). Both guidelines agree on the importance of key qualities like modularity and scalability. The News Hunter prototype which was designed to update a KG from multiple high-velocity, high-volume sources including social media platforms and aggregators like Reuters and AFP is reported to scale well (Berven, et al., 2020) – and the volumes involved in automating KG construction from in-house articles are much smaller, therefore scalability should be achievable.

Figure 3 (overleaf) summarizes the core components that may form the backbone of a typical KG pipeline. When **selecting models** to use for each component, the premise of work done on the PLUMBER system is that the best options will depend very much on the type of input text (Jaradeh, et al., 2023). It follows then that a reasonable approach to component selection and subsequent method refinement would be to evaluate performance on a sample of actual articles from the news corpus in question, rather than relying on

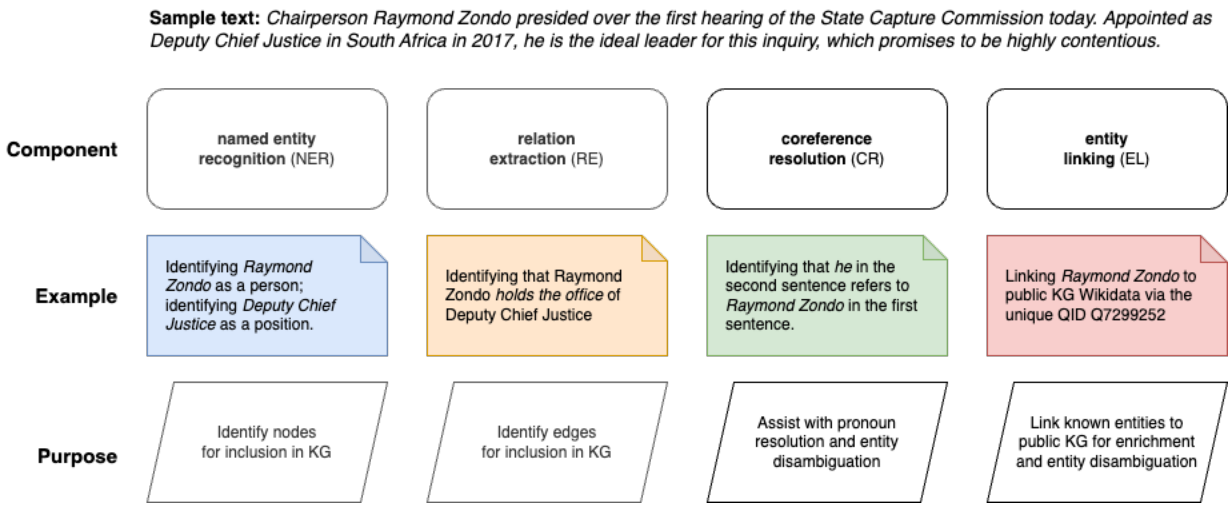

Figure 3: NLP components that form the backbone of the KG pipeline and examples of their use in the overall process

benchmarking datasets like the New York Times dataset (Riedel, et al., 2010). In fact using a partial gold-standard dataset is a popular choice for KG evaluation, even though these datasets tend to be small since they are time-consuming to produce (Paulheim, 2017). Using "human-in-the-loop" (HITL) techniques (Wu, et al., 2022), where initial model labels just need to be edited, can substantially reduce the work effort required to produce such a dataset. However, one should also bear in mind that this method can suffer from the cognitive bias known as WYSIATI (what you see is all there is) described by Kahneman (2011): it is much easier to evaluate existing labels than it is to spot missing labels.

Turning to the question of **evaluation**: assessing the quality of the full KG is challenging as, by definition, no ground truth labels are available. Weikum, et al. note the ideal KG "would have no errors, comprise all entities and types of interest, contain expressive statements that cover all aspects of interest, and ... be fully automated". They go on to state that this constitutes "wishful thinking": as with many machine learning projects trade-offs will be required (2021).

The primary metrics used for KG evaluation are precision (correctness) and recall (coverage). In the absence of ground truth labels, evaluating a sample of data points from the KG using subject-matter experts is suggested by more than one source (Paulheim, 2017), (Weikum, et al., 2021), and is a viable option that should be available to most newsrooms producing original content. Depending on use case requirements one may favour precision or recall in the design. Alternatively to strike a balance the harmonic mean of the two, or F1-score, is suggested.

Another important metric to bear in mind is efficiency: to function at production scale the KG components should operate within acceptable time and cost parameters, given expected volumes.

Finally, the **design science** paradigm is an ideal framework in which to explore how these many considerations come into play (Hevner, et al., 2004). An initial design search space is defined, with each iteration of the build-and-evaluate loop suggesting how model outputs can be refined and knowledge processing techniques employed to produce the final artifact: a KG automatically constructed from news articles.

3 Aim

The overall aim of the project was to produce a viable artifact in the form of: 1) a detailed model and methods for approaching KG construction from news, and 2) an instantiation of a news KG through which the proposed model and methods could be evaluated (Hevner, et al., 2004).

3.1 Objectives

1. Creating a small gold-standard dataset for evaluation purposes as suggested by Paulheim (2017) was the key starting point for this project. A sample of 30 articles were annotated for NER, RE and CR – using HITL techniques to reduce annotation effort (Wu, et al., 2022). This provided a concrete baseline against which to assess performance at each iteration.
2. Defining ontology requirements, using an available standard as recommended by Barrasa & Webber (2023), was the next key step in order to ensure standardization of entity and relation types.
3. Components for the initial KG pipeline were selected from affordable, open-source models using results obtained on the HITL dataset. Doing so ensured selection was based on performance on the actual corpus as suggested by Jaradeh, et al. (2023).
4. Subsequent iterations focused on improving baseline model results. 10 of the HITL articles were used for investigation and development, while 20 were reserved to assess if improvements were realised.
5. Thereafter knowledge processing techniques, including canonicalization (Weikum, et al., 2021) and EL (Hogan, et al., 2021, 6.2.3), were used to normalize and enrich the KG data points while constructing the KG.
6. Finally, a sample from the built KG instantiation was evaluated with the help of subject-matter experts (Paulheim, 2017), (Weikum, et al., 2021), thereby enabling evaluation of the quality of the finished product, as well as the model and methods used (Hevner, et al., 2004).

3.2 Research questions

Key questions considered during the project included:

RQ1. To what extent were available "out-the-box" models for each pipeline component suitable? And if there were shortcomings, to what extent could they be addressed, and how feasible was it to implement these solutions?

RQ2. Was the creation of a HITL gold-standard dataset feasible, and effective in assessing and improving model performance?

RQ3. Did the use of subject-matter experts meaningfully add to the evaluation process?

RQ4. Did the final proposed model and its methods produce a KG that was fit-for-purpose, and to what extent could these techniques inform how the news industry approaches automatic KG construction?

4 Methodology

This section outlines the data used and the methodology followed at a conceptual level. The final methods used in automatic KG construction were very much informed by results and observations emerging from each stage of the build-and-evaluate loop and are therefore discussed together in detail in section *5*.

4.1 Source data

The source data consisted of 2081 articles in English on the topic of the *Zondo Commission* published between 2018 and 2022 by News24. Permission was granted by Media24 for the articles to be used for the project, and a sample of publicly available content is available (see Appendix A). *Table 1* shows that articles fell into 3 categories: general reports, editorial analysis, and opinion pieces.

	Length in Sentences				
Category	**Median**	**Mean**	**Min**	**Max**	**# Articles**
General	28	34	2	300	1551
Analysis	52	76	2	1434	142
Opinion	59	70	4	1943	388

Table 1: Articles used for the project
showing the 3 main categories and key features

Opinion articles included guest columnists' writings (and therefore potentially more biased or emotional content). Evidence of the effect of these was seen in the final KG in relations like *state capture looting on an unparalleled scale >> participant >> Transnet*. For use cases where objectivity is important it may be preferred to exclude opinion content.

Table 2 shows the data points available for each article, along with their type and description:

Data point	**Type**	**Description**
ArticleId	String	Unique identifier per article
Permatitle	String	URL fragment, unique string identifier per article
PublishedDate	Date	Date article was published
Breadcrumb	String	Path to article on site (a proxy for category)
IsLocked	Bool	Subscriber-only content behind the paywall
Title	String	Article headline
Synopsis	String	Short summary of the content
Body	String	Main article content

Table 2: Data points available for articles
where Title, Synopsis and Body formed the key inputs

A concatenation of *Title*, *Synopsis* and *Body* were used as source inputs for extracting information. *ArticleId* and *Permatitle* were used to track the provenance of entities and relations as recommended by Ocaña & Opdahl (2023). The data was of high-quality as it is customer-facing content, but as expected with most corpora there were occasional spelling and grammar issues. Although dealing with these was beyond the scope of the project, from the point of view of canonicalization it would be recommended that post-processing steps be implemented to normalize variations that occur because of spelling issues like these: *state capture enquiry*, *state capture inquiry*. Weikum, et al. propose entity matching using similarity functions like Jaro-Winkler or Levenshtein as one possible solution (2021).

4.2 Candidate pipeline components

Rather than selecting components based on the best reported scores in the literature, two components for each task were identified for evaluation against the HITL dataset to determine which would perform best on this corpus as suggested by Jaradeh, et al. (2023). The components listed in *Table 3* were chosen based on a combination of accessibility and/or performance.

Task	**Component (metric if available)**	**Metric reference**
NER	flair/ner-english-large (F1 94.36)	(Schweter & Akbik, 2020)
	spaCy en_core_web_trf (F1 90.00) *	(spaCy, n.d.)
CR	fastcoref (avg F1 78.50) *	(Otmazgin, et al., 2022)
	LingMess (avg F1 81.40)	(Otmazgin, et al., 2023)
RE	Babelscape/rebel-large (F1 93.4 on NYT) *	(Cabot & Navigli, 2021)
	Flair Tagging Relations	
EL	OpenTapioca (uses Wikidata)	
	EntityLinker (uses Wikidata)	

Table 3: Components investigated for use in the final pipeline
** HITL annotations were based on indicated model (section 4.4)*

Setting up the experiments to test each of these 8 different libraries against the HITL dataset was time-consuming. However, as can be seen in section *5.1*, it ensured a good understanding of the implications of selecting one model over another in terms of quality and time/cost.

4.3 Ontology requirements definition

Initial assessment indicated that leveraging the Wikidata ontology (2023) would make sense, given that REBEL outputs Wikidata triples (Cabot & Navigli, 2021), that Flair relations could be mapped to Wikidata properties in all cases (with the exception of *alternate_name* which maps to Wikidata's *Also known as* data point), and

that the candidate EL options both use Wikidata as source. In a production environment this decision would require monitoring: Wikidata is a collaborative, evolving project and changes to the schema can occur over time (Weikum, et al., 2021).

An in-depth analysis was conducted as outlined in *Figure 4* (overleaf) to determine the final ontology used as the basis for the KG. REBEL was trained on Wikipedia text and Wikidata triples and claims to have kept "the 220 most frequent relations in the train split" (Cabot & Navigli, 2021). Flair was trained on "a modified version of TACRED" (Flair, 2024). The original TACRED paper lists 41 possible relations (Zhang, et al., 2017), however in practice outputs from Flair included several additions. In both cases it was difficult to obtain a definitive list of all possible relations that could be expected. Therefore RE was performed on a sample of 500 articles using both Flair and REBEL. The result was a list of relations and their frequencies from each model, which formed the basis for further review: only those relations deemed useful for IR purposes were included. For example, conceptual relations like *corruption >> facet of >> state capture* were excluded, as well as those deemed too high-level to be useful, like *Cyril Ramaphosa >> instance of >> human*.

To compare model performance later on, relations found by each model were mapped to each other. Identifying and completing inverse relations was also required to make comparisons. For example, Flair outputs "*owns"* as in *Atul Gupta >> owns >> ANN7*, whereas REBEL outputs "*owned by"* as in *ANN7 >> owned by >> Atul Gupta.* All Flair relations were covered by a corresponding REBEL relation, except for Flair's *alternate_ name* relation.

Obtaining inverse properties for Wikidata relations was automated using the Pywikibot library (Pywikibot, 2024), as was obtaining subject and object constraints for each relation. Results were manually reviewed – and a few corrections made to ensure accuracy and integrity.

The end-result was a list of 61 main relations identified for inclusion (with subject and object constraints), 47 of which had inverse relations, and just 20 of which were shared by both models. Although not a formally defined "ontology" using an ontology language, it nonetheless served as the basis to define which relations would be mapped into the KG and what type constraints should be applied to each to ensure standardization – fitting Barrasa & Webber's pragmatic suggestion "that a mix of standards and adaptive customization is most aligned with the reality of modern business" (2023, p. 47).

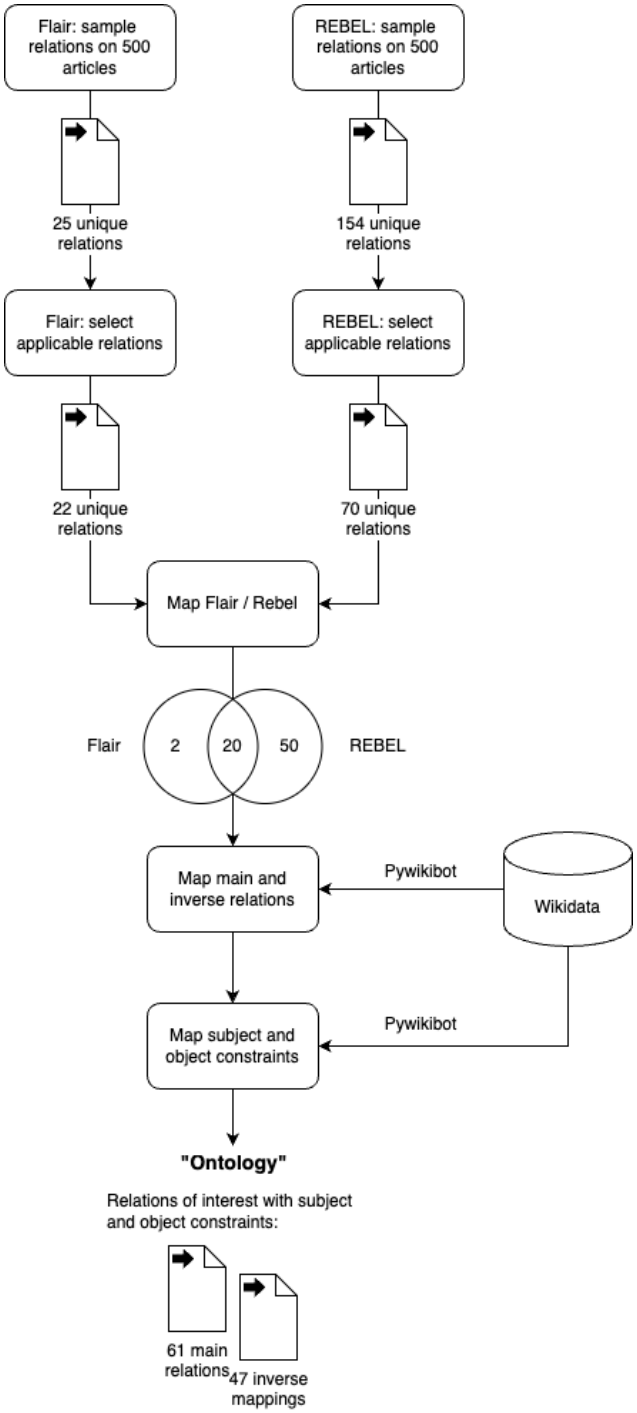

Figure 4: Preparatory steps for ontology definition to ensure alignment of models, completion of inverse relations, and subject and object type constraints

4.4 HITL dataset

The HITL gold-standard dataset was created using a stratified sample of 30 articles, ensuring a representative mix of the three main types of article shown in *Table 4* – since the writing styles, content, and length tend to be quite different across these 3 categories:

Category	*Length in sentences* **Median**	**# Articles in sample**
General	27	22
Analysis	49	2
Opinion	51	6

Table 4: HITL gold standard sample detail after stratified sampling was performed

30 articles was chosen as a minimum to ensure the task was achievable within a reasonable timeframe.

Labels were generated using outputs from the initial models selected (section *4.2*). These were imported into Label Studio for revision and correction. The final annotations from Label Studio formed the HITL dataset used for evaluation.

4.4.1 Named entity recognition annotation

Annotating 30 articles took 3 hours. *Figure 5* shows part of an article annotated for NER. The process was relatively straightforward, but required some decisions such as whether to include articles like *the* and titles like *Mr.*

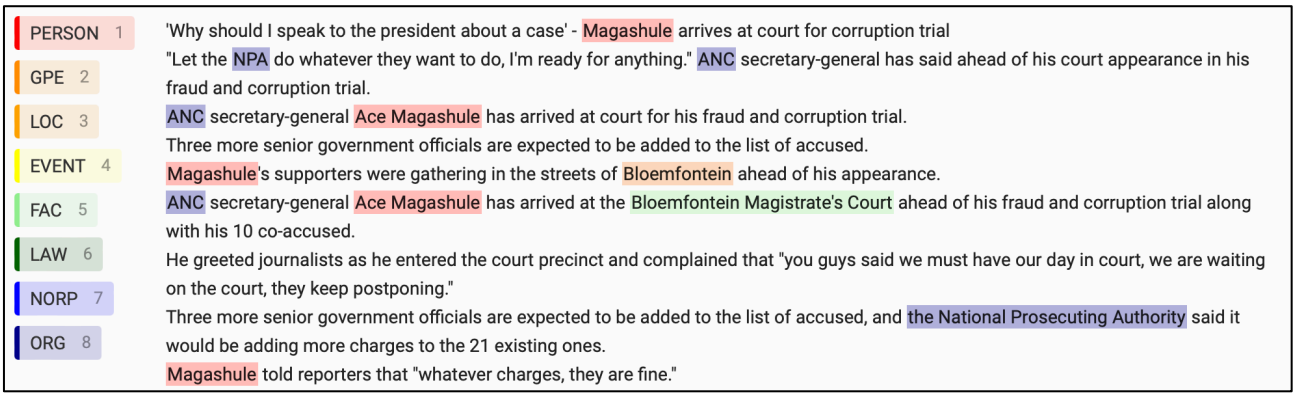

Figure 5: Sample of NER annotations

4.4.2 Coreference resolution annotation

Annotating 30 articles took 6 hours. *Figure 6* shows part of an article annotated for CR. The process was not trivial as it requires the annotator to find clusters with mentions that span the entire article. Some articles referred to multiple entities referenced by multiple pronouns – and it was occasionally hard to keep track of which pronoun referred to which entity. Furthermore, without subject-matter expertise, it was difficult to say whether terms like *Vrede Dairy Farm* and *Vrede Dairy Project* were synonymous – even in context.

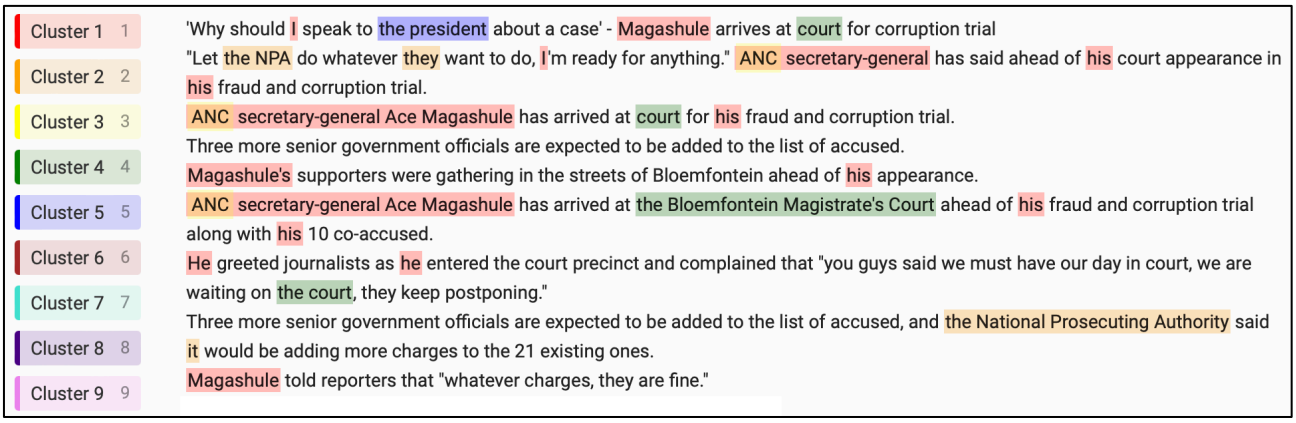

Figure 6: Sample of CR annotations

4.4.3 Relation extraction annotation

Annotating 30 articles took 12 hours. Annotations were completed using the 61 main relations identified – since inverse relations are entailed by the ontology (Hogan, et al., 2021, 4.1.3). The annotation process was time-consuming. Challenges included annotating all (cross-sentence) instances of a relation, deciding which relation type was most appropriate when there was ambiguity, and simply bearing in mind all 61 possibilities at once. The labelling interface was not ideal, with overlapping annotations making review visually tricky as seen in *Figure 7*:

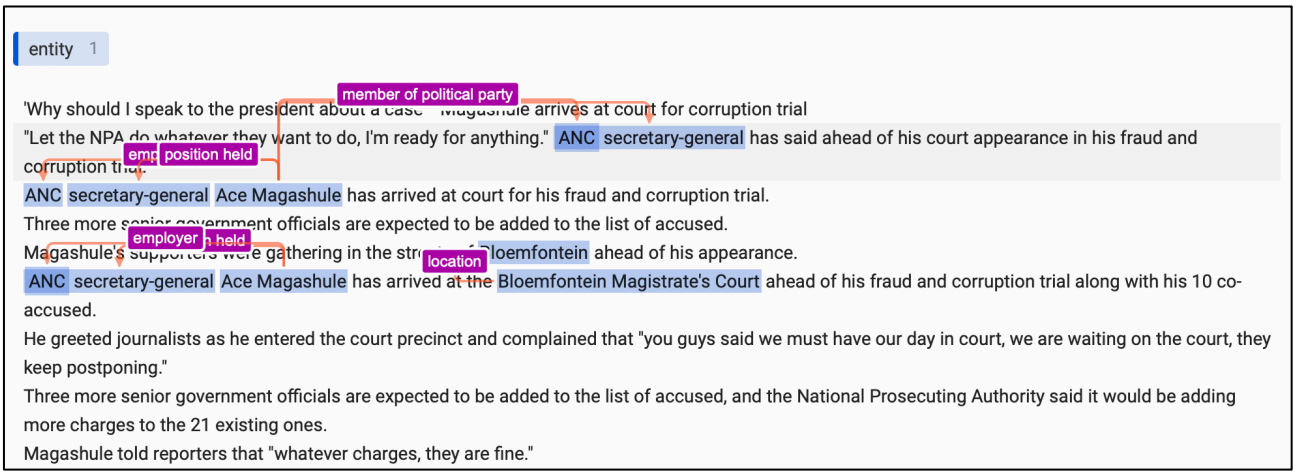

Figure 7: Sample of RE annotations

4.5 Component selection

Initial model selection was achieved by comparing the macro F1 performance of each model's outputs against the full set of 30 annotated articles (in other words the average of the F1 scores across the articles).

Article F1 score: $2 \times \frac{Precision \times Recall}{Precision+Recall}$

Macro F1 score: $\frac{1}{n}\sum_{i=1}^{n} Article\ F1_i$

In each case consideration had to be given to how to compare model outputs.

4.5.1 Named entity recognition evaluation

NER has two dimensions: matching of the entity tokens, and correct identification of the type. The MUC-5 evaluation method proposed by Chinchor & Sundheim suggests the categorizations shown in *Table 5* (1993):

Type	**Abbr.**	**Description**
Correct	COR	Prediction and annotation match exactly
Incorrect	INC	Prediction and annotation do not match at all
Partial	PAR	There is similarity between the prediction and annotation
Missing	MIS	An annotation was not predicted at all
Spurious	SPU	A prediction exists for which there is no annotation

Table 5: MUC-5 evaluation categories showing how each one is applied

"Partial boundary matching (regardless of type)" refined on this concept (Segura-Bedmar, et al., 2013), and the final evaluation metric used for the project classified *partial string matches*, as well as *exact string matches but incorrect entity type* both as partial matches (with a 50% weighting in the final scores) as shown below (Batista, 2018):

Precision denominator: $ACT(UAL) = TP + FP = COR + INC + PAR + SPU$

Precision: $\frac{COR + (0.5 \times PAR)}{ACT}$

Recall denominator: $POS(SIBLE) = TP + FN = COR + INC + PAR + MIS$

Recall: $\frac{COR + (0.5 \times PAR)}{POS}$

These enabled calculation of article F1 scores and macro F1 scores as described above.

4.5.2 Coreference resolution evaluation

An average of MUC, B^3 and CEAF F1 scores is typically used for evaluation of CR performance (Denis & Baldridge, 2009). A Python implementation *corefeval* (tollefj, 2023) is available to compile these scores and was used.

4.5.3 Relation extraction evaluation

Evaluation of RE was complicated by the two models under consideration having different relations available to predict. Inherently REBEL seemed like the preferred option as it offered many more relations of interest. However, reliability was also a factor and so the initial comparison of macro F1 scores was done on *just* the relations shared by both models. A second evaluation was then done on all relations of interest (section *5.1.3*).

4.5.4 Entity linking evaluation

Because EL involves searching for entities in public KGs, annotation was less viable as it would be too labour-intensive. EL was run on a selection of articles and the outputs compared to select the best option in terms of most entities linked and validity of those entities.

4.6 Model improvements

Having the HITL dataset available enabled detailed inspection of edge cases and issues at each round of the build-and-evaluate loop. The focus at this stage was on improving model precision – or fixing issues. Section *5.2* itemizes specific methods used to improve the output quality.

4.7 Processing knowledge into the final KG

4.7.1 Processing articles into the KG

Two steps, construction of the KG and processing knowledge into it (Tamašauskaitė & Groth, 2023), are combined in the KG article processing algorithm outlined in *Figure 8*. It is designed to handle articles one at a time, similar to a production environment. As such it was important to allow for updating of entities and relations as new information becomes available. For example two entities *Africa Global Operations* and *AGO* may have been created and their respective relations maintained in the KG. However, subsequent information reveals that *AGO* is an alias for *Africa Global Operations*. At this point the two entities need to be merged, and the relations updated, which is why change tracking is so crucial.

Method: New relations (and their associated entities) from an article are processed into the KG as follows:

- New articles are created as entity type ARTICLE, and the *entities_tracker* updated accordingly
- For each relation, check if it contains sufficient information for a relation to be defined
- If it does, process the head and tail entities into the KG (section *4.7.2* for details)
- If it is a new relation, create it and update the *relations_tracker* accordingly; otherwise update the relation count for the existing relation

- Create relations between each entity and the associated article, and update the *relations_tracker* accordingly

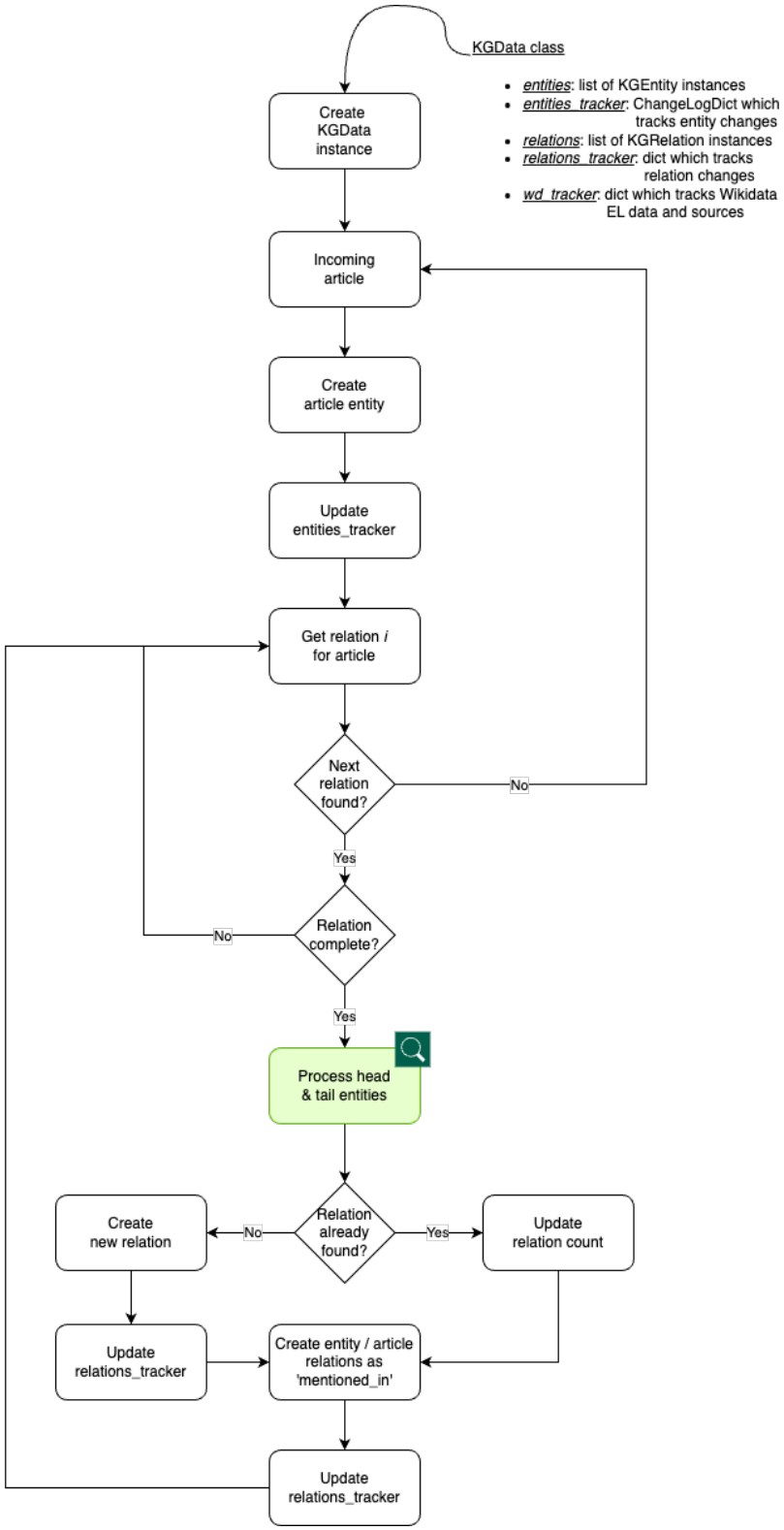

Figure 8: KG article processing algorithm in detail showing how articles, entities and relations are processed

4.7.2 Processing entities into the KG

The sub-algorithm for processing entities into the KG proved quite complex, with several inter-related dependencies as can be seen by the outline in *Figure 9* (overleaf).

Method: Key issues that are addressed by the entity-processing algorithm include:

- Whether to create a new entity, update an existing entity, or switch entities and merge information as aliases are disambiguated – either as a result of Flair's *alternate_name* outputs or through Wikidata
- Deciding whether to perform EL for the entity (based on whether any match is found in Wikidata)
- For entities with any match found on Wikidata EL was attempted using OpenTapioca up to a maximum of 5 retries
- If no result was obtained from OpenTapioca after 5 retries, and only one match was found via Pywikibot, that Wikidata entry is incorporated
- Extensive logging was implemented so that all edge cases could be traced back to their origins – which aided considerably in finalizing the logic

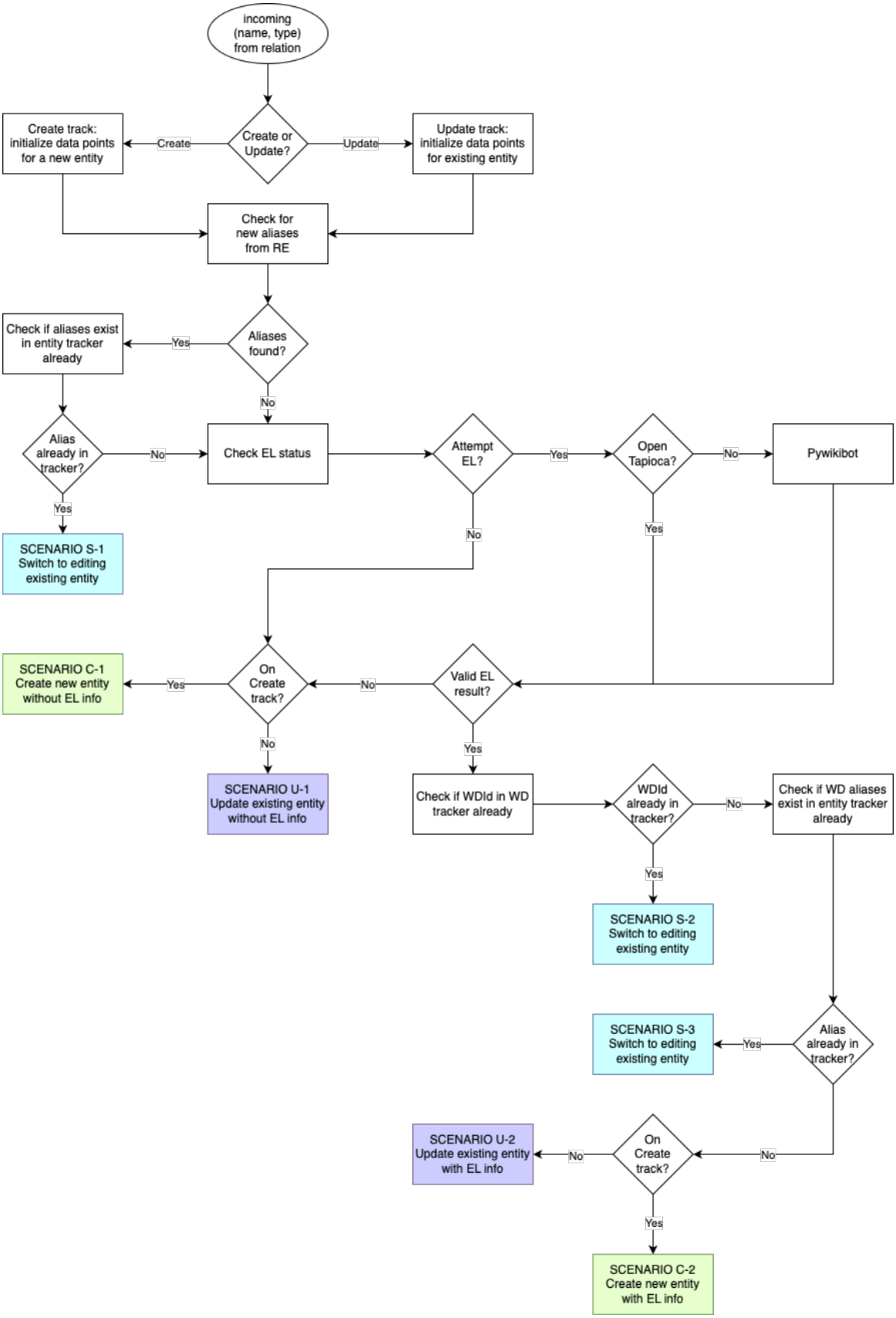

Figure 9: KG entity processing sub-algorithm in detail showing how EL is attempted and scenarios tracked

4.8 KG evaluation

Evaluating the built KG instance focused largely on precision, although some exploration of recall was done with the help of the journalists in the focus group (section *6.2*).

Entities evaluation method: The following data points were evaluated for KG entities:

- **Entity name and type**: Similar to the NER evaluation method (section *4.5.1*), the following criteria were used to score each entity:

Type	**Abbr.**	**Score**
Correct	COR	1
Incorrect	INC	0
Partial	PAR	0.5
Spurious	SPU	0

  The final score was **NER precision**: $Total\ Score/\#\ Entities$.

- **Entity aliases**: the following criteria were used to score aliases for each entity:

Type	**Score**
All aliases have been identified	1
No aliases and entity has no aliases	1
No aliases, entity should have aliases	0
Some aliases present, but incomplete	0.5
A single entity is split across multiple aliases	0.5

  The final score was **alias precision**: $Total\ Score/\#\ Entities$.

- **Entity linking**: the following criteria were used to score Wikidata matches for each entity where a link was found:

Type	**Score**
Linked entry correct for entity in context	1
Linked entry incorrect for entity in context	0

  The final score was **EL precision**: $Total\ Score/\#\ Entities\ with\ WD\ links$.

Relations evaluation method: KG relations were evaluated as True (with a score of 1) or False (with a score of 0). The final score was **RE precision**: $Total\ Score/\#\ Relations$.

5 Build-and-evaluate results

As recommended by Gregor & Hevner (2013), the following section describes the design search space and methods developed during each iteration of the build-and-evaluate loop.

5.1 Round 1 - component evaluation & selection

Each model was run against the 30 articles of the HITL dataset and then evaluated against the annotations. Jaradeh, et al.'s hypothesis (2023), is borne out in that the model with the better F1 score in the literature was not always the model with the better F1 score on this corpus.

5.1.1 Named entity recognition models

Table 6 shows spaCy's macro F1 results were marginally better than Flair's but its performance was considerably faster, so it was selected as the base on which to build further. Macro F1 on the test set of 20 articles is noted for reference.

		Full set (30 articles)		Test set (20 articles)
Component	**& Reported F1**	**Runtime**	**F1**	**F1**
spaCy en_core_web_trf	0.90000	6.89772 sec	0.91592	0.92278
flair/ner-english-ontonotes-large	0.90930	12.48451 sec	0.91078	

Table 6: NER Metrics for spaCy and Flair (spaCy, n.d.) and (Schweter & Akbik, 2020)

5.1.2 Coreference resolution models

Table 7 shows that fastcoref performed better on the sample set and was much faster, so it was selected as the base on which to build further. Macro F1 on the test set is noted for reference.

		Full set (30 articles)		Test set (20 articles)
Component	**& Reported F1**	**Runtime**	**F1**	**F1**
fastcoref	0.78500	24.03070 sec	0.71688	0.73203
LingMess	0.81400	36.43618 sec	0.68682	

Table 7: CR Metrics for fastcoref and LingMess (Otmazgin, et al., 2022) and (Otmazgin, et al., 2023)

5.1.3 Relation extraction models

Table 8 shows that, on the relations shared by the two models, REBEL outperformed Flair on macro F1 score and speed by a large margin, confirming it as the preferred choice in terms of coverage and performance.

		Full set (30 articles)		Test set (20 articles)
Component	**& Reported F1**	**Runtime**	**F1**	**F1**
Babelscape/rebel-large	0.93400 (NYT)	25.44956 sec	shared 0.49206 all 0.53658	all 0.52550
Flair	n/a	47.52002 sec	shared 0.22977	

Table 8: RE Metrics for shared REBEL and Flair relations (Cabot & Navigli, 2021) as well as across all relations for REBEL

Although it fell outside the scope of this project, in the face of a low starting point like this it would be highly recommended to investigate additional models or alternative methods at this point.

5.1.4 Entity linking models

The initial results from EntityLinker were so obviously poor that it was discarded almost immediately as an option, and OpenTapioca selected in preference. *Table 9* illustrates this using the sentence: *Jacob Zuma, former president, remarked that it was going to be a long day.*

Component	**Outputs**
OpenTapioca	*Found*: Jacob Zuma *Wikidata reference*: https://www.wikidata.org/entity/Q57282 *Wikidata description*: 4th President of South Africa (2009–2018)
EntityLinker	*Found*: Jacob *Wikidata reference:* https://www.wikidata.org/wiki/Q289957 *Wikidata description:* Patriarch son of Isaac, God renamed him as Israel. Father of the Israelites *Found:* president *Wikidata reference:* https://www.wikidata.org/wiki/Q1255921 *Wikidata description:* non-political leader of an organization, company, community, club, trade union, university or other group *Found:* Long Day *Wikidata reference:* https://www.wikidata.org/wiki/Q6672495 *Wikidata description:* 1996 single by Matchbox Twenty

Table 9: Comparison of EL outputs for OpenTapioca and EntityLinker (Delpeuch, 2020) and (Gerber & Mensio, 2023)

5.2 Round 2 – improving baseline model outputs

Round 2 focused on carrying forward the outputs from Round 1, investigating edge cases and issues, and devising solutions to improve overall results. *Figure 10* summarizes the measures taken – described in detail below:

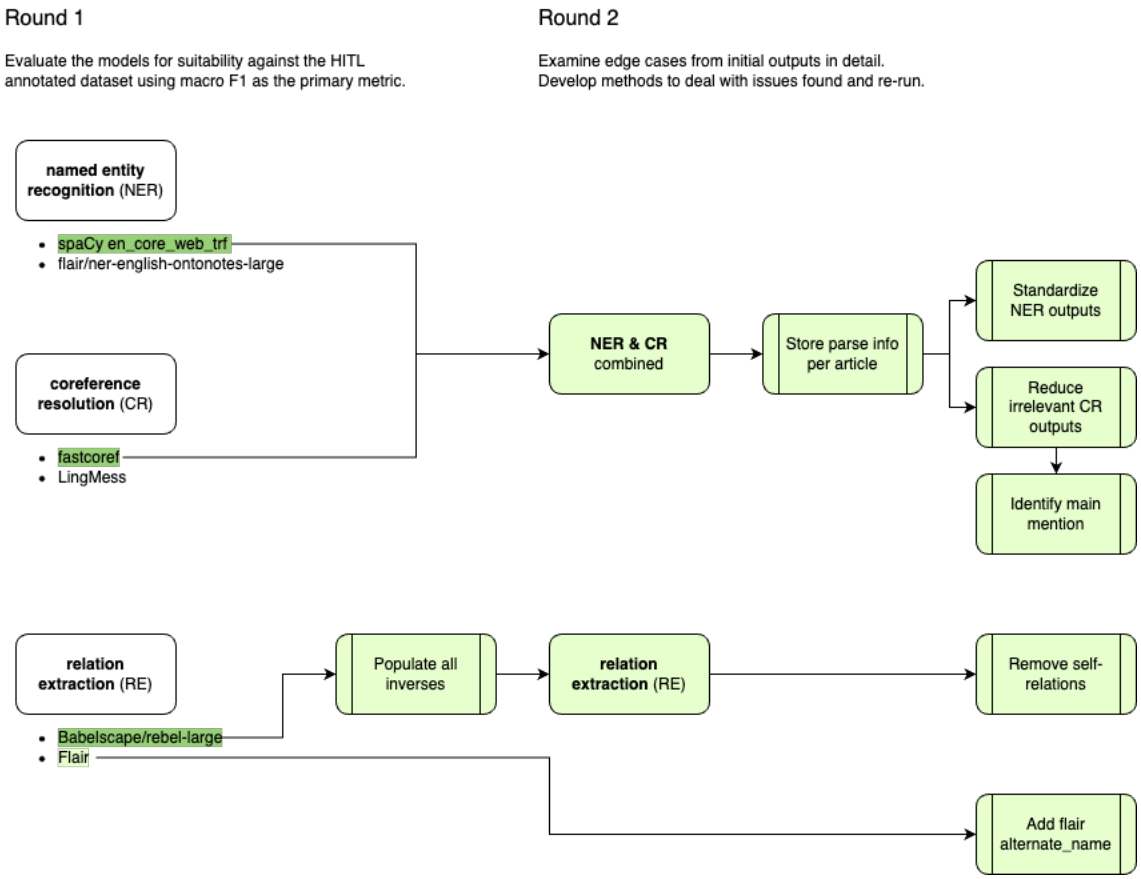

Figure 10: Round 1 > 2, build & evaluate loop for pipeline components

Because spaCy was used as the foundation for both NER and CR, these models were combined resulting in a runtime improvement of nearly 4 seconds for the 30 articles (see *Table 10*).

Standardization of NER outputs: All titles were removed from names so that instances like *Advocate Paul Pretorius SC* resolved to *Paul Pretorius*. Furthermore spaCy's named entity outputs were "noisy" in some cases – for example inclusion of extraneous punctuation, or only outputting part of a name like *Home Affairs* instead of *Department of Home Affairs*, requiring standardization.

Method: Collect entity type and part-of-speech tags per token ("parse info" in *Figure 10*). Assemble clean entities according to the following rules:

- include only entities of interest (PERSON, GPE, LOC, EVENT, FAC, LAW, ORG)
- include adpositions (ADP) in terms like *Department of Home Affairs*

- include possessives (PART) and coordinating conjunctions (CCONJ) in terms like *Liesel's Cakes and Bakes*
- include punctuation (PUNCT) along with opening and closing brackets in terms like *Price Waterhouse Coopers (PWC)*
- exclude extraneous determiners (DET) in terms like *the Sunday Times*
- exclude extraneous trailing possessives in terms like *Advocate Gcabashe's*
- exclude extraneous closing brackets where no opening bracket was identified as well as extraneous spaces

Adding this post-processing method increased the NER F1 score by 1.4% (see *Table 10*), ensuring higher levels of consistency.

Reducing irrelevant CR outputs and identifying main mention(s): CR resolves *all* mentions in text, not just those related to named entities. The first challenge was to only extract clusters that referred to named entities. This is not as simple as looking for phrases that contain a named entity – for example with a phrase like *the journalists who worked for the Sunday Times* the main subject is the journalists (not a named entity). For CR to be useful one also needs to establish a clean, unambiguous entity name. A phrase like *My former colleague Pearlie Joubert of the Sunday Times* contains two named entities, so it is necessary to use dependency parsing to establish that *Pearlie Joubert* is the main mention of this phrase. Furthermore some mentions may legitimately refer to multiple entities, like *PWC and SAA*.

Method: Collect parts-of-speech tags, dependency, and dependency head tags per token ("parse info" in *Figure 10*). Filter coreferences according to the following rules and attempt to resolve to a single entity:

- include only entities of interest (PERSON, GPE, LOC, EVENT, FAC, LAW, ORG)
- include only dependencies of interest – those that will form the noun head or be closely associated with it (nsubj, dobj, csubj, nsubjpass, csubjpass, xsubj, iobj, pobj, compound, appos, nmod)
- find entities that are a match for coreference clusters
- where there is ambiguity find the main subject, and if that (or any of its compounds) is a named entity select it (the example in *Figure 11*) or if not look for appositional modifiers that may constitute the named entity (the example in *Figure 12*)

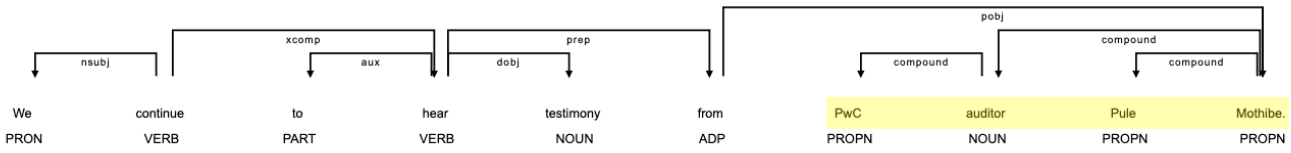

Figure 11: Retrieving noun head and associated terms

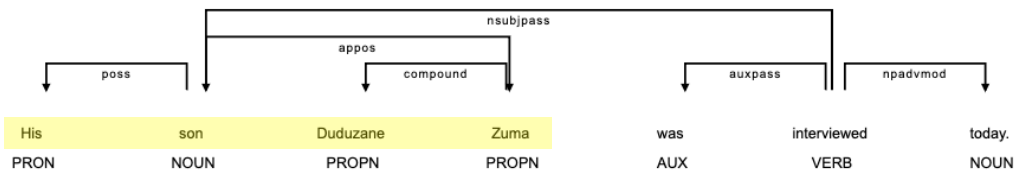

Figure 12: Finding appositional modifiers

Adding this post-processing method increased the CR F1 score by 14% (see *Table 10*), however edge cases remained in terms of obtaining a definitive main mention. Furthermore, because CR gathers entire noun phrases, alternative references were also still messy. At this point significant development work remained to obtain truly useful outputs.

Refining RE outputs: REBEL was used as the main RE model, and some basic improvements were identified including removing self-relations and populating inverses entailed by the Wikidata ontology. An important change introduced was retaining Flair for its *alternate_name* relation. Because of the complexities that emerged with CR during round 2 – and given that Flair's *alternate_name* proved reliable on the HITL dataset at a threshold > 0.999 – it was decided to include it just for this one relation. This went a long way to aiding with disambiguation. It came at the cost of longer runtimes but entity disambiguation is such an important part of the process that it was deemed an acceptable trade-off.

> **Method**: Update REBEL outputs by removing self-relations and populating inverses entailed by the ontology. Add Flair's *alternate_name* relation where confidence > 0.999.

Overall these changes resulted in a modest 0.686% improvement in the F1 score for relation extraction (see *Table 10*).

NER, CR & RE Improvements realized: *Table 10* summarizes the improvements realized through the above methods:

		F1 Test set (20 articles)	
Component	**Runtime**	**Round 1**	**Round 2**
NER spaCy en_core_web_trf	26.10637 sec	0.92278	0.93697
CR fastcoref		0.73203	0.87199
RE Babelscape/rebel-large Flair	23.49133 sec 43.18754 sec	0.52550	0.53236

Table 10: Round 2 summary of improvements

5.3 Round 3 – building the first KG on the HITL dataset

The first KG was built on the foundation of the model outputs obtained from Round 2, using all 30 articles of the HITL dataset. Doing so allowed a detailed evaluation of the veracity of the KG. Using the evaluation criteria described in section *4.4.1*, the following results were obtained:

Result type	Runtime	Metric
Number of entities	03 min 20 sec	277
Number of linked entities		57
Number of relations		283
NER precision		0.88628
Alias precision		0.93502
EL precision		0.94737
RE precision		0.51590

Table 11: Round 3 metrics using the KG evaluation metrics

As can be seen, the KG precision metrics align quite well with the F1 metrics from Round 2. Note the runtime includes API calls to OpenTapioca and Wikidata via Pywikibot.

5.4 Round 4 – improving the KG

Two key factors were noted in Round 3 which caused poorer performance: ambiguous relations where the same entities had more than one possible relation; and the issue of overlapping names, for example *Zondo >> position held >> Deputy Chief Justice* vs *Raymond Zondo >> position held >> Deputy Chief Justice*.

Method: the following additional post-processing steps were added to the RE pipeline:

- Removal of all ambiguous relations
- Cleanup overlapping relations by first trying to get the best matching entity found by NER, and failing that taking the longest name found by RE

The addition of these methods resulted in improvements to precision as shown in *Table 12* – however at a slight cost to recall as evidenced by the lower numbers of entities and relations found overall:

Result type	Runtime	Metric
Number of entities	03 min 32 sec	242
Number of linked entities		47
Number of relations		222
NER precision		0.91322
Alias precision		0.95248
EL precision		0.98000
RE precision		0.59009

Table 12: Round 4 metrics using the KG evaluation metrics

6 Final results

As recommended by Gregor & Hevner (2013), the following section describes the final instantiation of the news KG, and discusses the subject-matter expert evaluation process including its results and implications for the design process itself.

6.1 The built KG

The final KG identified 6944 entities and 10829 relations (excluding articles). *Figure 13* shows an excerpt: the ego-graph for *Transnet* and entities directly related to it:

Figure 13: Excerpt from KG – ego-graph for Transnet (excluding related articles)

Given the content it was unsurprising that people and organizations featured most prominently: *Figure 14* shows the top 10 node types identified (see *Appendix C* for full breakdown).

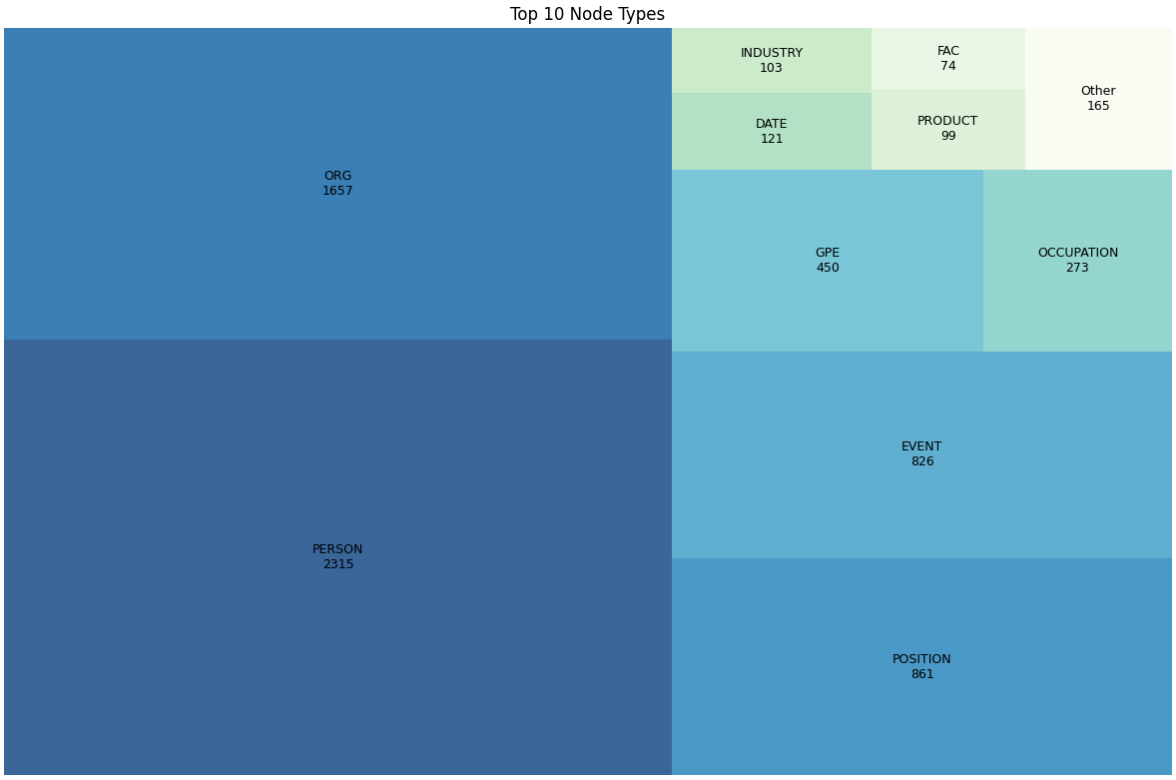

Figure 14: Top 10 node types

Figure 15 shows the top 10 nodes by number of connections. These are all entities that were important in the coverage so it is expected they yielded a large number of relations:

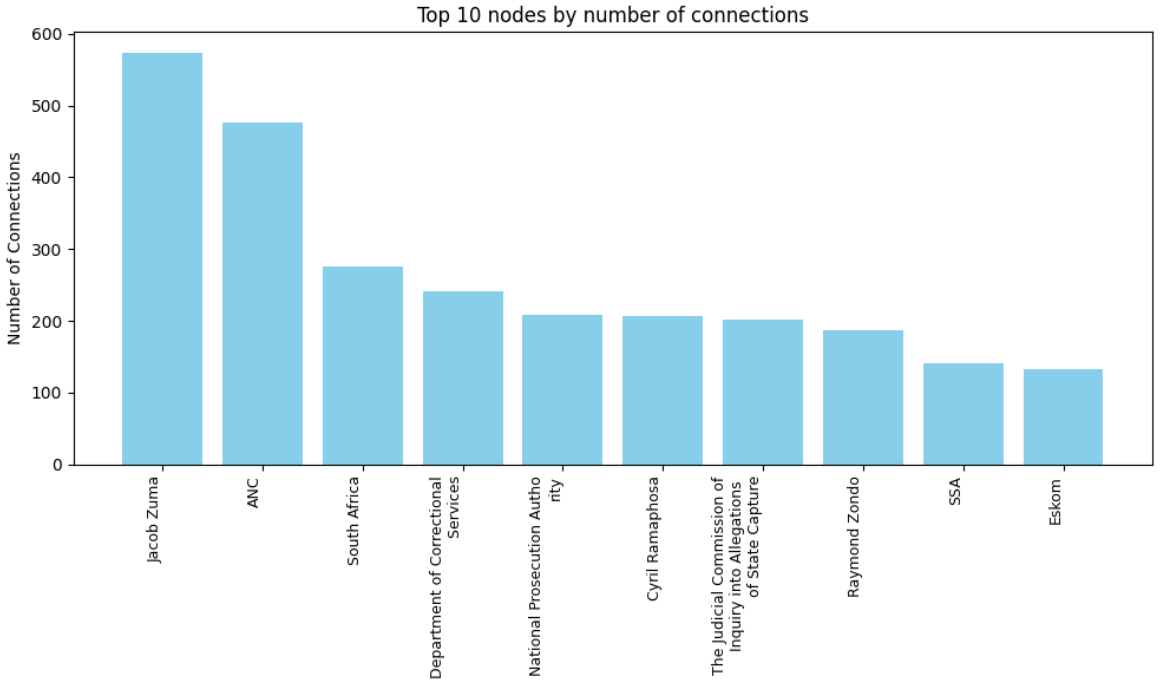

Figure 15: Top 10 nodes by number of connections

616 entities had aliases associated with them. 811 entities were matched to Wikidata (540 through contextual matching via OpenTapioca, and 271 through direct matching via Pywikibot). There are 557 weakly connected components, and while one expects a real-world network like this one to be sparse, a density of 0.00022 is relatively high – with 5645 nodes or 81.29% occurring in the largest of the weakly connected components, indicating many inter-related connections identified which is positive.

Figure 16 shows the top 10 relation types identified (see *Appendix C* for full breakdown).

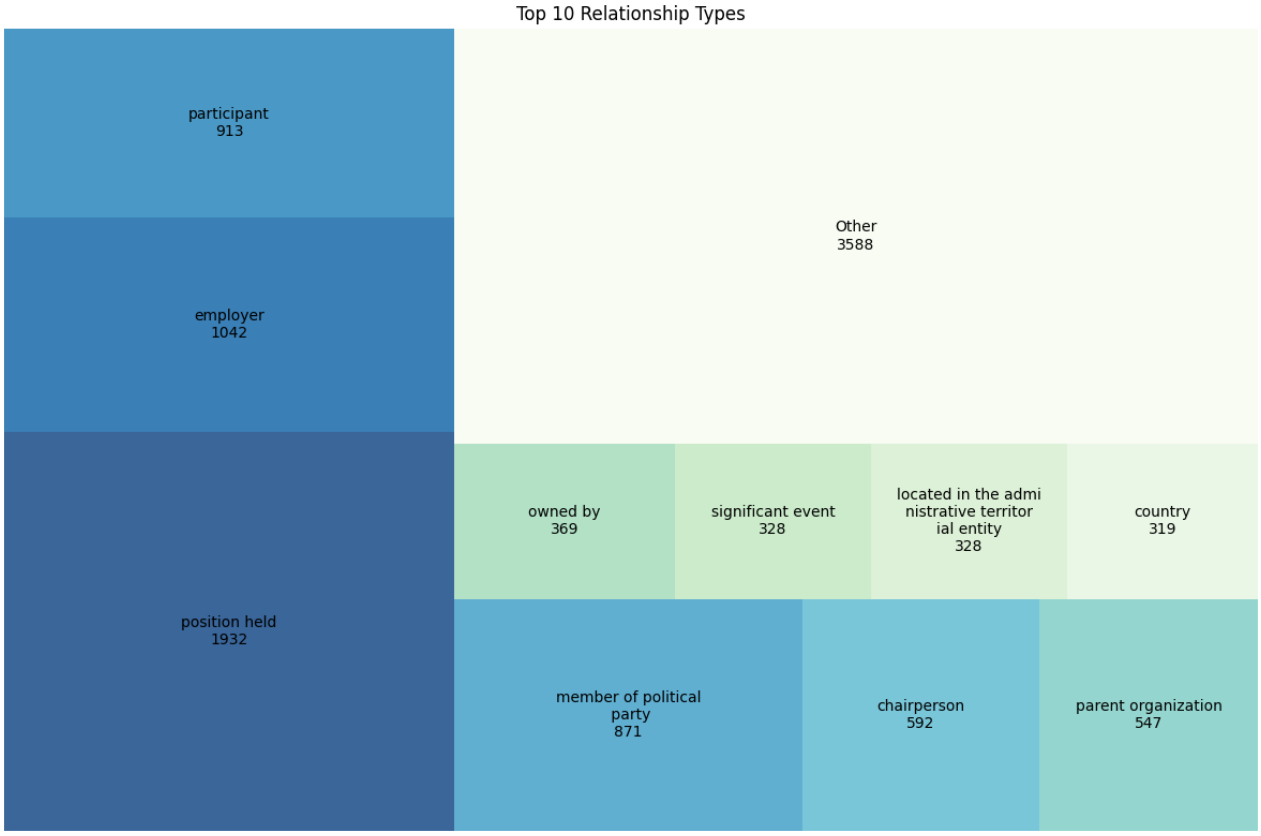

Figure 16: Top 10 relation types

The distribution of relation weights (number of times a relation was identified in the corpus) in *Figure 17* shows that only 1796 (or 16.6%) were found more than once which meant it was not possible to use frequency as a quality measure in determining which relations to include.

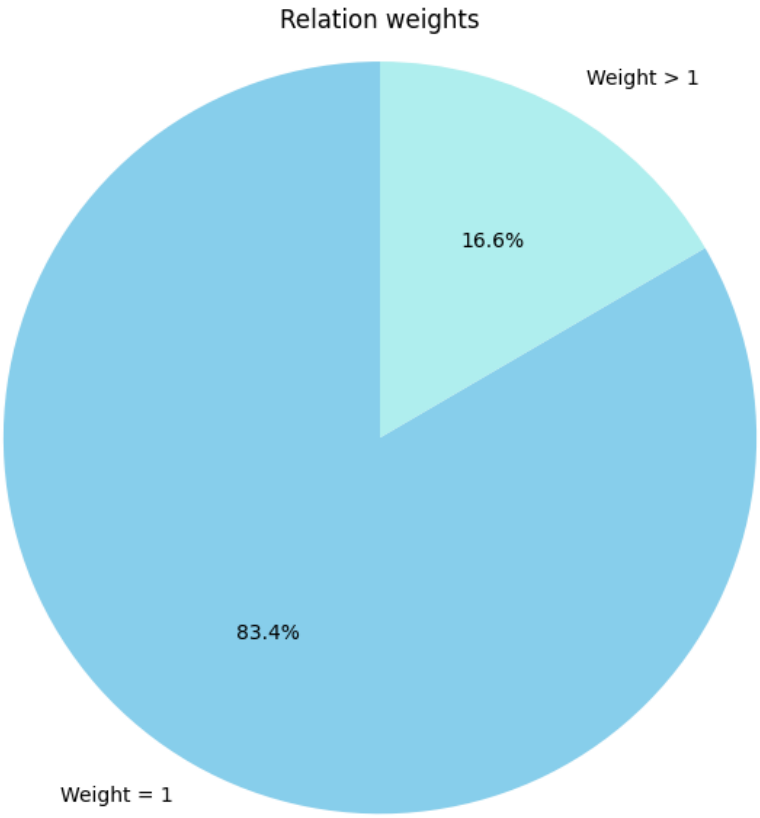

Figure 17: Relation weights = 1 / > 1 demonstrating why it was not possible to use frequency as a quality filter on relations

6.2 Editorial evaluation

News24 is fortunate enough to have two of the country's foremost authorities on the *Zondo Commission* and state capture on the editorial team. **Adriaan Basson** (AB) is editor-in-chief at News24. Aside from extensive coverage for the site on this topic, he is also co-author of the book *Enemy of the People: How Jacob Zuma stole South Africa and how the people fought back*[1] and author of *Blessed by BOSASA: Inside Gavin Watson's State Capture Cult*[2]. **Kyle Cowan** (KC) is an investigative journalist and also author of the book *Sabotage: Eskom Under Siege*[3].

To ensure that evaluation was not biased, AB was asked to select 5 organizations and 5 people of interest for review. Two focus group sessions were held: the first with AB, the second with KC as outlined in *Table 13*.

[1] Basson, A. & Du Toit, P. (2017) *Enemy of the People: How Jacob Zuma stole South Africa and how the people fought back*. Johannesburg: Jonathan Ball Publishers
[2] Basson, A. (2019) *Blessed by BOSASA: Inside Gavin Watson's State Capture Cult.* Johannesburg: Jonathan Ball Publishers
[3] Cowan, K. (2022) *Sabotage: Eskom Under Siege.* Cape Town: Penguin Random House

The 2 sessions together totalled 4 hours, of which 3½ hours were spent in review, while the remaining time was used to introduce the reviewers to the concept of KGs and explain the review methodology (section *4.8*). In the allotted time the 6 smallest egos and their relations were reviewed in their entirety. There was insufficient time to complete the remaining 4.

Ego	**Entities**	**Relations**	**Reviewed by**
Angelo Agrizzi	24	33	AB
Estina	28	30	AB
Atul Gupta	21	35	AB
Transnet	65	78	AB
Brian Molefe	26	29	KC
Eskom	118	137	KC
Tom Moyane	28	31	N/A
Bosasa	219	276	N/A
Jacob Zuma	490	592	N/A
ANC	449	482	N/A

Table 13: Egos selected for review and their review status
6/10 egos selected were reviewed in the available time

Reviewing these data points was easy for the journalists and progress was quick once the first few examples were understood. Because provenance was built into the KG design, where there was doubt about the veracity of a triple it was easily traced back to the article and sentence of origin in the corpus for further inspection.

Table 14 shows the results from the final review, which are discussed in detail in section *6.3*:

Result type	**Metric**	**Comparison to Round 4**
Total entities reviewed	251	+9
Total linked entities	60	+13
Total relations reviewed	333	+111
NER precision	0.86454	−0.04868
Alias precision	0.87251	−0.07997
EL precision	0.85000	−0.13000
RE precision	0.48649	−0.10360

Table 14: Final KG sample metrics

The reviewers were also asked to consider important missing relations. *Transnet* was deemed sufficient. However, gaps could be found for all the remaining egos. Interestingly many of these fell outside the scope of the 61 relation types included – and indeed outside the scope of REBEL generally. Key omissions included who reports to whom within organizations, colleagues in peer relationships, shareholders, and supplier or service-provider relationships.

6.3 Evaluation implications

There was a reduction in performance on all metrics in the final KG compared to round 4 which used the HITL dataset. The observed differences are, however, within reasonable bounds of expectation, given the disparity in the size of the datasets, with the full corpus providing a much more representative test of the model's performance under varied circumstances. The editorial review enabled identification of key issues:

6.3.1 Relation extraction

An RE precision of 0.48649 is unsatisfactory. A very recent survey from July 2024 on RE techniques still lists REBEL as the 2nd best performer on the NYT dataset with its F1 score of 0.93400 (Zhao, et al., 2024). It is worth noting, however, that REBEL's own reported results on CONLL04 (*also* a news-based dataset) were much lower with an F1 of just 0.71970 (Cabot & Navigli, 2021) – once again highlighting how performance varies across datasets. One theory to account for the degradation of performance is this corpus containing many South African names, which REBEL is unlikely to have seen. Ensembling model results was contemplated to improve outputs, but the poorer F1 of Flair precluded this option. It was further envisaged that precision could be increased by using statistical methods, for example only including relations identified $> n$ times. However, as seen in *Figure 17* of section *6.1*, this was also not possible given 83.4% of relations occurred only once – and perhaps news corpora in general would face this issue as coverage on many topics would be broad rather than deep.

To obtain satisfactory results on RE alternative models should be evaluated and / or alternative methods incorporated. Weikum, et al. are optimistic about results to be had from pattern- and rule-based methods (2021). Furthermore, because these methods would be constructed on the actual corpus, they would likely deal well with local names. Jurafsky & Martin confirm that pattern-based techniques can yield high-precision results but are labour-intensive. They suggest bootstrapping methods that build on seed patterns to increase recall by learning new patterns in a semi-supervised learning context (2024, p. 7). Given that both subject-matter experts expressed a strong desire to see additional relations not included in the base models, this additional effort would be justified.

6.3.2 Entity disambiguation

The most serious issues that reduced NER precision arose from attempts at entity disambiguation. The first was where EL resulted in an erroneous link, effectively "poisoning" the entire entity. The following examples hint at a first-world bias in Wikidata contents:

- *The New Age* newspaper is also referred to in the corpus as *TNA*. OpenTapioca linked this to the *UK National Archives* which is also known as *The National Archives* or *TNA*.
- South Africa has a *Department of Justice*, also referred to as *DOJ* in the corpus. Open Tapioca linked this to the *United States Department of Justice*, also known as *DOJ*.

Weikum, et al. discuss use of a "contextual profile" to deal with this "out-of-KB" entity problem. Multiple contextual mentions of an entity are gathered, which together can be used to evaluate whether the entity should be linked or not (2021). One method to do this could be to gather a companion set of mentions from the associated entity in Wikipedia (which maps to Wikidata) and use embeddings to calculate cosine similarity between the 2 sets of mentions, creating a proxy confidence-score for the proposed match.

The second issue arose from erroneous *alternate_name* outputs from Flair – effectively conflating 2 entities that should have remained separate, for example:

- *Bosasa* was returned as an alternate name for the *Department of Correctional Services*, where *Bosasa* was actually a supplier to the department.

Re-incorporating CR into the pipeline would add value here, either as a means to cross-check suggested aliases from Flair, or as an entirely alternative method to produce the same information.

6.3.3 Canonicalization of entities

Closely linked to entity disambiguation is the canonicalization of entities, where the objective is that all the different ways in which an entity can be referred to in the corpus (and public KGs) are wrangled into a single entity. The alias precision score was gratifyingly high, but the edge cases examined showed (once again) that re-incorporating CR into the pipeline would be beneficial. The most obvious example is that of the *Zondo Commission* itself. Many variants were left isolated, e.g. *state capture inquiry*, *state capture commission*, *state capture commission of inquiry*, which could have been connected through CR.

7 Discussion and conclusions

Having produced an instantiation of a news KG through which the proposed model and methods could be evaluated, the original research questions (RQ) can now be revisited.

In terms of assessing the feasibility of using "out-the-box" models for each pipeline component (RQ1), testing multiple models showed that results from the literature are unlikely to match those on a specific corpus. It is therefore worthwhile setting up experiments to test multiple models. Some models performed significantly faster than others, but for the volumes involved in a typical newsroom (Majid, 2023) all ran within acceptable parameters. **NER** models were fit-for-purpose as-is, and were improved with simple NLP post-processing methods. **CR** results showed considerable post-processing is required to transform the outputs into a format that is useful for KG processing. These issues are not insurmountable, but do require non-trivial NLP developments. **RE** models were sub-par on this corpus. It is possible alternative models would perform better. However, more than one source also suggests looking at rule- and pattern-based methods, and perhaps bootstrapping these to improve recall as well as precision. It would therefore be recommended to combine one or more of these methods with the best baseline model available. This would enable inclusion of useful relations (beyond the closed ontology derived from RE). It could also improve performance – particularly when it comes to local names. This is no longer "out-the-box" territory, however, and would require additional development time. Only one of the **EL** models evaluated was viable and seemed to suffer from a bias towards confirming erroneous results from first-world entries available in the database. Additional methods like the contextual profiling described in section *6.3.2* could assist in obtaining confidence scores for link proposals.

Creation of the HITL gold-standard dataset was both feasible and effective (RQ2). Annotating 30 articles took only 21 hours for all 3 tasks (NER, CR and RE). Involving a subject-matter expert for cases where there was uncertainty about labels would have been ideal, and a second annotator would also have been preferred from a quality perspective. It is worth spending time on this stage, and even to consider annotating a larger dataset. The annotation process itself revealed many issues and decisions that would have to be considered during the project – serving as an exploratory data analysis of sorts.

The HITL dataset not only supported quantitative evaluation at each iteration so that appropriate decisions could be made about how to navigate the design search space, it also assisted with surfacing specific issues and edge-cases – which in turn suggested the next round of improvement ideas. While it is unsurprising that the metrics on the final KG were rather lower than those obtained on the HITL dataset, it nonetheless provided a very adequate indication of which models and methods were likely to be sufficient and where further development would be required.

Involving subject-matter experts for the final KG review also added value to the process (RQ3). They accurately reviewed a large number of data points in a short time. This enabled a quantitative assessment of a sample of the final KG, and again highlighted key issues that should be addressed in successive iterations. It effectively also produced a *second* partial gold standard dataset against which future builds could be judged. Furthermore, getting stakeholder feedback at this point in the process was useful in that it showed the need to expand the relation types included. As an aside, one might wonder whether, instead of using subject-matter experts, an LLM could be used – but a small test on 20 of the triples reviewed by the journalists only yielded a score of 17/20 for the LLM.

Although the assessed quality of the final KG indicates further iterations on the build-and-evaluate cycle are required, the model used to approach the project proved largely sound (RQ4). For newsrooms wanting to introduce KG technology, use of the **design science methodology** is a good choice. Its principles are aligned with the way projects typically run in a business setting, and it allows for the iterative exploration and evaluation of possibilities within the design search space. The specific methods developed from this exploration (section *5*) resulted in measurable incremental improvements, and could be investigated for use on other corpora. The use of **open-source components** is still recommended as a foundation on which to build the KG but with important caveats. Sufficient time should be allowed for investigating components that achieve a reasonable baseline. In addition custom developments using NLP techniques that require at least intermediate-level linguistic expertise should be anticipated.

Creation and use of the **HITL dataset** formed a solid quantitative foundation for the project, as did the use of **subject-matter experts** in the final evaluation. Not only did these enable comparative assessments, but also informed the way forward at each iteration. It is recommended to incorporate both of these approaches into the KG building process to ensure that progress is measurable and requirements are met. Having used these approaches, it is currently clear what the main issues are, and which directions should be explored next to improve upon the final result (section *6.3*).

One of the limitations of the project is that the content is mainly representative of business and politics reporting. Other news categories like lifestyle and sport may have different requirements. In sport, for example, specific methods might need to be developed to deal with aspects like reporting team statistics. Writing styles may also play a significant role: the formal writing style of a business publication is very different from the colloquial style used by tabloids. Another limitation is that the project focused on building the minimum viable product. In a production-grade KG, further enrichment with additional properties would ideally be incorporated, particularly temporal properties like *when* Jacob Zuma was South Africa's president, and not just the fact that he was.

Key directions for taking this research further should include: 1) an exploration of the effectiveness of incorporating rule- and pattern-based semi-supervised methods to enhance RE; 2) re-incorporating CR outputs along with NLP techniques to improve entity disambiguation and the extraction of aliases; 3) the addition of entity matching techniques based on similarity of mentions to improve the confidence of EL; 4) testing the applicability of the methods on other news categories; 5) including techniques to enrich the KG with additional properties including temporal properties.

The model and methods proposed for approaching KG construction from news articles in this project will form a solid foundation from which to explore these areas.

9915 words

8 Ethics

Media24 owns the content that was used for this project (2081 articles on the *Zondo Commission* published on the News24 site between 2018 and 2022). Written permission was obtained from the CIO of Media24 on 16 November 2023 to use the content for purposes of this research, on condition that only a limited sample of unlocked content would be made publicly available, and that no locked content would be shared. All experiments were conducted within the Media24 GCP environment where the data is housed.

Two members of the News24 editorial team were recruited (on a purely voluntary basis) to help evaluate a sample of the KG outputs – both are experts on the topic of state capture and the *Zondo Commission*. The evaluation sessions were held online. Media24's IT team, the CIO, the Project Head of AI, and the 2 evaluators will all receive a report on the final project outcomes as well as recommendations for how to go forward with the research internally.

Appendix A – Code repository

All code developed for this project is available at https://github.com/shotleft/zondo-knowledge-graph.git which has been shared with user *GoldSmithsMScDataScience* (login details available from module leader). The **README.md** file provides an overview of how the various components relate to the sections in this report.

Appendix B – Hardware specifications

The base model components (Rounds 1 & 2) were run on an N1 standard machine on Google Cloud Platform (GCP) with 8 vCPU, 4 core, 30GB memory and 1 Nvidia T4 GPU. Practically speaking 4 CPU's and 15Gb would suffice as monitoring showed that no more than 4 CPUs were used at a time and memory utilization seldom exceeded 6GB. The remainder of the project was run on an E2 standard machine on GCP (4 vCPU, 2 core, 16GB memory).

Appendix C – KG type counts

Number of nodes per type in the final KG:

Entity type	**Count**
PERSON	2315
ORG	1657
POSITION	861
EVENT	826
GPE	450
OCCUPATION	273
DATE	121
INDUSTRY	103
PRODUCT	99
FAC	74
LAW	68
PUBLICATION	34
FAMILY	28
LOC	24
AWARD	11

Special entity type for articles:

Entity type	**Count**
ARTICLE	2081

Number of relations per type in the final KG:

Relation Type	Count
position held	1932
employer	1042
participant	913
member of political party	871
chairperson	592
parent organization	547
owned by	369
significant event	328
located in the administrative territorial entity	328
country	319
member of	270
field of work	254
sibling	213
spouse	210
occupation	199
inception	187
headquarters location	181
family	137
industry	129
office held by head of the organization	127
country of citizenship	122
product or material produced	115
residence	102
child	96
contains administrative territorial entity	95
founded by	73
father	71
appointed by	71
applies to jurisdiction	59
shares border with	59
date of death	55

Relation Type	Count
capital	54
legislated by	54
head of government	53
candidacy in election	52
office held by head of government	45
dissolved, abolished or demolished date	44
military rank	43
head of state	40
legislative body	40
place of birth	39
educated at	28
replaces	28
place of death	27
presenter	27
relative	21
organizer	20
cast member	19
editor	17
location	16
member of sports team	16
award received	15
operator	14
mother	12
religion	9
location of formation	7
military branch	6
country of origin	6
affiliation	4
director	4
authority	3

Special relation type for relation mentions in articles:

Relation type	Count
mentioned_in	28060

Appendix D – Glossary

adposition

A grammatical term including prepositions and postpositions

apposition

A grammatical term indicating multiple terms that refer to the same item

coreference resolution (CR)

Identifying which entities are being referred to in a piece of text – especially, but not limited to use of pronouns.

disambiguation

Distinguishing between entities or concepts with similar names or linking entities or concepts that occur under different names.

edge

A link between 2 nodes in a graph, typically specifying the relationship between them and, optionally, the direction of that relationship.

ego-graph

All the nodes and relationships associated with a single entity (the ego).

entity linking (EL)

Linking multiple mentions of an entity to a single entity or node in a KG that uniquely represents that entity and disambiguates it from similar entities.

graph

"A set of elements, which we call nodes, along with a set of connections between pairs of nodes, which we call links" (Menczer, et al., 2020, p. 15).

human-in-the-loop (HITL)

Machine learning techniques are combined with human knowledge to create datasets or improve model performance at lower cost (Wu, et al., 2022).

information extraction (IE)

The process of extracting structured information from unstructured data sources like text.

knowledge graph (KG)

"a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities" (Hogan, et al., 2021, 1).

label

a node label defines its type, for example person, organization, location, etc.

large language model (LLM)

Large language models process input text and generate output text, on the basis of having been trained on extremely large corpora of text.

named entity recognition (NER)

Extracting entities from text – typical entities include people, places, organizations, but more broadly other items such as dates, quantities, currency, and so on.

News24

An English-language South African news site.

node

A node in a graph represents an entity – which may be a physical object, person, organization, location in the real world, or a more abstract concept.

ontology

A "formal representation of what terms mean within the scope in which they are used" (Hogan, et al., 2021, 4.1).

property

Node or edge properties augment the basic information available in order to enhance retrieval and reasoning capabilities – for example the start and end dates a position was held.

relation extraction (RE)

Identifying relations between entities or concepts, also sometimes referred to as 'triples extraction'.

retrieval augmented generation (RAG)

Grounding the responses of an LLM so that answers are generated from relevant documents from a knowledge base outside of the original training data.

semantic search

Search that attempts to identify meaning and intent, rather than simple text matching.

subject and object constraints

In a relational triple specifies what types a subject and object can take on in a specific relation

triple

A relation expressed as subject >> predicate >> object

References

Al-Moslmi, T. & Ocaña, M. G. (2020) *Lifting News into a Journalistic Knowledge Platform.* Galway, CEUR.

Allen, B. P. & Ilievski, F. (2024) Standardizing Knowledge Engineering Practices with a Reference Architecture. *Transactions on Graph Data and Knowledge,* 2(1).

Barrasa, J. & Webber, J. (2023) *Building Knowledge Graphs.* Kindle ed. Sebastopol: O'Reilly Media.

Batista, D. S. (2018) *Named-Entity evaluation metrics based on entity-level* [Online]. Available at: https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/ [July 2024].

BBC (2023) *Ontologies* [Online]. Available at: https://www.bbc.co.uk/ontologies [December 2023].

Berven, A., Christensen, O. A., Moldeklev, S. & Opdahl, A. L. (2020) A knowledge-graph platform for newsrooms. *Computers in Industry,* Volume 23. doi:https://doi.org/10.1016/j.compind.2020.103321.

Cabot, P.-L. H. & Navigli, R. (2021) *REBEL: Relation Extraction By End-to-end Language generation.* Punta Cana, Association for Computational Linguistics, p. 2370–2381. doi:https://doi.org/10.18653/v1/2021.findings-emnlp.204.

Chinchor, N. & Sundheim, B. (1993) *MUC-5 Evaluation Metrics.* Baltimore, Association for Computational Linguistics. doi:https://doi.org/10.3115/1072017.1072026.

Delpeuch, A. (2020) *OpenTapioca: Lightweight Entity Linking for Wikidata* [Online]. Available at: https://arxiv.org/abs/1904.09131 [July 2024].

Denis, P. & Baldridge, J. (2009) *Global joint models for coreference resolution and named entity classification.* Jaén, Sociedad Española para el Procesamiento del Lenguaje Natural., pp. 87-96.

Fernández, N., Fuentes, D., Sánchez, L. & Fisteus, J. A. (2010) The NEWS ontology: Design and applications. *Expert Systems with Applications*, 37(12), pp. 8694-8704. doi:https://doi.org/10.1016/j.eswa.2010.06.055.

Flair (2024) *Tagging Relations* [Online]. Available at: https://flairnlp.github.io/docs/tutorial-basics/other-models#tagging-relations [July 2024].

Gerber, E. & Mensio, M. (2023) *Spacy Entity Linker* [Online]. Available at: https://pypi.org/project/spacy-entity-linker/ [July 2024].

Gregor, S. & Hevner, A. R. (2013) Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), pp. 337-335. doi:https://doi.org/10.25300/misq/2013/37.2.01.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004) Design science in information systems research. *MIS Quarterly*, 28(1), pp. 75-105. doi:https://doi.org/10.2307/25148625.

Hogan, A. et al. (2021) *Knowledge Graphs* [Online]. Available at: https://kgbook.org/ [September 2024].

Jaradeh, M. Y. et al. (2023) Information extraction pipelines for knowledge graphs. *Knowledge and Information Systems,* 65(5). doi:https://doi.org/10.1007/s10115-022-01826-x.

Jurafsky, D. & Martin, J. H. (2024) *Speech and Language Processing* [Online]. Available at: https://web.stanford.edu/~jurafsky/slp3/20.pdf [September 2024].

Kahneman, D. (2011) *Thinking, Fast and Slow.* London: Penguin Books.

Li, J., Sun, A., Han, J. & Li, C. (2022) A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), p. 61. doi:https://doi.org/10.1109/tkde.2020.2981314.

Mahouachi, M. E. & Suchanek, F. (2020) *Extracting Complex Information from Natural Language Text: A Survey*. Galway, CEUR.

Majid, A. (2023) *At 1,500 stories per day, Mail Online is UK's most prolific news website* [Online]. Available at: https://pressgazette.co.uk/media-audience-and-business-data/at-1500-stories-per-day-mail-online-is-uks-most-prolific-news-website/ [September 2024].

Menczer, F., Fortunato, S. & Davis, C. A. (2020) *A First Course in Network Science*. Cambridge: Cambridge University Press.

Melnyk, I., Dognin, P. & Das, P. (2022) *Knowledge Graph Generation From Text*. Abu Dhabi, Association for Computational Linguistics. doi:https://doi.org/10.18653/v1/2022.findings-emnlp.116.

Ocaña, M. G. & Opdahl, A. L. (2023) A Software Reference Architecture for Journalistic Knowledge Platforms. *Knowledge-Based Systems,* 276. doi:https://doi.org/10.1016/j.knosys.2023.110750.

Opdahl, A. L. et al. (2022) Semantic Knowledge Graphs for the News: A Review. *ACM Computing Surveys,* 55(7), pp. 1-38. doi:https://doi.org/10.1145/3543508.

Otmazgin, S., Cattan, A. & Goldberg, Y. (2022) *F-coref: Fast, Accurate and Easy to Use Coreference Resolution.* Taipei, Association for Computational Linguistics, pp. 48-56.

Otmazgin, S., Cattan, A. & Goldberg, Y. (2023) *LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution.* Dubrovnik, Association for Computational Linguistics, p. 2752–2760. doi:https://doi.org/10.18653/v1/2023.eacl-main.202.

PA Media (n.d.) *SNaP Ontologies* [Online]. Available at: https://iptc.org/thirdparty/snap-ontology/ [February 2024].

Pan, J. Z., Kalo, J.-C. & Chen, J. (2023) Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge (TGDK)*, 1(1).

Paulheim, H. (2017) Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web,* 8(3), pp. 489-508. doi:https://doi.org/10.3233/SW-160218.

Peng, C., Xia, F., Naseriparsa, M. & Osborne, F. (2023) Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56 (11), p. 13071–13102. doi:https://doi.org/10.1007/s10462-023-10465-9.

President of the Republic of South Africa (2018) *Proclamation No.3 of 2018* [Online]. Available at: https://gazettes.africa/akn/za/officialGazette/government-gazette/2018-02-09/41436/eng@2018-02-09 [February 2024].

Pywikibot (2024) *pywikibot 9.3.1* [Online]. Available at: https://pypi.org/project/pywikibot/ [August 2024].

Riedel, S., Yao, L. & McCallum, A. (2010) *Modeling Relations and Their Mentions without Labeled Text.* Berlin, Heidelberg, Springer. doi:https://doi.org/10.1007/978-3-642-15939-8_10.

Schweter, S. & Akbik, A. (2020) *FLERT: Document-Level Features for Named Entity Recognition* [Online]. Available from: https://arxiv.org/pdf/2011.06993.pdf [10 April 2024].

Segura-Bedmar, I., Martínez, P. & Herrero-Zazo, M. (2013) *SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)*. Atlanta, Georgia, Association for Computational Linguistics, p. 341–350.

Singhal, A. (2012) *Introducing the Knowledge Graph: things, not strings* [Online]. Available at: https://blog.google/products/search/introducing-knowledge-graph-things-not/ [January 2024].

spaCy (n.d.) *English* [Online]. Available at: https://spacy.io/models/en [February 2024].

Tamašauskaitė, G. & Groth, P. (2023) Defining a Knowledge Graph Development Process Through a Systematic Review. *ACM Transactions on Software Engineering and Methodology,* 32(1), pp. 1-40. doi:https://doi.org/10.1145/3522586.

tollefj (2023) *corefeval - A super simple coreference evaluation tool* [Online]. Available at: https://github.com/tollefj/coreference-eval [June 2024].

Weikum, G., Dong, X. L., Razniewski, S. & Suchanek, F. (2021) Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Foundations and Trends® in Databases*, 10(2-4), pp. 108-490.

Wikidata (2023) *Wikidata:WikiProject Ontology/Modelling* [Online]. Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Modelling [December 2023].

Wikidata (2024) *Wikidata:WikiProject Ontology* [Online]. Available at: https://www.wikidata.org/w/index.php?title=Wikidata:WikiProject_Ontology&oldid=2175854176 [July 2024].

Wilton, P., McGinnis, J. & Harman, P. (2012) *SNaP Stuff Ontology* [Online]. Available at: https://iptc.org/thirdparty/snap-ontology/stuff/ [February 2024].

Wu, X. et al. (2022) A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems,* 135, pp. 364-381. doi:https://doi.org/10.1016/j.future.2022.05.014.

Zhang, Y. et al. (2017) *Position-aware Attention and Supervised Data Improve Slot Filling*. Copenhagen, Association for Computational Linguistics.

Zhao, X., Deng, Y., Yang, M. & Wang, L. (2024) A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers. *ACM Computing Surveys*, 56(11), pp. 1-39.