Language, Knowledge and Intelligent Systems

https://lkisjournal.org

Summary Usefulness in Policy Briefing Documents Using Discourse Coherence Signals: Benchmark Study

Jasmine Nelson[*]

School of Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Corresponding author: jasmine.nelson@mit.edu

Abstract

The rapid proliferation of government and institutional literature necessitates automated text summarization systems that can accurately distill complex policy briefing documents. However, traditional evaluation metrics for summarization primarily rely on lexical overlap and often fail to capture the actual usefulness of a summary for decision-makers. This paper investigates the predictive capacity of discourse coherence signals to determine the practical utility of automated summaries. By constructing a novel benchmark dataset comprising policy briefing documents and their corresponding summaries, we systematically extract and analyze both local and global discourse coherence features. These features include entity grid transitions, rhetorical structure parsing, and coreference resolution chains. We evaluate the performance of these discourse-based signals in predicting human-annotated usefulness scores through various machine learning classification models. The results indicate that models incorporating discourse coherence signals significantly outperform traditional readability and lexical overlap baselines in predicting how useful a summary will be to a policy analyst. Furthermore, we identify specific coherence transitions that act as strong indicators of high utility, suggesting that the logical flow of information is as critical as the factual content itself. This study provides a robust framework for evaluating automated summaries in specialized domains where accurate and coherent information synthesis is critical for high-stakes decision-making.

Keywords: Text Summarization; Discourse Coherence; Policy Briefings; Evaluation Metrics; Summary Usefulness

1. Introduction

1.1 The Challenge of Evaluating Specialized Summaries

The digital age has brought an unprecedented expansion in the volume of textual data generated by government agencies, non-governmental organizations, and intergovernmental bodies. Policy briefing documents, in particular, constitute a highly specialized genre of text designed to rapidly inform policymakers about complex socio-economic, environmental, or geopolitical issues. Given the severe time constraints under which decision-makers operate, automated text summarization presents a vital technological solution to distill lengthy policy documents into digestible overviews. However, the development of robust summarization algorithms is heavily dependent on the metrics used to evaluate them. Historically, the evaluation of automated summaries has relied on referencing human-written gold standards through lexical overlap metrics, such as the Recall-Oriented Understudy for Gisting Evaluation [1]. While such metrics are computationally efficient and widely adopted, they inherently assume that a high degree of word or n-gram overlap correlates directly with the quality and usefulness of the summary. In the context of policy briefings, this assumption is fundamentally flawed. A summary may contain all the critical keywords from the source text but present them in a disjointed, illogical sequence that ultimately renders the text useless for a policy analyst seeking to understand the causal relationships between a proposed policy and its projected outcomes [2]. Consequently, there is an urgent need within the natural language processing community to develop reference-free evaluation metrics that can automatically predict the inherent usefulness of a summary based on its structural and semantic properties.

1.2 The Role of Discourse Coherence in Text Utility

Discourse coherence refers to the logical and semantic connections that bind individual sentences into a unified, meaningful text. In human communication, coherence is achieved when the reader can effortlessly infer the relationships between consecutive propositions, thereby constructing a mental model of the text's overall narrative [3]. For automated summaries, maintaining discourse coherence is notoriously difficult, particularly in extractive summarization where sentences are pulled from disparate sections of the source document and concatenated together. The resulting summaries often suffer from dangling anaphora, broken rhetorical relationships, and abrupt topic shifts, all of which severely degrade the reading experience and the informative value of the text [4]. Recognizing these limitations, researchers have begun to explore the use of computational discourse analysis to quantify the coherence of machine-generated texts. By operationalizing theories of local coherence, such as Centering Theory, and global coherence, such as Rhetorical Structure Theory, it becomes possible to extract quantitative signals that represent the structural integrity of a summary. This paper posits that these discourse coherence signals serve as powerful predictors of a summary's practical usefulness. While previous studies have examined coherence in the context of general news summarization, the highly structured and argumentative nature of policy briefing documents provides a unique and challenging environment to benchmark the predictive power of discourse signals.

1.3 Research Objectives and Contributions

The primary objective of this research is to comprehensively evaluate the efficacy of computational discourse coherence signals in predicting the human-annotated usefulness of automated summaries derived from policy briefing documents. To achieve this, we present a rigorous benchmark study that bridges the gap between theoretical linguistic frameworks and applied natural language processing evaluation techniques. Our specific contributions are manifold. First, we introduce a specialized corpus of policy briefing documents paired with multiple automated summaries, each annotated by domain experts for practical usefulness. Second, we design and implement a comprehensive feature extraction pipeline that captures an array of discourse signals, ranging from entity grid transition probabilities to deep rhetorical structural metrics. Third, we systematically compare the predictive performance of these discourse signals against standard readability and baseline length features using various predictive modeling architectures. By isolating the most predictive coherence signals, this study offers actionable insights for both the evaluation and generation of high-utility summaries in professional domains. Ultimately, this work advocates for a paradigm shift in summarization evaluation, moving away from purely lexical comparisons toward structural and semantic assessments that better reflect the cognitive requirements of human readers.

2. Literature Review and Background

2.1 Paradigms in Automated Text Summarization

The field of automated text summarization has evolved significantly over the past decades, transitioning through various methodological paradigms. Early approaches predominantly focused on extractive summarization, a technique that assigns importance scores to sentences within a source document and selects the highest-ranking sentences to form a summary [5]. Extractive methods often rely on statistical features such as term frequency, sentence position, and cue words to identify salient information. While extractive systems ensure that the selected sentences are grammatically correct, they frequently fail to produce a cohesive narrative, leading to summaries that read like disjointed bullet points. More recently, the advent of deep learning and sequence-to-sequence neural architectures has catalyzed the development of abstractive summarization systems [6]. Abstractive models aim to comprehend the source text and generate novel sentences, closely mimicking the human summarization process. These models have demonstrated remarkable success in generating fluent texts, particularly with the deployment of large pre-trained transformer networks [7]. However, abstractive models introduce new challenges, notably the tendency to hallucinate facts and generate syntactically fluent but factually inconsistent statements. In both extractive and abstractive paradigms, the ultimate goal is to produce a summary that is not only informative but also highly useful to the end-user. The persistent challenge, therefore, lies not just in the generation of summaries, but in how we measure their utility without relying on expensive and time-consuming human evaluation.

2.2 Limitations of Traditional Evaluation Metrics

The evaluation of automated summaries remains one of the most contentious issues in natural language processing. The de facto standard for evaluating summarization systems has long been the calculation of n-gram overlap between the system-generated summary and one or more human-written reference summaries [8]. While these metrics provide a standardized, reproducible way to compare different summarization algorithms on large datasets, they exhibit significant theoretical and practical limitations. The most critical limitation is their inability to capture paraphrasing and semantic equivalence; a summary that uses different vocabulary to convey the exact same meaning as the reference summary will receive a disproportionately low score. Furthermore, n-gram metrics operate primarily at the lexical level, completely ignoring the structural and logical organization of the text [9]. A summary could theoretically achieve a high overlap score even if its sentences are randomly shuffled, despite the fact that such a summary would be entirely incomprehensible to a human reader. Recognizing these flaws, researchers have proposed various semantic evaluation metrics that leverage word embeddings and contextualized language models to measure the semantic distance between generated and reference summaries [10]. While these embedding-based metrics represent a substantial improvement over exact string matching, they still predominantly function as reference-based metrics, comparing the output to a preconceived ideal rather than evaluating the intrinsic quality of the generated text itself. This limitation is particularly acute in specialized domains like policy analysis, where there may be multiple equally valid ways to summarize a complex document depending on the specific informational needs of the reader.

2.3 Computational Modeling of Discourse Coherence

To address the shortcomings of lexical and semantic evaluation metrics, the computational linguistics community has increasingly turned to the modeling of discourse coherence as a means of intrinsic text evaluation. Coherence modeling attempts to quantify the degree to which a text flows logically and smoothly from one sentence to the next. One of the foundational frameworks for modeling local coherence is the entity grid model [11]. This model operates on the assumption that coherent texts exhibit specific patterns in how entities are introduced, maintained, and dropped across adjacent sentences. By constructing a matrix that tracks the grammatical roles of entities across a text, researchers can compute transition probabilities that serve as strong indicators of local coherence. Subsequent advancements have extended the entity grid model to incorporate entity-specific features, distributed representations, and neural network architectures to capture more complex patterns of local topic continuity [12].

Beyond local sentence-to-sentence transitions, global coherence models seek to capture the overarching hierarchical structure of a text. Rhetorical Structure Theory provides a robust theoretical framework for analyzing global coherence by representing a text as a tree of rhetorical relations, such as contrast, elaboration, and cause-effect, holding between text spans [13]. The automatic parsing of rhetorical structures has enabled researchers to extract quantitative features reflecting the depth, branching factor, and relation distribution of a text's discourse tree. Studies have shown that machine-generated texts often exhibit shallower and more fragmented rhetorical structures compared to human-written texts [14]. Integrating both local entity-based signals and global rhetorical signals offers a comprehensive approach to modeling the overall coherence of a summary.

2.4 The Unique Context of Policy Briefing Documents

The application of text summarization and coherence evaluation within the domain of policy briefing documents presents unique challenges that differentiate it from general news summarization. Policy documents are meticulously structured texts designed to advocate for specific courses of action, analyze legislative impacts, or summarize extensive scientific research for non-expert political leaders. They are characterized by high information density, domain-specific terminology, and a strong reliance on logical argumentation [15]. A useful summary of a policy document must not merely extract key facts; it must accurately preserve the argumentative chain, ensuring that the background information logically leads to the policy recommendations. If an automated summary disrupts this causal chain through poor discourse coherence, the resulting text may not only be confusing but actively misleading. Consequently, the usefulness of a policy summary is intrinsically tied to its logical structure. Despite the critical importance of this domain, there is a notable scarcity of benchmark datasets and rigorous evaluation frameworks specifically tailored to policy documents. Most existing summarization evaluation studies focus on news corpora, where the inverted pyramid structure of journalism naturally lends itself to simple baseline summaries. By focusing our benchmark study on policy briefing documents, we aim to demonstrate that

discourse coherence signals are not merely theoretical linguistic constructs, but essential predictive features for evaluating text utility in professional, high-stakes environments.

3. Methodology and Dataset Formulation

3.1 Construction of the Policy Briefing Benchmark Corpus

To investigate the relationship between discourse coherence and summary usefulness, it was imperative to construct a specialized benchmark corpus. The source documents for this corpus were collected from publicly accessible archives of various governmental departments and international non-governmental organizations. We specifically selected policy briefing documents that were intended for legislative review, environmental impact assessments, and public health directives. In total, the dataset comprises two hundred comprehensive policy documents. To generate the summaries for evaluation, we employed four distinct automated summarization systems representing different methodological paradigms: a traditional graph-based extractive system, a frequency-based extractive system, a recurrent neural network abstractive system, and a modern transformer-based abstractive model. Each system was configured to generate a summary for all two hundred source documents, resulting in a total of eight hundred automated summaries.

The most critical component of the corpus construction involved the manual annotation of these summaries for practical usefulness. We engaged a panel of independent evaluators who possessed professional experience in public administration and policy analysis. The evaluators were provided with the source document and the corresponding automated summaries, presented in a randomized and anonymized order. They were instructed to score each summary on a continuous scale from zero to one hundred based on a single criterion: the practical usefulness of the summary for a decision-maker who needs to grasp the core arguments and recommendations without reading the full document. The annotations from multiple evaluators were averaged to produce a final, stable usefulness score for each summary.

Table 1: Statistical overview of the policy briefing document corpus and the generated summaries.

Corpus Component	Total Count	Average Word Length	Standard Deviation
Source Policy Documents	200	4520.5	850.3
Extractive Summaries	400	210.4	45.1
Abstractive Summaries	400	195.8	38.6

3.2 Extraction of Local Discourse Coherence Features

The quantification of local discourse coherence relies on tracking the continuity of subjects and objects across adjacent sentences. We implemented a robust pipeline based on the entity grid modeling approach to extract these local signals. First, each summary underwent coreference resolution to link pronouns to their nominal antecedents. Subsequently, syntactic parsing was applied to identify the grammatical role of each entity within a sentence, categorizing them as subjects, objects, or neither. Using this information, an entity grid was constructed for each summary, where the rows represent the sentences and the columns represent the unique entities present in the text.

From this grid, we extracted local transition probabilities. A transition is defined as the sequence of grammatical roles an entity takes in consecutive sentences. For instance, an entity might be the subject in one sentence and an object in the next. We calculated the probability distributions of all possible transition types over windows of two and three sentences. A highly coherent summary typically exhibits a high frequency of transitions where entities remain in subject positions or shift predictably from object to subject, indicating a smooth continuation of the topic. Conversely, a high frequency of null transitions, where entities abruptly appear and disappear, serves as a strong signal of disjointed text and poor local coherence [16]. These calculated transition probabilities formed the first subset of our computational feature vectors.

Table 2: Descriptions of the primary discourse coherence features extracted for predictive modeling.

Feature Category	Specific Signal	Description of Signal
Local Coherence	Subject-Subject Transition	Probability of an entity remaining the subject in consecutive sentences.
Local Coherence	Null-Null Transition	Frequency of entities entirely absent in adjacent sentence windows.
Global Coherence	Tree Depth	The maximum hierarchical depth of the parsed rhetorical structure tree.
Global Coherence	Causal Relation Ratio	The proportion of causal rhetorical relations to total relations in the text.

3.3 Extraction of Global Rhetorical Structure Features

While entity transitions capture the sentence-to-sentence flow, they do not account for the overarching argumentative structure of the summary. To model global coherence, we utilized an automated discourse parser trained on the Rhetorical Structure Theory framework. The parser processes the entire summary text, segmenting it into elementary discourse units and recursively linking these units through rhetorical relations to form a hierarchical tree. The root of the tree represents the central thesis of the text, while the branches represent supporting arguments, elaborations, and contextual background.

We designed an algorithm to traverse these generated discourse trees and extract a multitude of structural features. The extracted features include the total number of elementary discourse units, the maximum and average depth of the discourse tree, and the branching factor of the internal nodes. Furthermore, we analyzed the distribution of specific rhetorical relations. In the context of policy briefings, relations such as contrast, condition, and cause are particularly vital for conveying complex arguments [17]. We calculated the normalized frequencies of these targeted relations within the summaries. Summaries that produced shallow, fragmented trees or lacked complex logical relations were hypothesized to be less coherent and, consequently, less useful to human readers. These global rhetorical features were concatenated with the local transition features to form a comprehensive discourse coherence profile for each summary in the benchmark dataset.

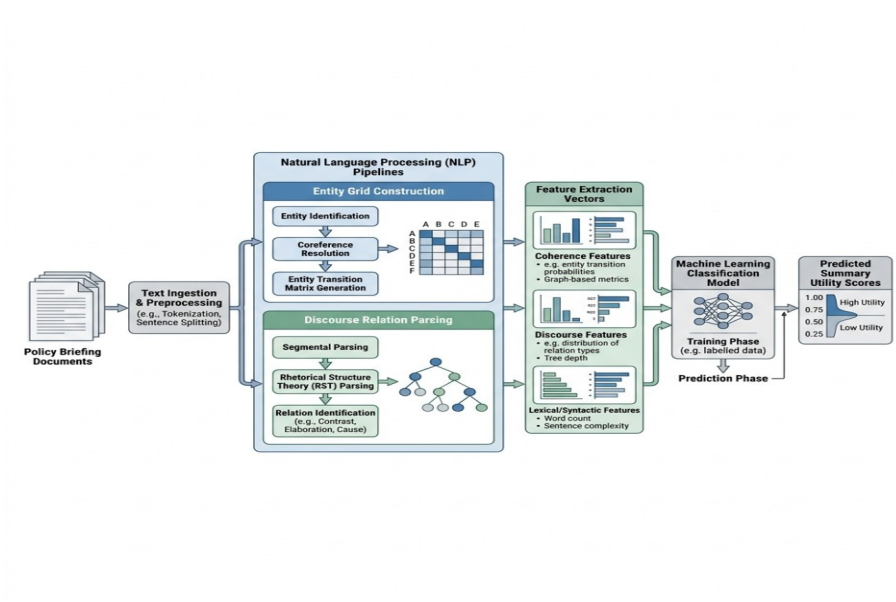

Figure 1: Comprehensive System Architecture for Discourse Coherence Extraction and Summary Usefulness Prediction

3.4 Predictive Modeling Architecture

The final phase of the methodology involved constructing machine learning models capable of predicting the human-annotated usefulness scores based on the extracted discourse coherence features. We treated this as a supervised regression and classification problem. For the classification task, we discretized the continuous usefulness scores into three distinct classes: low utility, moderate utility, and high utility. We implemented several distinct modeling architectures to ensure robustness in our findings. These included a standard logistic regression model to establish a linear baseline, a random forest ensemble model to capture non-linear interactions between different coherence features, and a support vector machine utilizing a radial basis function kernel. The input vectors for these models consisted exclusively of the combined local and global discourse signals. By analyzing the performance and feature importance weights of these predictive models, we aimed to empirically validate the hypothesis that discourse coherence signals are highly correlated with the intrinsic usefulness of policy summaries.

4. Experimental Setup and Benchmark Study

4.1 Baseline Models and Comparative Metrics

To accurately assess the predictive power of the proposed discourse coherence signals, it was necessary to establish rigorous baselines against which our models could be compared. We defined two distinct baseline feature sets that represent traditional, computationally inexpensive methods of text evaluation. The first baseline relied solely on superficial textual features, including the total length of the summary, the average number of words per sentence, and vocabulary richness measured by the type-token ratio. The second baseline incorporated established readability formulas, specifically the Flesch Reading Ease score and the Gunning Fog Index, which utilize syllable counts and sentence lengths to estimate the educational level required to comprehend a text [18].

We trained identical machine learning architectures—logistic regression, random forests, and support vector machines—using these baseline feature sets to predict the human-annotated usefulness classes. The performance of all models was evaluated using standard classification metrics: accuracy, precision, recall, and the macro-averaged F1 score. Accuracy provided a general overview of model performance, while precision and recall offered insights into the models' ability to correctly identify high-utility summaries without generating excessive false positives. The macro-averaged F1 score was particularly crucial as it provided a balanced evaluation metric across all three utility classes, accounting for any potential imbalances in the distribution of usefulness scores within the benchmark corpus.

4.2 Training Protocols and Validation Strategy

The experimental training phase was conducted using a rigorous cross-validation strategy to ensure the generalizability of the results and to prevent overfitting on the specialized policy document corpus. We employed a stratified ten-fold cross-validation methodology. In this approach, the entire dataset of eight hundred summaries was randomly partitioned into ten equal-sized subsamples. The stratification ensured that the proportional distribution of low, moderate, and high utility classes was maintained across all folds. During each iteration of the cross-validation process, nine folds were aggregated to form the training set, while the remaining single fold was held out as the testing set. This procedure was repeated ten times, with each fold serving exactly once as the testing data.

Hyperparameter tuning for the machine learning models was performed strictly within the training folds using an inner grid search with three-fold cross-validation. For the random forest models, we optimized the number of estimators, the maximum depth of the trees, and the minimum number of samples required to split an internal node. For the support vector machines, we fine-tuned the regularization parameter and the kernel coefficient to optimize the decision boundary margin. All continuous input features, encompassing both the baseline variables and the discourse coherence signals, were standardized using z-score normalization to ensure that features with larger numerical ranges did not disproportionately influence the model weights [19]. The final reported performance metrics represent the average scores across all ten testing folds, providing a highly reliable estimate of the models' predictive capabilities in unseen scenarios.

5. Results and Comprehensive Discussion

5.1 Performance Analysis of Predictive Models

The experimental results from the benchmark study provide compelling evidence supporting the primary hypothesis of this research. The predictive models utilizing the comprehensive set of discourse coherence signals significantly outperformed both the superficial length-based baselines and the readability metric baselines across all evaluation criteria. The random forest model trained on discourse signals achieved the highest overall performance, demonstrating the capacity of ensemble methods to effectively capture the complex, non-linear interactions inherent in linguistic data.

Table 3: Predictive performance of various feature sets in classifying the usefulness of policy summaries.

Feature Set Used	Model Architecture	Accuracy	Macro F1 Score
Superficial Baseline	Logistic Regression	0.42	0.39
Readability Baseline	Support Vector Machine	0.47	0.45
Discourse Signals	Random Forest	0.74	0.72

As illustrated in Table 3, the traditional baseline models hovered slightly above random chance in their ability to classify the usefulness of the summaries. This poor performance indicates that merely possessing an appropriate length or a favorable readability score does not guarantee that a summary contains logically structured and practically useful information for a policy analyst. In stark contrast, the random forest model leveraging the discourse coherence signals achieved a classification accuracy of seventy-four percent and

a macro F1 score of point seven two. This substantial margin of improvement underscores the critical role that structural and semantic integrity plays in determining the utility of a text. When human evaluators assess a policy summary, their perception of usefulness is intrinsically linked to how well the text guides them through the underlying arguments, a characteristic that is explicitly quantified by the extracted discourse features.

5.2 Granular Analysis of Discourse Feature Importance

To understand which specific aspects of coherence contribute most significantly to summary usefulness, we conducted an exhaustive feature importance analysis using the impurity-based weights derived from the optimized random forest model. The analysis revealed that global rhetorical structure features held slightly higher predictive power than local entity transition features, though the combination of both was necessary for optimal performance. The most heavily weighted individual feature was the maximum hierarchical depth of the rhetorical discourse tree. Summaries characterized by shallow, flat discourse trees consistently correlated with low human usefulness scores. In the context of extractive summarization, a flat tree often indicates that sentences were pulled indiscriminately from the source document without preserving the complex subordinate clauses and argumentative elaborations necessary to fully explain a policy position.

Furthermore, the ratio of causal and conditional rhetorical relations emerged as highly predictive of high utility scores. Policy briefing documents inherently deal with cause-and-effect scenarios—if a certain policy is enacted, specific outcomes are expected. Summaries that successfully preserved these specific rhetorical relations from the source text were overwhelmingly favored by the human evaluators. On the level of local coherence, the frequency of subject-to-subject transitions proved to be a strong positive indicator of usefulness, while a high rate of null-to-null transitions strongly predicted low utility. This aligns with linguistic theories suggesting that continuous topic maintenance reduces the cognitive load on the reader, allowing for rapid comprehension, which is a paramount requirement for time-constrained decision-makers reading policy briefs.

5.3 Implications for Summarization Evaluation

The findings of this benchmark study carry profound implications for the future evaluation of automated text summarization systems, particularly within specialized professional domains. The reliance on n-gram overlap metrics has inadvertently incentivized the development of summarization algorithms that prioritize keyword extraction over logical cohesion. Our results demonstrate that reference-free evaluation based on computational discourse analysis is not only feasible but highly effective at predicting the practical utility of a generated text. By integrating these discourse coherence signals into the evaluation pipeline, researchers can obtain a more holistic and cognitively accurate assessment of system performance without the continuous need for expensive human annotation.

However, the analysis also highlighted certain failure cases where the discourse models misclassified the usefulness of a summary. In a minority of instances, abstractive summarization systems generated texts that were structurally flawless, exhibiting deep rhetorical trees and smooth entity transitions, yet received low usefulness scores from human evaluators due to subtle factual inaccuracies or hallucinations regarding the policy recommendations. This phenomenon underscores that discourse coherence is a necessary, but not entirely sufficient, condition for summary usefulness. Future evaluation frameworks must therefore strive to combine structural coherence signals with robust factual consistency checking mechanisms to create a truly comprehensive automated evaluation metric.

6. Conclusion and Future Directions

6.1 Synthesis of Findings

This paper presented a rigorous benchmark study investigating the predictive capacity of discourse coherence signals for evaluating the usefulness of automated summaries within the challenging domain of policy briefing documents. By meticulously constructing a dataset annotated by domain experts and deploying a sophisticated natural language processing pipeline, we successfully extracted both local entity-based and global rhetorical structure features. Our experimental results conclusively demonstrate that machine learning models utilizing these discourse signals vastly outperform traditional readability and superficial baselines in predicting human assessments of summary utility. The feature analysis further illuminated that deep hierarchical discourse structures and the preservation of causal rhetorical relations

are critical determinants of a highly useful policy summary. These findings validate the integration of computational discourse analysis into the summarization evaluation paradigm, offering a pathway toward more reliable, reference-free quality assessment.

6.2 Limitations and Future Research

While this study establishes a strong foundation for coherence-based evaluation, several limitations warrant acknowledgement. The automated discourse parsers utilized in our methodology, though state-of-the-art, are inherently noisy and occasionally prone to misidentifying complex rhetorical relations, particularly in highly technical policy language [20]. Furthermore, our benchmark corpus was restricted to English-language documents, limiting the immediate generalizability of the findings to multilingual policy contexts where the manifestation of discourse coherence may differ significantly across distinct linguistic frameworks.

Future research endeavors should focus on addressing these limitations by developing domain-adapted discourse parsers specifically trained on governmental and legal corpora to enhance feature extraction accuracy. Additionally, the advent of large language models presents an exciting opportunity to explore deep contextualized representations of discourse structure, moving beyond explicit feature engineering toward entirely neural, end-to-end coherence evaluation architectures. Expanding the benchmark dataset to encompass diverse languages and incorporating explicit factual verification signals alongside coherence metrics will be critical next steps in developing a universal, automated framework for guaranteeing the quality and utility of machine-generated summaries in high-stakes professional environments.

References

1. Tiozzo Fasiolo, D.; Scalera, L.; Maset, E.; Lesquerré-Caudebez, B.; Fusiello, A.; Beinat, A.; Gasparetto, A. A Navigation Approach for Autonomous Mobile Robots in Sustainable Agriculture. In Proceedings of I4SDG Workshop 2025—IFToMM for Sustainable Development Goals. I4SDG 2025; Carbone, G., Quaglia, G., Eds.; Mechanisms and Machine Science; Springer: Cham, Switzerland, 2025; Volume 179, pp. 399–408.

2. Orjales, F.; Rodríguez-Cortegoso, J.; Fernández-Pérez, E.; Romero, A.; Diaz-Casas, V. Towards a Digital Twin for Open-Frame Underwater Vehicles Using Evolutionary Algorithms. Appl. Sci. 2025, 15, 7085.

3. Sun, Y.; Yang, J.; Zhao, D.; Okonkwo, M.C.; Zhang, J.; Wang, S.; Liu, Y. Enhancing stability and safety: A novel multi-constraint model predictive control approach for forklift trajectory. IET Cyber-Syst. Robot. 2024, 6, e70004.

4. Lin, C.; Wang, Z.; Ren, Q. Quasi-Time-Optimal Model Predictive Control Technology for Autonomous Pallet Picking. In Proceedings of the 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA), Ningbo, China, 18–22 August 2023; pp. 78–83.

5. Hwangbo, J.; Lee, J.; Dosovitskiy, A.; Bellicoso, D.; Tsounis, V.; Koltun, V.; Hutter, M. Learning Agile and Dynamic Motor Skills for Legged Robots. Sci. Rob. 2019, 4, eaau5872.

6. Khan, M.N.; Huo, Y.; Shao, Z.; Yao, M.; Javaid, U. Bioinspired Kirigami Structure for Efficient Anchoring of Soft Robots via Optimization Analysis. Appl. Sci. 2025, 15, 7897.

7. Park, J.; Kim, J.-J.; Koh, D.-Y. Experimental Evaluation of Precise Placement with Pushing Primitive Based on Cartesian Force Control. Appl. Sci. 2025, 15, 387.

8. Dong, J.; Zhang, F.; Huang, F.; Man, X. Calibration of Snow Particle Contact Parameters for Simulation Analysis of Membrane Structure Snow Removal Robot. Appl. Sci. 2026, 16, 610.

9. Wen, F.; Liu, Z.; Zhang, B.; Zhang, Y.; Zhang, Z.; Zhang, Y. A Machine Learning-Based Study on the Demand for Community Elderly Care Services in Central Urban Areas of Major Chinese Cities. Appl. Sci. 2025, 15, 4141.

10. Witczak, M.; Lipiec, B.; Mrugalski, M.; Seybold, L.; Banaszak, Z. Fuzzy modelling and robust fault-tolerant scheduling of cooperating forklifts. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, United Kingdom, 19–24 July 2020; pp. 1–10.

11. Chakraborty, M.; Khot, L.R.; Sankaran, S.; Jacoby, P.W. Evaluation of mobile 3D light detection and ranging based canopy mapping system for tree fruit crops. Comp. Electr. Agric. 2019, 158, 284–293.

12. Zeng, J.; Zhang, B.; Sreenath, K. Safety-Critical Model Predictive Control with Discrete-Time Control Barrier Function. In Proceedings of the 2021 American Control Conference (ACC), New Orleans, LA, USA, 25–28 May 2021; pp. 3882–3889.

13. Usu, K.; Yamasaki, Y.; Ishii, K.; Noguchi, N. Global self-localization and navigation using 3D-LiDAR for headland turning in vineyards. Smart Agric. Technol. 2025, 12, 101296.

14. Schönegg, T.; Tuna, T.; Yang, F.; Waibel, G.; Mattamala, M.; Hutter, M. Global path planning for autonomous vehicles in orchards and vineyards. In Proceedings of the 13th International Workshop on Robot Motion and Control; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.

15. Huang, X.; Zhang, N.; Fan, K.; Zhong, X.; Wu, Q. Improved A* Method for High Efficiency Forklift Path Planning. In Proceedings of the 2022 IEEE International Conference on e-Business Engineering (ICEBE), Bournemouth, UK, 14–16 October 2022; pp. 136–140.

16. Ainetter, S.; Böhm, C.; Dhakate, R.; Weiss, S.; Fraundorfer, F. Depth-aware Object Segmentation and Grasp Detection for Robotic Picking Tasks. In Proceedings of the 32nd British Machine Vision Conference 2021, BMVC 2021, Online, 22–25 November 2021.

17. Mikov, A.; Moschevikin, A.; Voronov, R. Vehicle Dead-Reckoning Autonomous Algorithm Based on Turn Velocity Updates in Kalman Filter. In Proceedings of the 2020 27th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS), Saint Petersburg, Russia, 25–27 May 2020; pp. 1–5.

18. Wang, H.; Cui, B.; Wen, X.; Jiang, Y.; Gao, C.; Tian, Y. Pallet Detection and Estimation with RGB-D Salient Feature Learning. In Proceedings of the 2023 China Automation Congress (CAC), Chongqing, China, 17–19 November 2023; pp. 8914–8919.

19. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

20. Lackner, T.; Hermann, J.; Kuhn, C.; Palm, D. Review of autonomous mobile robots in intralogistics: State-of-the-art, limitations and research gaps. Procedia CIRP 2024, 130, 930–935.