Making Revisions Understandable: A Survey of Edit Intentions, Methods, and Applications

Fangping Lan **Qi Zhang** **Eduard C. Dragut**
Temple University
{fangping.lan, qi.zhang, edragut}@temple.edu

Abstract

Text revision is a core process in document creation, capturing how authors iteratively refine, reorganize, and improve written content. With the increasing availability of large-scale revision histories from platforms such as Wikipedia and arXiv, NLP research has begun to move beyond modeling what changes are made to understanding why they are made, i.e., the underlying edit intentions. To our knowledge, this is the *first* survey[1] that synthesizes text revision research through the lens of edit intentions, providing a unified view of datasets, taxonomies, identification methods, and applications. We review prior work across the full revision workflow, including revision corpus construction, edit intention taxonomy design, and edit intention identification. We further categorize representative datasets and methods, summarize downstream applications such as writing assistance and document edit summarization, and highlight key open research directions.

1 Introduction

Text revision is a fundamental process in document creation, reflecting how authors iteratively refine content, correct errors, and reshape meaning over time. Unlike final drafts, revisions preserve explicit traces of authors' decision-making, making them a valuable source for studying writing behavior and textual evolution (Faigley and Witte, 1981a). The widespread availability of large-scale revision histories from platforms such as Wikipedia (Daxenberger and Gurevych, 2012) and arXiv (Du et al., 2022a; Jiang et al., 2022) has further enabled empirical study of revision processes in naturalistic settings (Zhang et al., 2017; Kashefi et al., 2022).

Beyond identifying *what* textual changes occur, understanding *why* authors make these changes, referred to as *edit intentions*, is crucial for interpreting revisions at a semantic and pragmatic level (Yang et al., 2017). Edit intentions provide a high-level abstraction over surface-level differences (Zhang et al., 2016), capturing the underlying goals of revisions such as content elaboration, factual correction, stylistic refinement, or structural reorganization (Spangher et al., 2022). This perspective has catalyzed a range of NLP applications, including writing assistance (Zhang and Litman, 2015), text simplification (Laban et al., 2023), and document edit summarization (Ruan et al., 2024). Beyond revision settings, recent work shows that intent taxonomies can generalize to other forms of content interaction, such as modeling the communicative role of hyperlinks in social media posts (Lan et al., 2026).

Prior works cover writing processes, collaborative editing, and text generation, but none treat *edit intentions* as a first-class object across the revision workflow. Early composition studies distinguish surface-level versus meaning-changing edits yet predate computational models, large-scale corpora, and standardized evaluation protocols (Sommers, 1980; Faigley and Witte, 1981a). Surveys of collaborative writing (e.g., Wikipedia) primarily examine editor behavior and quality control, using edit types mainly as auxiliary signals rather than modeling intentions (Pfeil et al., 2006; Halfaker et al., 2013). More recent NLP surveys emphasize *single-shot* transformations such as style transfer or simplification, without considering revision histories, alignment and differencing, or taxonomy design (Jin et al., 2022). Meanwhile, dataset- and method-specific studies introduce corpora or models for edit intention identification or revision summarization, but do not synthesize shared design choices, annotation practices, or methodological trade-offs across domains and settings (Zhang and Litman, 2015; Spangher et al., 2022; Ruan et al., 2024).

Thus, the literature remains fragmented along several recurring axes. First, choices in revision workflows (e.g., segmentation granularity, version

[1]The GitHub repository is available at https://github.com/TUDMLab/MakeRevisionsUnderstandable

Findings of the Association for Computational Linguistics: ACL 2026, pages 35003–35019

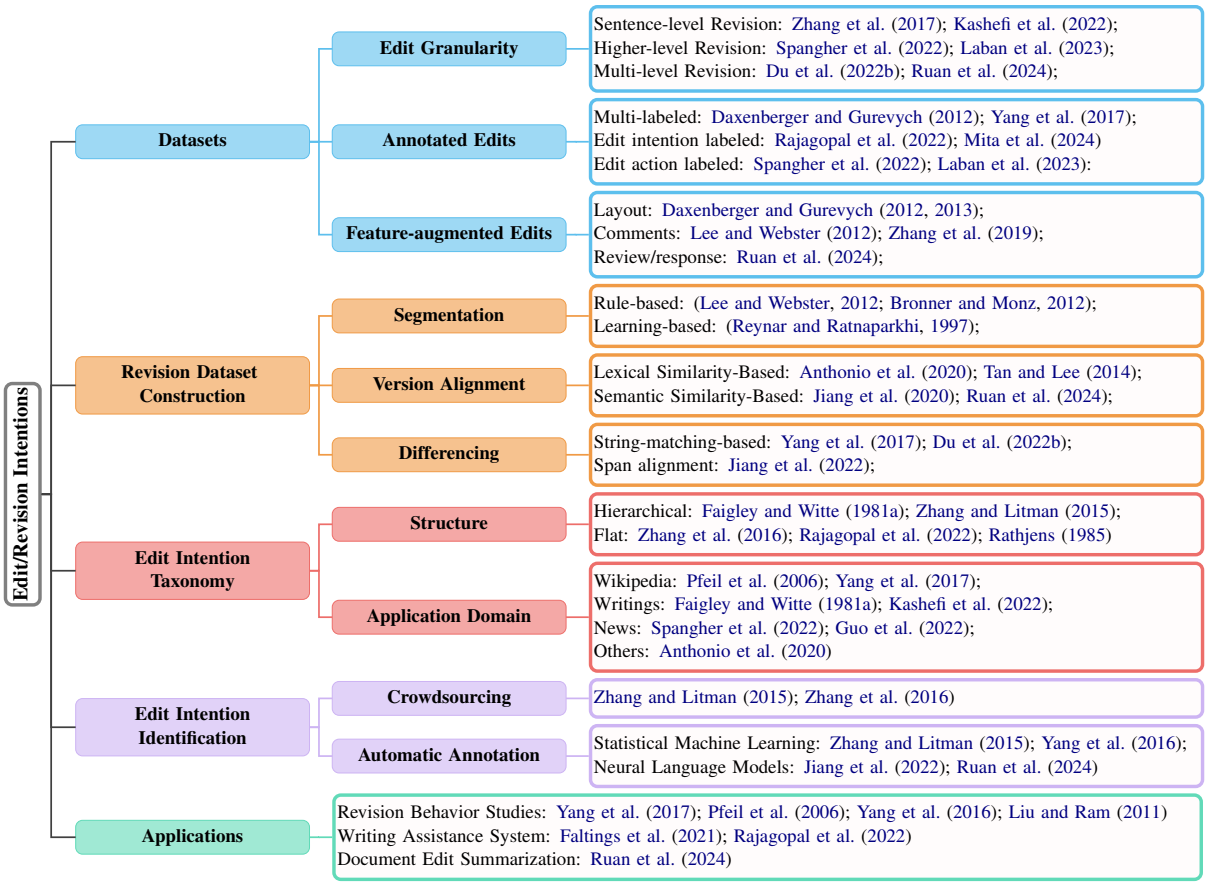

Figure 1: Taxonomy of edit intention-related research. We only list representative works for each kind of task, and for a more complete version, please refer to Appendix A).

alignment, and differencing) directly determine what can be reliably labeled and learned downstream. Second, edit intention taxonomies (EITs) are often domain-specific and defined with varying levels of granularity and scope, hindering cross-corpus comparability. Third, identification methods and evaluations are frequently tied to particular label sets and protocols, limiting reproducibility and cross-domain generalization. To address these challenges, we adopt an *edit intention–centric* perspective that integrates revision corpus construction, EIT design, edit intention identification, and downstream applications under a unified view. To our knowledge, this is the *first* survey that synthesizes text revision research explicitly through the lens of edit intentions. We consolidate terminology and systematically organize prior work across datasets, corpus construction workflows, taxonomy design, identification methods, applications, and shared open challenges. We first identified seed papers using keyword-based searches (e.g., "edit intention", "revision taxonomy", "writing revision") across major digital libraries. We then performed backward and forward snowballing by examining references and citing papers to expand coverage. We included papers that either (1) propose, refine, or analyze an EIT, or (2) develop downstream tasks that explicitly leverage edit intentions. We iterated this process until no additional eligible works were identified.

This survey aims to provide a unified and structured view of text revision research through the lens of edit intentions. Specifically, we: (i) establish consistent terminology for revision-related concepts; (ii) systematically categorize datasets and methods across the revision pipeline; (iii) analyze approaches to edit intention taxonomy construction and identification; (iv) summarize empirical findings across key application domains; and (v) identify promising open directions for future research. In summary, our contributions are:

- **Unified framework.** We consolidate terminology and organize prior work across (1) the full revision workflow, (2) datasets and corpus construction, (3) EITs, (4) identification methods, (5) applications, and (6) open challenges, into a uni-

fied *six-dimensional* view. Figure 1 presents the first 5 views.

- **Lineage-aware taxonomy analysis.** We introduce a *lineage-aware* perspective to characterize how EITs evolve across domains and granularities, and summarize common operations such as merging, splitting, and refinement.
- **Methods-to-applications map.** We review edit intention identification approaches spanning manual annotation, crowdsourcing, neural models, and LLM-based methods, and connect them to downstream uses including writing assistance, revision behavior analysis, and document edit summarization, highlighting methodological trade-offs and evaluation pitfalls.

2 Formulation

We introduce core definitions and notation to clarify key concepts in revision analysis and edit intention modeling and to support consistent comparison across datasets, methods, and applications.

An *edit* is a single, coherent change to a document, formalized as the insertion, deletion, or substitution of a sub-expression p such that both the original text s^{t-1} and the revised text $s^t = e(s^{t-1})$ are well-formed semantic constituents (MacCartney, 2009; Jiang et al., 2022). For example, inserting p = "in 1949" into s^{t-1} = "She died from an illness" yields s^t = "She died in 1949 from an illness." This formulation captures the semantic contribution of p independent of the underlying representation.

A *revision* occurs when an editor saves changes to a document (Yang et al., 2016, 2017; Du et al., 2022b). Revisions may occur at the sentence (S), paragraph (P), or document (D) level, and a single revision can contain multiple edits. We denote a revision at granularity $g \in \{D, P, S\}$ between versions $t-1$ and t as $R^{t,g}$, relating the original expression g^{t-1} to its revised form g^t.

A *document-level revision* $R^{t,D}$ corresponds to a pair of document versions (D^{t-1}, D^t). It comprises $|R^{t,D}|$ paragraph-level revisions, denoted $\{R_i^{t,P}\}$, where i indexes paragraph-level revisions. Each paragraph-level revision $R_i^{t,P}$ may in turn contain $|R_i^{t,P}|$ sentence-level revisions, denoted $\{R_{ij}^{t,S}\}$. Finally, each sentence-level revision $R_{ij}^{t,S}$ consists of one or more edits $\{e_k\}$.

An *edit action* a_k specifies the operation applied to a text object during an edit (Du et al., 2022b), such as *insert*, *delete*, *merge*, or *split*, where k indexes the set of possible edit actions.

An *edit intention* I_k represents the editor's underlying goal when performing a specific edit. We assume that each edit action a_k is associated with a single edit intention I_k. Section 5 discusses the categorization of edit intentions.

Version alignment establishes correspondences between two versions of the same text. Given an original document or paragraph at version $t-1$ and its revised version at version t, the goal is to map each paragraph or sentence in the original text to its corresponding unit in the revised text, if such a correspondence exists.

Given an edit intention taxonomy $EIT = \{I_1, \ldots, I_k\}$ and an original-revised text pair (g^{t-1}, g^t) with m edits, *edit intention identification* aims to assign the most likely intention $I_i \in EIT$ to each edit e_i. At the sentence level, this is typically framed as a multi-class classification task, whereas at the paragraph and document levels it is modeled as multi-class multi-label classification, allowing multiple intentions per revision.

3 Datasets

We review publicly released datasets for text revision and edit intention research, excluding datasets if they do not contain annotated edit actions or edit intentions. The remaining corpora are primarily derived from Wikipedia, academic writing, and student essays. In this section, we organize existing datasets along three dimensions: *edit granularity*, which determines the level at which revisions are analyzed; *ground-truth labels*, which are essential for modeling edit actions and intentions; and *feature augmentation*, which provides additional contextual signals useful for edit intention identification and revision modeling. Table 1 summarizes the statistics of the collected datasets[2].

3.1 Edit Granularity

Sentence-level Revision Early work focuses on sentence-level revisions, constructing sentence-pair corpora, primarily from Wikipedia, to distinguish meaning-preserving and meaning-changing edits (Daxenberger and Gurevych, 2012, 2013). Subsequent studies move beyond surface operations to annotate semantic edit intentions, capturing the motivations behind revisions (Yang et al., 2017), including atomic insertion edits collected across

[2]While we aim to cover all publicly available datasets, we acknowledge the possibility of missing some relevant work.

Paper	Gran.	# Pairs	H.	I.	A.	Source	Features
Daxenberger and Gurevych (2013)♦	S	2K	✓	✓	✓	Wikipedia	Including layout
Yang et al. (2017)	S	5.7K	✓	✓	✗	Wikipedia	Multiple labels
Faruqui et al. (2018)	S	43M	✓	✓	✗	Wikipedia	Multilingual
Kashefi et al. (2022)♦	S	3K	✓	✓	✓	Student Essays	Revision history, Multiple labels
Kim et al. (2022)♦	P, S	367K	✓	✓	✓	ArXiv, Wikipedia WikiNews	Across domain, Revision history, Including edit tagging
Zhang et al. (2019)	Sec	786K	✗	✗	✓	Wikipedia	Revision history, Comment
Rajagopal et al. (2022)	D	9.3K	✓	✓	✗	Wikipedia	Revision history, Comment
Laban et al. (2023)	D	145K*	✓	✓	✓	Wikipedia	/
Spangher et al. (2022)	D	40M	✗	✗	✓	News	Multilingual
Jiang et al. (2022)	D, S	13K	✓	✓	✓	ArXiv	/
Mita et al. (2024)	D, P	64*	✓	✓	✗	ACL anthology	Comment
Ruan et al. (2024)	Sec, P, S, SS	11.6K	✓	✓	✓	F1000Research	Review&response included

Table 1: Text-revision datasets. **Gran.** denotes revision granularity (S: sentence, SS: subsentence, P: paragraph, D: document, Sec: section); **#Pairs** is the number of revised sentence pairs; **H.** indicates human annotation; **I./A.** indicate edit intention/action labels; **Source** gives corpus origin; and **Features** lists additional information. $*$ marks revised document pairs, and ♦ denotes the most recent corpus in a series. See Appendix B for the complete table.

multiple languages (Faruqui et al., 2018). Sentence-level datasets are also widely used in educational contexts, where student essay revisions support fine-grained analysis of writing development, with annotations ranging from per-edit multi-label intentions to single labels per sentence pair (Zhang et al., 2017; Kashefi et al., 2022).

Higher-level Revision While sentence-level analysis captures fine-grained edits, it can obscure broader revision patterns in extensively edited documents (Hashemi and Schunn, 2014). Paragraph- and document-level datasets therefore enable more interpretable analysis of structural revision strategies, such as sentence reordering and content reorganization (Zhang and Litman, 2014; Spangher et al., 2022). Existing higher-level revision corpora span Wikipedia, scientific, and news domains, preserving paragraph or document structure with annotations for edit actions and, in some cases, edit intentions (Du et al., 2022b,a; Kim et al., 2022; Zhang et al., 2019; Rajagopal et al., 2022). More recent resources support multi-granularity or full-document analysis, distinguishing sentence-level intentions from document-level actions and enabling holistic modeling of complex revision workflows (Jiang et al., 2022; Mita et al., 2024).

3.2 Ground-truth Labels

Most revision datasets rely on human annotation, which provides reliable supervision for model training and evaluation. In contrast, some datasets forego manual edit intention annotation in favor of automated labeling of edit actions, enabling large-scale analysis at reduced annotation cost (Zhang et al., 2019; Spangher et al., 2022).

3.3 Feature-augmented Datasets

Revision History Many datasets preserve full revision histories, providing longitudinal signals for studying document evolution and iterative writing behavior (Zhang et al., 2017; Kashefi et al., 2022). Such histories enable analysis of how edit intentions and actions change over time and support modeling of iterative revision processes across domains (Du et al., 2022b,a; Kim et al., 2022).

Comments Revision comments provide explicit signals of editors' intentions and are therefore valuable supervision for modeling revision behavior and downstream tasks. Several datasets pair edits with corresponding comments and associated metadata, enabling discriminative modeling and richer analysis of editing practices (Zhang et al., 2019; Rajagopal et al., 2022; Mita et al., 2024).

Reviews and Responses Collaborative text production typically follows cycles of drafting, peer review, revision, and response, where response documents explicitly describe the changes made in reaction to reviewer feedback (Cheng et al., 2020; Kuznetsov et al., 2022). Recently introduced datasets model this complete collaborative cycle, providing annotations for both edit actions and edit intentions in the context of scholarly publishing and peer review (Ruan et al., 2024).

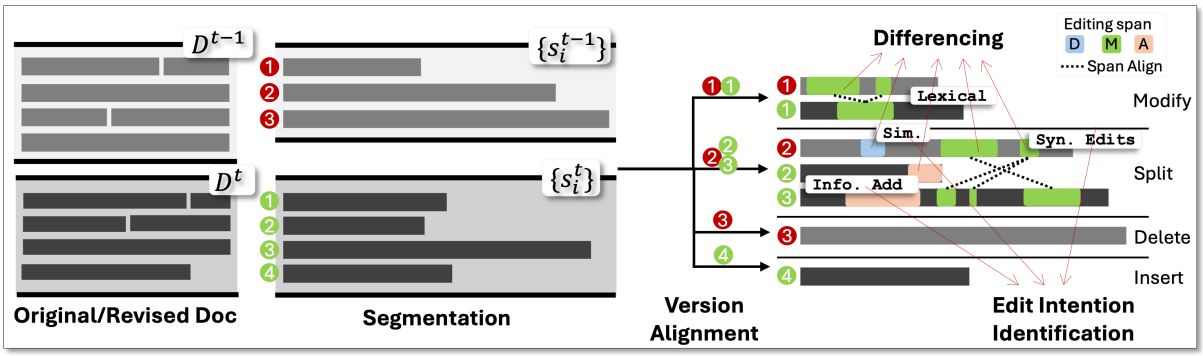

Figure 2: The workflow of revision dataset construction. See Appendix C for a more detailed example.

4 Revision Dataset Construction

Figure 2 illustrates the core steps of revision dataset construction: segmentation, version alignment, and differencing. We briefly summarize each step and highlight the key design choices that affect downstream edit intention analysis.

Segmentation Segmentation partitions documents into revision units at a chosen granularity $g \in \{P, S\}$, enabling alignment, differencing, and annotation. Finer-grained segmentation supports associating edits with coherent intentions and improves annotation reliability, while coarser segmentation can conflate heterogeneous revising goals. Existing approaches broadly fall into rule-based methods using punctuation heuristics (Lee and Webster, 2012) and learning-based models that treat boundary detection as a structured prediction problem (Reynar and Ratnaparkhi, 1997).

Version Alignment Version alignment establishes correspondences between texts at the same granularity across versions, $g^t - 1$ and g^t, identifying unchanged, revised, deleted, and added units. Most alignment methods rely on similarity estimation followed by a decision procedure. Similarity can be computed using surface-level lexical measures or semantic representations derived from language models, with the latter better handling paraphrasing and contextual reformulation (Zhang and Litman, 2014; Jiang et al., 2020). To improve scalability and consistency, prior work adopts hierarchical alignment (e.g., paragraph-to-sentence) or joint structured prediction to avoid exhaustive pairwise matching (Jiang et al., 2020; Ruan et al., 2024).

Differencing Differencing identifies the revised text spans ($\{p_k\}$) between aligned texts. Classical approaches apply string-matching-based algorithms such as longest common subsequence or edit distance to extract insertions, deletions, and substitutions (Myers, 1986). While efficient, these methods operate purely at the surface level and often conflate semantically distinct edit actions or fail to capture paraphrastic substitutions and text movement. Recent approaches reformulate differencing as semantic span alignment, producing finer-grained and more interpretable edit units that better support edit intention identification (Jiang et al., 2022). See Appendix F for an example.

Overall, revision corpus construction requires coordinated design choices across segmentation, alignment, and differencing, as errors or mismatches at early stages directly propagate to edit intention annotation and modeling.

5 Edit Intention Taxonomy

We now turn to how the motivations underlying revisions are conceptualized and organized. Edit intention categorization refers to the process of identifying the full range of intentions underlying textual revisions and organizing them into an EIT. Beyond providing a vocabulary of revision purposes, the design of an EIT also shapes how edit intention identification is formulated (e.g., label granularity and hierarchy), how annotations are collected, and how results are compared across datasets. In this section, we focus on the construction principles and design choices behind EITs, rather than enumerating specific edit intention labels.

Consider the following simplified EIT commonly used in revision studies. Revisions are first distinguished by whether they preserve or change meaning. *Surface-level edits* include FLUENCY (grammar and spelling correction) and FORMATTING. *Meaning-changing edits* include ELABORA-

TION (adding new information), VERIFICATION (correcting factual errors or adding citations), and SIMPLIFICATION (removing redundancy or reducing complexity). This structure reflects a widely adopted taxonomy lineage in academic writing and Wikipedia revision analysis.

5.1 Characterizing Edit Intention Taxonomies

EITs characterize the purposes behind text revisions. In this survey, we review prior EITs along four dimensions: taxonomy structure, category definitions and revision examples, application domains, and availability of data and resources (see the last one in Appendix D.2; Table 3 summarizes the surveyed taxonomies in Appendix D.1).

Taxonomy Structure Most EITs adopt a tree-based organization, ranging from flat category lists to hierarchical structures with up to three levels. Hierarchical EITs commonly distinguish high-level revision purposes from lower-level edit actions. Three-level taxonomies typically separate meaning-preserving from meaning-changing revisions (Faigley and Witte, 1981a; Zhang and Litman, 2015; Yang et al., 2016), sometimes further refining meaning changes based on discourse scope or edited objects. Two-level taxonomies retain similar high-level distinctions but often omit explicit modeling of edit operations (Jones, 2008; Daxenberger and Gurevych, 2012; Du et al., 2022b; Ruan et al., 2024). Flat taxonomies either collapse earlier hierarchies (Zhang and Litman, 2016; Rajagopal et al., 2022) or focus on specialized edit phenomena, such as edit actions (Spangher et al., 2022), clarity (Rathjens, 1985), or headline revisions (Guo et al., 2022). Despite structural differences, many EITs capture overlapping revision concepts.

Definitions and Revision Examples Clear definitions and representative examples are critical for EIT usability (Dilnutt, 2004). Omissions are common, particularly when category names appear self-explanatory (Anthonio et al., 2020), and some categories lack both definitions and examples, leading to ambiguity (e.g., *Style and Readability* (Jones, 2008)). Definition quality also varies, with some categories being overly broad (e.g., *Copy-Editing*) or inconsistent with their descriptions. Examples are typically provided only for leaf categories, and coverage is uneven across taxonomies (Faigley and Witte, 1981a; Jones, 2008; Liu and Ram, 2011; Faltings et al., 2021), complicating interpretation and cross-taxonomy comparison.

Application Domains EITs have been developed across diverse application domains. A substantial portion of the literature focuses on Wikipedia, examining collaborative behaviors, editor roles, and semantic edit intentions (Pfeil et al., 2006; Jones, 2008; Daxenberger and Gurevych, 2012; Yang et al., 2016, 2017). Writing-focused studies analyze revisions in student, technical, and academic texts to understand revision intent, clarity, and argumentative development (Faigley and Witte, 1981a; Rathjens, 1985; Zhang and Litman, 2015; Kashefi et al., 2022). Other work targets news articles and headlines (Spangher et al., 2022; Guo et al., 2022) or instructional texts such as wikiHow (Anthonio et al., 2020). More recent EITs aim to generalize across multiple domains (Du et al., 2022b).

Overall, existing EITs exhibit substantial conceptual overlap but differ in structure, definition quality, domain coverage, and resource support.

5.2 EIT Construction

We analyze relationships among EITs through their citation network (Figure 3; see Appendix D.3 for details), which reveals an EIT lineage: later EITs often build on earlier ones via merging, splitting, and renaming categories.

Constructing an EIT is non-trivial because EITs evolve with task objectives, application domains, and annotation requirements. For example, the EIT proposed for academic writing (Zhang and Litman, 2015) is later adapted to analyze semantic edit intentions in Wikipedia revisions (Yang et al., 2017). A follow-up study collapses all subcategories under *Surface Changes*, reflecting a shift toward augmentative revisions (Zhang and Litman, 2016).

A lineage-aware perspective shows that many EITs evolve via systematic transformations rather than purely bottom-up or top-down design (Lan et al., 2025). For instance, synthesizing prior EITs and iteratively refining them through batch annotation of samples enables the emergence of document-level categories not observable in sentence-level revision analysis (Laban et al., 2023). Common evolution patterns include: (1) *reuse-based pattern*; (2) *merge-based pattern*; (3) *refinement-based pattern*; and (4) *hierarchy adaptation*. These patterns highlight that bottom-up and top-down strategies are not mutually exclusive.

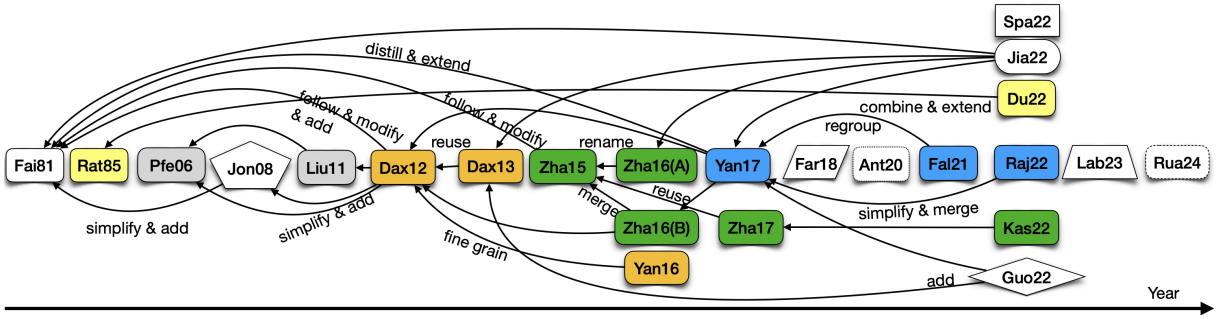

Figure 3: Lineage of existing EITs. Each block (labeled by paper alias; see Table 3) represents an EIT, with dashed arrows indicating inheritance relationships. Link labels denote how child EITs derive from parent EITs, and blocks sharing the same color indicate highly similar taxonomies (e.g., Pfe06 and Liu11)

6 Methods: Edit Intention Identification

For the edits in a revision, $\{e_k\} = R^{t,g}$, we may annotate one or more edit intentions from an EIT, $\{I_k\} \in EIT$, that most capture the editor's intention. Due to the semantic complexity and implicit nature of edit intentions, most existing studies rely on manual annotation to construct edit intention taxonomies, create revision datasets with gold labels, and analyze revision behavior, despite the substantial cost and limited scalability of this approach (Yang et al., 2017; Kashefi et al., 2022).

Manual Annotation Manual annotation is typically conducted via crowdsourcing platforms such as Amazon Mechanical Turk or through trained students, with annotator selection emphasizing language proficiency and, in some cases, domain or editing expertise (Zhang et al., 2017; Daxenberger and Gurevych, 2012). For example, Wikipedia-based corpora often require familiarity with platform conventions and policies (Yang et al., 2017). To ensure annotation consistency, studies commonly employ structured training procedures that include detailed guidelines, illustrative examples, live demonstrations, and iterative practice with feedback (Yang et al., 2017; Du et al., 2022b; Lan et al., 2025). Annotation reliability is assessed using inter-annotator agreement metrics, with Krippendorff's α and MASI frequently adopted in multi-label or overlapping intention settings (Daxenberger and Gurevych, 2012; Passonneau, 2006).

Automatic Annotation Automatic edit intention identification has been explored using both traditional machine learning and neural approaches. Early work formulates the task as binary or multi-label classification using feature-based models, often employing one-vs-rest strategies and designed features derived from textual differences, discourse cues, and metadata (Zhang and Litman, 2015; Daxenberger and Gurevych, 2013). While interpretable and computationally efficient, such models struggle with paraphrasing and context-dependent semantic changes. More recent approaches adopt neural and LLMs that encode original and revised text jointly and better capture semantic intent (Jiang et al., 2022). In particular, in-context learning with LLMs enables flexible intent generation without task-specific training, but introduces challenges related to computational cost, reproducibility, and output consistency (Ruan et al., 2024). Appendix E gives the annotation evaluation metrics.

Trade-offs Manual and automatic annotation methods exhibit complementary strengths and limitations. Manual annotation provides high-fidelity data grounded in human judgment and is essential for EIT development and further analysis, but is expensive and hard to scale (Yang et al., 2017). Automatic approaches enable large-scale analysis across domains and revision histories, but remain sensitive to domain shift, annotation guidelines, and model assumptions (Zhang and Litman, 2015). Thus, many studies adopt hybrid strategies in which manually annotated datasets serve as gold standards for training, evaluation, or calibration of automatic models, balancing annotation quality with scalability (Jiang et al., 2022; Ruan et al., 2024).

7 Application

Edit intention has been adopted as a unifying abstraction across a range of applications that seek to analyze, support, or summarize revision behavior beyond surface-level textual diffs.

Revision Behavior Analysis Edit intentions have been widely used to study collaboration dynamics in online collaboration systems. Prior work shows

that different intention distributions correlate with editor retention, participation patterns, and quality outcomes (Yang et al., 2017). For example, maintenance- and integration-oriented edits are associated with long-term editor survival, whereas simplification- or vandalism-related edits correlate with higher revert rates and early disengagement. Edit intentions have also been shown to contribute differently across an article's lifecycle (Yang et al., 2017), with content-expanding intentions playing a larger role in early stages and refinement-oriented intentions becoming more prominent as articles mature.

Beyond individual behavior, edit intentions provide a lens for analyzing sociocultural variation and division of labor in collaborative writing. Cross-lingual studies reveal differences in revision strategies across cultural contexts, indicating that revision behavior is shaped by social norms and task demands (Pfeil et al., 2006). Other work models editor roles by aggregating intention distributions over revision histories, uncovering interpretable mixtures of roles such as substantive contributors, copy editors, and vandal fighters, and demonstrating how different roles contribute to quality over time (Yang et al., 2016; Liu and Ram, 2011).

Writing Assistance Systems Edit intention has also been used as a control signal in writing assistance systems. Intention-aware editors model revision as a goal-directed process, enabling users to request specific types of changes (e.g., improving fluency versus adding content) and supporting iterative, interactive revision workflows. Empirical results show that conditioning generation on explicit edit intentions leads to more effective revisions than intent-agnostic baselines (Faltings et al., 2021). More approaches combine edit intention with free-form edit descriptions, using intention as a coarse semantic scaffold and natural language explanations to capture finer-grained rationale. This hybrid representation improves revision generation, reinforcing edit intention as a foundational abstraction for controllable and interpretable writing support (Rajagopal et al., 2022).

Document Edit Summarization Document edit summarization represents a higher-level application of edit intention analysis, aiming to abstract low-level diffs into concise summaries that explain what changed and why. Early work aggregates categorized edits to produce structured summaries of revision activity (Fong and Biuk-Aghai, 2010), while more recent formulations treat summarization as an intent-aware generation task (Ruan et al., 2024). By leveraging edit actions, intentions, and document structure, models generate natural-language summaries that resemble human revision descriptions. Results show that edit intention provides essential semantic guidance for summarization, though challenges remain in coverage, factual grounding, and discourse coherence for large documents.

Additional Uses Edit intention has also been applied in educational writing assessment, where revision types are used to characterize student writing development and support targeted feedback (Sommers, 1980; Faigley and Witte, 1981b; Zhang et al., 2017; Kashefi et al., 2022). In collaborative platforms, certain intentions (e.g., vandalism or revert-triggering edits) serve as signals for moderation, quality control, and editor modeling (Potthast et al., 2008; Halfaker et al., 2013). More recently, edit intention has been explored as a semantic abstraction for evaluating and interpreting model-generated revisions, enabling intent-aware comparison between human and automated editing behavior (Jiang et al., 2022; Ruan et al., 2024).

8 Future Research Directions

Evaluation and Benchmark Design Progress in edit intention research would benefit from evaluation frameworks that balance standardization with flexibility. Rather than enforcing a single universal taxonomy, future benchmarks could define shared core intention sets alongside domain-specific extensions, enabling both comparability and expressiveness. Designing evaluation protocols that account for overlapping or multi-intention edits remains an important research direction.

Evolving and Generalizable Edit Intention Taxonomies Future work may move toward edit intention taxonomies that are explicitly designed to evolve across domains, tasks, and revision granularities. Rather than treating taxonomies as fixed label inventories, research may emphasize extensible designs that support refinement, aggregation, and specialization while preserving core conceptual distinctions. Lineage-aware taxonomy construction, where new categories are introduced through principled reuse, merging, or refinement of existing ones, offers a path toward improving comparability and cumulative progress across studies.

Multi-intention and Cross-granularity Modeling Most existing work assigns a single intention to each edit at a fixed granularity, typically the sentence level. Future research should develop models that support multi-intention labeling and explicitly link sentence-level edits to paragraph- and document-level revising goals, such as restructuring, argument development, or narrative flow. Such representations would better reflect real-world revision behavior.

Process-level Modeling of Revision Dynamics Revisions are inherently sequential and iterative, yet are often modeled as independent edits in isolation. Capturing how edit intentions evolve over time, and how earlier revisions constrain or enable later ones, offers a promising direction for understanding revision strategies and improving downstream applications such as writing assistance and revision planning.

Future Applications Edit intention modeling has demonstrated value in revision behavior analysis, writing assistance, and document edit summarization, and continues to open new application opportunities. Promising directions include intent-aware software documentation maintenance, incremental summary updating, educational feedback and writing assessment, moderation and quality control in collaborative platforms, and evaluation of model-generated revisions. In high-stakes domains such as legal, medical, or policy drafting, edit intentions may further support auditability, justification, and explainability of revisions. While many of them remain exploratory, they underscore edit intention as a unifying abstraction for understanding, guiding, and evaluating text revision across domains. Edit intentions also offer a principled lens to study LLM evolution by systematically characterizing how factual knowledge, temporal awareness, and writing behavior change across model versions through analyzing intentional differences in responses to the same prompts over time.

9 Conclusion

This survey synthesizes text revision research through the unified lens of *edit intentions*, organizing prior work across datasets, corpus construction workflows, EIT design, identification methods, and downstream applications. We further introduce a structured categorization framework and highlight how EITs evolve over time and across domains through a citation network of existing EITs, helping capture EIT construction principles and design choices in the literature. By reviewing both manual and automatic annotation paradigms and mapping them to downstream applications, we show how edit intention serves as a practical semantic abstraction beyond surface diffs. Finally, we outline key challenges and dedicate future research to exploring this area in promising directions.

10 Acknowledgements

This work was supported in part by the U.S. National Science Foundation awards III-2107213, 2026513, and ITE-2333789.

11 Limitations

This survey has several limitations that reflect both the scope of the paper and the current state of edit intention research.

First, our coverage is necessarily bounded by the availability of publicly released datasets and published work. Although we made a best effort to include representative studies across domains and revision granularities, relevant datasets or methods may have been overlooked, particularly in non-English settings or in proprietary writing systems that do not release revision histories.

Second, the diversity of edit intention taxonomies poses an inherent challenge for comprehensive synthesis. Edit intentions are often domain-dependent, overlapping, and evolving, which limits direct comparability across studies. While we emphasize lineage-aware analysis and highlight common construction patterns, our categorization cannot fully reconcile differences in label definitions or annotation guidelines across datasets.

Finally, recent advances in large language models are evolving rapidly, and some findings summarized in this survey may be affected by future model releases or methodological shifts. Although we aim to capture stable research directions and enduring challenges, specific modeling approaches or empirical conclusions may require re-evaluation as the field progresses.

References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 5721–5729, Marseille, France. European Language Resources Association.

Irshad Bhat, Talita Anthonio, and Michael Roth. 2020. Towards modeling revision requirements in wikiHow instructions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8407–8414, Online. Association for Computational Linguistics.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Amit Bronner and Christof Monz. 2012. User edits classification using document revision histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366, Avignon, France. Association for Computational Linguistics.

Danielle Brown and Vinicio Sinta. 2016. Six things you didn't know about headline writing: Sensational form in viral news of traditional and digitally native news organizations. *Research Journal of ISOJ*, 6.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

Rod Dilnutt. 2004. The role of taxonomy in knowledge management. *The International Journal of Knowledge, Culture, and Change Management: Annual Review*, 3.

Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022a. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 96–108, Dublin, Ireland. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022b. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981a. Analyzing revision. *College Composition and Communication*, 32(4):400–414.

Lester Faigley and Stephen Witte. 1981b. Analyzing revision. *College Composition and Communication*, 32(4):400–414.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Peter Kin-Fong Fong and Robert P. Biuk-Aghai. 2010. What did they do? deriving high-level edit histories in wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, New York, NY, USA. Association for Computing Machinery.

Xingzhi Guo, Brian Kondracki, Nick Nikiforakis, and Steven Skiena. 2022. Verba volant, scripta volant: Understanding post-publication title changes in news outlets. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 588–598, New York, NY, USA. Association for Computing Machinery.

Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*.

Robert A Harris. 2017. *Writing with clarity and style*, 2 edition. Routledge, Second edition. | New York, NY : Routledge, [2018] | Previous edition published in 2003.

Homa B. Hashemi and Christian D. Schunn. 2014. A tool for summarizing students' changes across drafts. In *12th International Conference on Intelligent Tutoring Systems - Volume 8474*, ITS 2014, page 679–682, Berlin, Heidelberg. Springer-Verlag.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

John Jones. 2008. Patterns of revision in online writing: A study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. ArgRewrite v.2: an annotated argumentative revisions corpus. *Lang. Resour. Eval.*, 56(3):881–915.

Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Improving iterative text revision by learning where to edit from other revision tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9986–9999, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Fangping Lan, Abdullah Aljebreen, and Eduard Dragut. 2025. UniT: One document, many revisions, too many edit intention taxonomies. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23005–23024, Vienna, Austria. Association for Computational Linguistics.

Fangping Lan, Abdullah Aljebreen, and Eduard C. Dragut. 2026. Why they link: An intent taxonomy for including hyperlinks in social posts. *Preprint*, arXiv:2601.17601.

John Lee and Jonathan Webster. 2012. A corpus of textual revisions in second language writing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 248–252, Jeju Island, Korea. Association for Computational Linguistics.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2).

Bill MacCartney. 2009. Natural language inference.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2024. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 251–265, Mexico City, Mexico. Association for Computational Linguistics.

Eugene W Myers. 1986. AnO(ND) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Martin Potthast, Benno Stein, and Teresa Holfeld. 2008. Overview of the 1st international competition on wikipedia vandalism detection. In *Proceedings of the 1st International Workshop on Wikipedia Vandalism Detection*.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Dietrich Rathjens. 1985. The seven components of clarity in technical writing. *IEEE Transactions on Professional Communication*, PC-28(4):42–46.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington, DC, USA. Association for Computational Linguistics.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Re3: A holistic framework and dataset for modeling collaborative document revision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4635–4655, Bangkok, Thailand. Association for Computational Linguistics.

Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College Composition and Communication*, 31(4):378–388.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. NewsEdits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 403–408, Baltimore, Maryland. Association for Computational Linguistics.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):446–455.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 37–41, San Diego, California. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2014. Sentence-level rewriting detection. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 149–154, Baltimore, Maryland. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California. Association for Computational Linguistics.

Xuchao Zhang, Dheeraj Rajagopal, Michael Gamon, Sujay Kumar Jauhar, and ChangTien Lu. 2019. Modeling the relationship between user comments and edits in document revision. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5002–5011, Hong Kong, China. Association for Computational Linguistics.

A Complete Research Taxonomy

We present the complete taxonomy of research in text revision in the lens of edit intention in Figure 4.

B Complete Revision Datasets

We list the complete list of the corpora of text revision with edit intentions in Table 2.

C An Example of Workflow

We give a concrete example of revision dataset construction in a desired scenario, i.e., every step works well, in Figure 5. Given two paragraphs of the same Wikipedia article - Mariinsky Theater, segment them into sentences. The P^{t-1} and P^{t} are divided into 3 and 4 sentences, respectively. Then, we do a Cartesian product of sentences from the two versions, and apply the version alignment algorithm to find correspondence, resulting in 3 matches, i.e., 1 to 1, 2 to 2&3, and 3 to 4. The match of 2 to 2&3 indicates a finer edit action of split, while the other two matches are revised. Using 2 to 2 as a demonstration for differencing, we highlight the results of the differencing algorithm with green (Substitution), orange (Addition), and blue (Deletion) blocks. The dashed lines connect the editing spans that indicate one substantial edit. Then, we label those edits with an EIT.

D EIT Characteristics

D.1 Complete EITs Statistics

Table 3 summarizes the characteristics.

D.2 Application Domains and Availability

Data Sources and Resource Availability Prior work varies considerably in the availability of publicly released datasets, annotation guidelines, and source code. Some studies release substantial resources, including annotated corpora and preprocessing pipelines (Daxenberger and Gurevych, 2012; Yang et al., 2017; Du et al., 2022b), while others provide datasets without aligned taxonomies

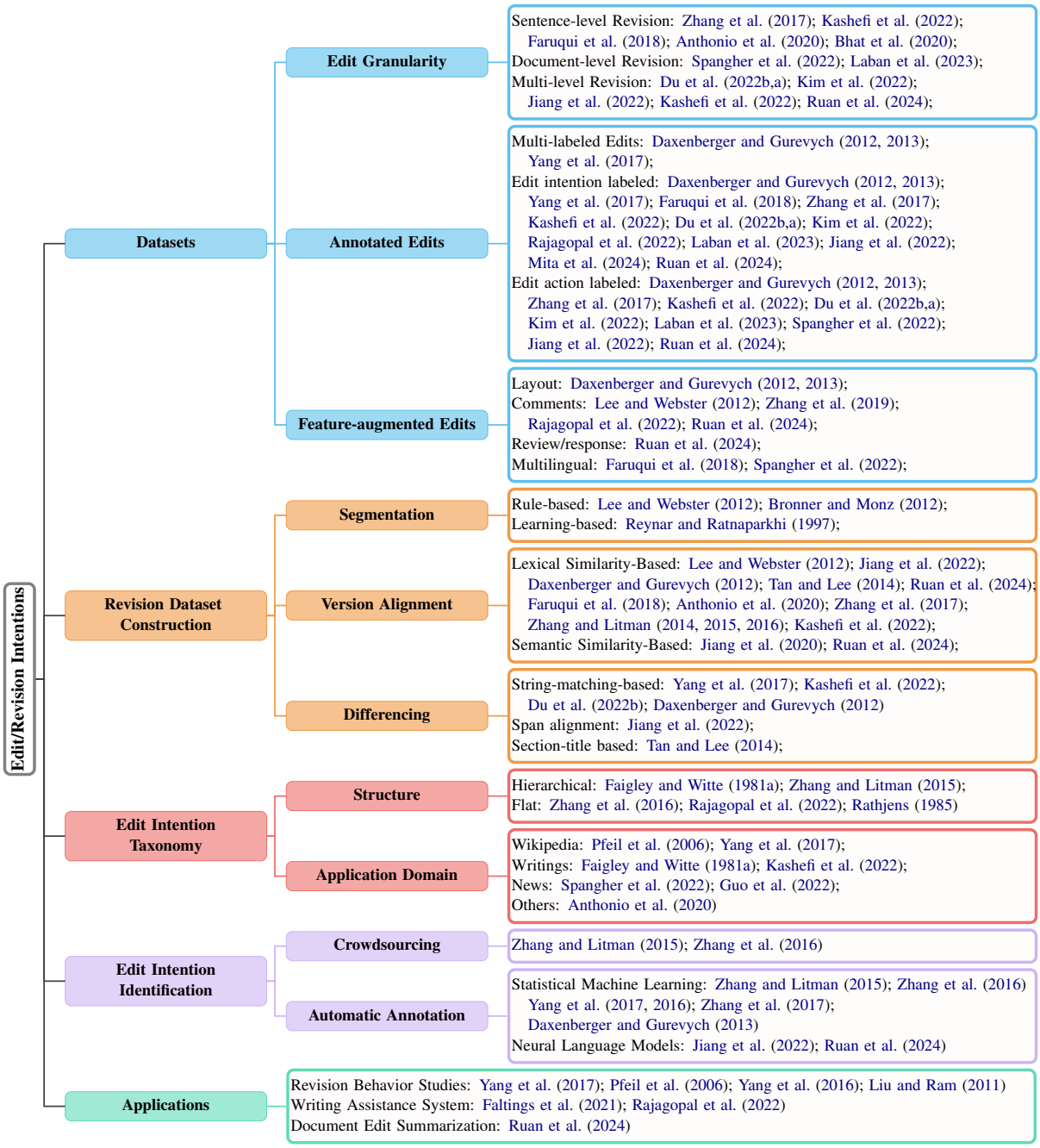

Figure 4: Taxonomy of edit intention-related research.

(Guo et al., 2022) or release trained models and crawlers (Anthonio et al., 2020). However, broken or outdated links remain common Daxenberger and Gurevych, 2013, limiting reproducibility and reuse.

D.3 Observations on EIT Lineage

Fai81[3] is the earliest EIT in our survey. Jon08 extends it by simplifying Macro/Micro-structure subcategories and adding Wikipedia-specific categories. Dax12 and Zha15 both inherit the high-level split of Surface Changes vs. Text-Base Changes from Fai81, while adapting subcategories to their respective settings: Dax12 emphasizes Wikipedia editing objects (e.g., file, template), and Zha15 targets structural elements of academic writing (e.g., claims/ideas, introductory materials).

Du22 reframes the taxonomy around meaning impact via Meaning-changed vs. Non-meaning-changed, defining Non-meaning-changed subcategories by building on Raj85 and Harris (2017).

[3]For compactness, we use aliases for the papers to reference their EITs. See Table 3 for full reference.

Paper	Gran.	# pairs	Human	Source	Intent	Action	Features
Daxenberger and Gurevych (2012)	S	2K	✓	Wikipedia	✓	✓	Including layout
Daxenberger and Gurevych (2013)	S	2K	✓	Wikipedia	✓	✓	Including layout
Yang et al. (2017)	S	5.7K	✓	Wikipedia	✓	✗	Multiple labels
Faruqui et al. (2018)	S	43M	✓	Wikipedia	✓	✗	Multilingual
Zhang et al. (2017)	S	180	✓	Student Essays	✓	✓	Revision history
Kashefi et al. (2022)	S	3K	✓	Student Essays	✓	✓	Revision history; Multiple labels
Du et al. (2022b)	S, P	31K*	✓	ArXiv Wikipedia WikiNews	✓	✓	Across domain; Revision history
Du et al. (2022a)	S, P	367K	✓	ArXiv Wikipedia WikiNews	✓	✓	Across domain; Revision history
Kim et al. (2022)	S, P	367K	✓	ArXiv Wikipedia WikiNews	✓	✓	Across domain; Revision history; Including edit tagging
Zhang et al. (2019)	Sec	786K	✗	Wikipedia	✗	✓	Revision history; Comment
Rajagopal et al. (2022)	D	9.3K	✓	Wikipedia	✓	✗	Revision History; Comment
Laban et al. (2023)	D	145K*	✓	Wikipedia	✓	✓	/
Spangher et al. (2022)	D	40M	✗	News	✗	✓	Multilingual
Jiang et al. (2022)	S, D	13K	✓	ArXiv	✓	✓	/
Mita et al. (2024)	D, P	64*	✓	ACL anthology	✓	✗	Comment
Ruan et al. (2024)	Sec, P, S, SS	11.6K	✓	F1000Research	✓	✓	Review&response included

Table 2: Text-revision datasets. **Gran.** denotes revision granularity: S (Sentence), SS (Subsentence), P (Paragraph), D (Document), and Sec (Section). **#Pairs** reports the number of revised sentence pairs. **Human** indicates the presence of human-annotated labels. **Source** specifies the origin of the corpus. **Intent** and **Action** indicate whether edit intentions and edit actions are annotated. **Features** lists additional information included in each dataset. Note: (1) $*$ indicates the number of revised document pairs; (2) datasets grouped without separating lines correspond to reused and expanded corpora in follow-up work.

In parallel, Liu11 compresses Pfe06's detailed scheme to improve feasibility for automatic classification, and further augments Pfe06 by introducing references-related categories alongside its link-related ones, motivated by their role in article quality.

Zha16(B) simplifies Zha15 by collapsing Surface subcategories into a single Surface category to focus on augmentative changes, and merges Rebuttal into Warrant based on the rarity of rebuttal edits observed in Zha15. Yan16 extends Dax12 by further fine-graining Reference, using edit intentions to identify editor roles, whereas Dax12 uses them to study collaborative behavior.

Yan17 distills and extends categories from Fai81, Dax12, and Zha16(B), organizing top-level categories by whether revisions are general or Wikipedia-specific, and introducing many new subcategories. Fal21 then reorganizes Yan17's leaf categories into Fluency, Content, and Other, distinguishing grammar/structure edits from meaning-changing edits. Raj22 further refines Yan17 by collapsing similar categories and merging Wikipedia-specific categories into Other.

Finally, Guo22 draws on Yan17 and Dax13 but tailors the taxonomy to news headline revisions, replacing unsuitable categories and adding journalism-driven ones (Blom and Hansen, 2015; Brown and Sinta, 2016). Ant20 is derived from manual inspection of a sampled set of sentence revision pairs.

E Annotation Evaluation Metrics

We summarize the evaluation metrics commonly used to assess the performance of edit intention identification models under these different formulations and annotation settings.

In multi-label classification, each instance (here, an edit) can belong to multiple labels. We consider example-based and label-based evaluation. The key difference between them is what gets averaged. *Example-based Metrics* are weighting each edit equally for multi-label classification. It can evaluate the overall usefulness on real instances.

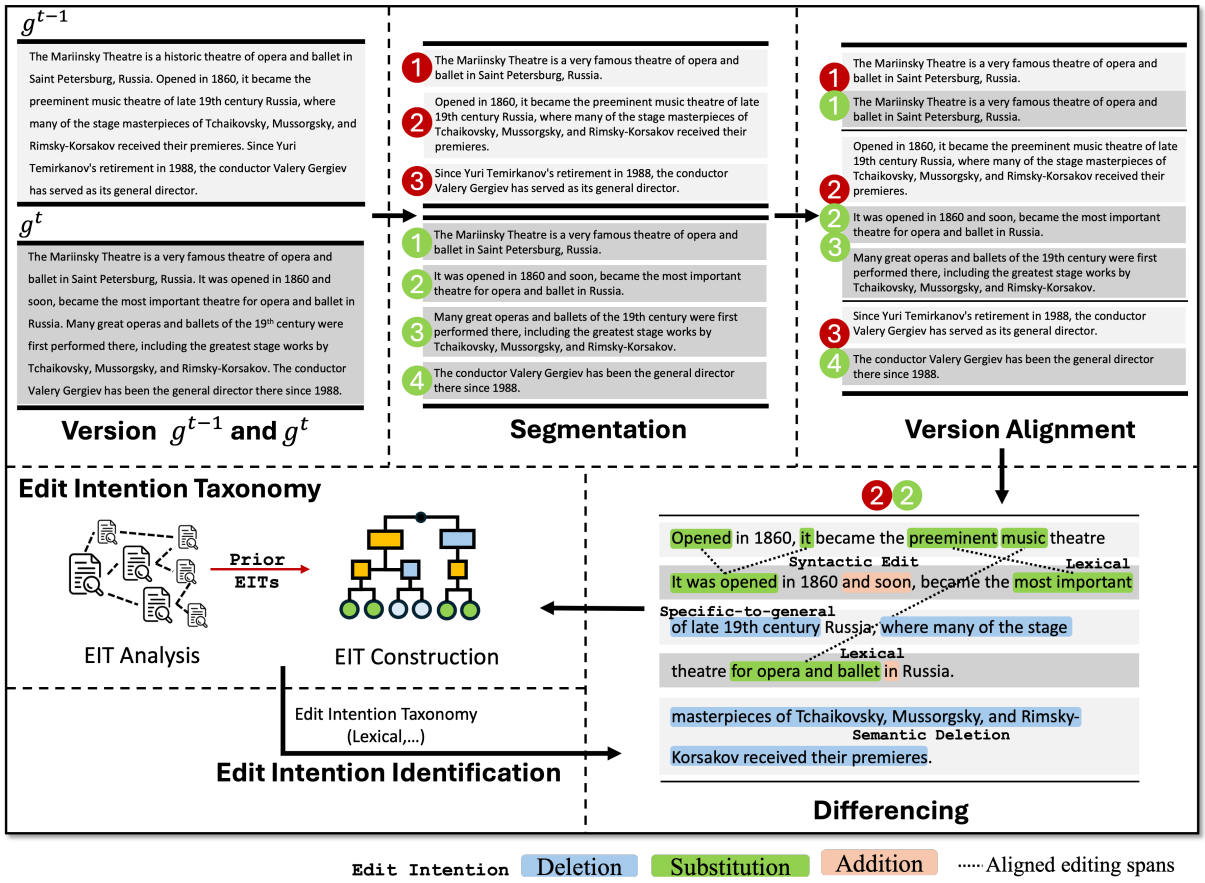

Figure 5: An example of the workflow of revision dataset construction.

We denote the set of relevant categories for each edit $e_i \in E$ as $y_i \in C$ and the set of predicted categories as $h(e_i)$. The accuracy of a multi-label classifier is defined as

$$ACC = \frac{1}{|E|} \sum_{i}^{|E|} \frac{|h(e_i) \cap y_i|}{h(e_i \cup y_i)} \quad (1)$$

which corresponds to the Jaccard similarity of $h(e_i)$ and y_i averaged over all edits. Example-based precision and recall are defined as

$$P = \frac{1}{|E|} \sum_{i}^{|E|} \frac{|h(e_i) \cap y_i|}{h(e_i)} \quad (2)$$

$$R = \frac{1}{|E|} \sum_{i}^{|E|} \frac{|h(e_i) \cap y_i|}{y_i} \quad (3)$$

$$F1 = \frac{1}{|E|} \sum_{i}^{|E|} \frac{2 \times |h(e_i) \cap y_i|}{|h(e_i)| + |y_i|} \quad (4)$$

We report macro- and micro-averaged F1 scores for label-based measures for *Label-based metrics*, which report robustness across categories and imbalance. It treats each category equally (macro) or each label occurrence equally (micro).

Exact Match evaluates whether the predicted labels are the same as the actual labels, calculated as

$$ACC_{exact} = \frac{1}{|E|} \sum_{i=1}^{|E|} I \quad (5)$$

where $I = 1$ if $h(e_i) = y_i$ and $I = 0$ otherwise.

F Demonstration: Differencing Algorithms

Figure 6 demonstrates the results of the string-matching-based LCS algorithm and semantic-aware span alignment. For example, deleting the phrase "went on to conduct" should ideally be decomposed into deleting "went on to" and substituting "conduct" with "performed", yet LCS-based differencing treats this as a single deletion. Similarly, semantically equivalent substitutions such as "conduct" ↔ "performed" or "simulations" ↔ "experiments" are not explicitly aligned. This behav-

alias	paper	domain	#le.	has_definition			has_examples		link?
				l1	l2	l3	extents	from	
Fai81	Faigley and Witte, 1981a	Writings	3	Yes	Yes	Partial	Partial	Inside	/
Zha15	Zhang and Litman, 2015	Writings	3	Yes	Yes	No	Full	Inside	/
Zha16(A)	Zhang et al., 2016	Writings	2	Yes	Yes		Full	Outside	Active
Zha17	Zhang et al., 2017	Writings	2	Yes	Yes	/	Full	Both	Active
Fal21	Faltings et al., 2021	Writings, Wikipedia	2	Yes	Yes	/	Partial	Inside	Broken
Du22	Du et al., 2022b	Writings, Wikipedia, News	2	Yes	Yes	/	Full	Both	Active
Jia22	Jiang et al., 2022	Writings	2	Yes	Yes	/	Full	Both	Active
Kas22	Kashefi et al., 2022	Writings	2	Yes	Yes	/	Full	Both	Active
Rua24	Ruan et al., 2024	Writings	2	Yes	Yes	/	Full	Both	Active
Rat85	Rathjens, 1985	Writings	1	Yes	/	/	Full	Inside	/
Zha16(B)	Zhang and Litman, 2016	Writings	1	Yes	/	/	Partial	Inside	/
Yan16	Yang et al., 2016	Wikipedia	3	Yes	Yes	No	No	/	/
Jon08	Jones, 2008	Wikipedia	2	Yes	No	/	No	/	/
Dax12	Daxenberger and Gurevych, 2012	Wikipedia	2	Yes	Yes	/	Full	Inside	Moved
Dax13	Daxenberger and Gurevych, 2013	Wikipedia	2	Yes	Yes	/	No	/	Broken
Yan17	Yang et al., 2017	Wikipedia	2	Yes	Yes	/	Full	Outside	Active
Lab23	Laban et al., 2023	Wikipedia	2	Yes	Yes	/	Full	Both	Active
Pfe06	Pfeil et al., 2006	Wikipedia	1	Yes	/	/	Full	Inside	/
Liu11	Liu and Ram, 2011	Wikipedia	1	Yes	/	/	No	/	/
Far18	Faruqui et al., 2018	Wikipedia	1	Yes	/	/	Full	Both	Active
Raj22	Rajagopal et al., 2022	Wikipedia	1	Yes	/	/	Full	Both	Active
Ant20	Anthonio et al., 2020	WikiHow	1	No	/	/	Full	Both	Active
Spa22	Spangher et al., 2022	News	1	Yes	/	/	Full	Both	Active
Guo22	Guo et al., 2022	News	1	Yes	/	/	Full	Inside	Active

Table 3: It shows the application **domain** of the taxonomy in a **paper**; **#le.** is the number of levels and whether it is flat (**#le.**=1) or hierarchical (**#le.**>1); whether the categories in level # (**l#**) has definitions (**has_definition**); whether 'Full', 'Partial' or 'No' leaf categories (**extents**) have concrete revision examples (**has_examples**) and they are **from** 'Inside' of the paper, 'Outside' of the paper through provided link (**link?**) or 'Both'. '/' denotes not applicable.

ior mixes deletions, substitutions, and additions at the string level, making subsequent edit alignment and edit intention identification difficult. Unlike LCS-based differencing, the span align algorithm captures semantic substitutions (e.g., "conduct" ↔ "performed" in Figure 3) and directly aligns corresponding edit spans.

G The Use of Large Language Model

To enhance readability, we employed OpenAI GPT-5.2 strictly as a language editing tool for grammar correction and stylistic refinement. Its use was limited to functions analogous to conventional proofreading and did not contribute to the conception, methodology, analysis, or scientific content of this work.

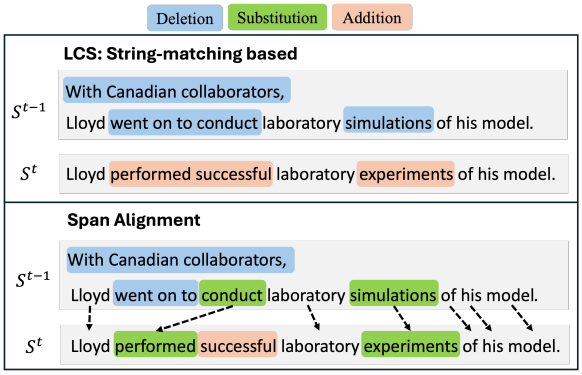

Figure 6: Comparison of string-matching–based differencing and semantic span alignment. While LCS-based differencing highlights surface-level changes, span alignment captures semantically meaningful edit operations such as deletions, substitutions, and additions.

paper	code?	data?	sources	link?
Fai81	No	No	(In)experienced student and expert revisions	No
Zha15	No	No	Student written papers	No
Zha16(A)	No	No	Student writings	Yes
Zha17	No	Ye	Student writings	Yes
Fal21	Yes	Ye	Wikipedia revision histories	Yes
Du22	Yes	Ye	Formally human-written text(Wikipedia, ArXiv, Wikinews)	Yes
Jia22	Yes	Ye	ArXiv Papers Revisions	Yes
Kas22	/	Ye	argumentative writing essays	Yes
Rua24	Yes	Ye	F1000RD dataset in Kuznetsov et al., 2022, ARR-22 subset of the NLPeer corpus in Dycke et al., 2023	Yes
Rat85	/	/	/	No
Zha16(B)	No	No	High school student written papers	No
Yan16	No	No	Three datasets from English edition of Wikipedian	No
Jon08	No	No	(Non-)featured Wikipedia articles revision histories	No
Dax12	No	Ye	(Non-)featured Wikipedia articles revision histories	Yes
Dax13	No	Ye	Wikipedia revision histories	Yes
Yan17	Yes	Ye	Wikipedia revision histories	Yes
Lab23	Yes	Ye	Wikipedia's revision history	Yes
Pfe06	No	No	Wikipedia page revision histories	No
Liu11	No	No	Different quality-level Wikipedia articles based on quality	No
Far18	No	Yes	revision histories	Yes
Raj22	No	Yes	Wikipedia revision histories	Yes
Ant20	Yes	Yes	wikiHow (non-)featured articles revision histories	Yes
Spa22	Yes	Yes	News revision histories	Yes
Guo22	No	Yes	News headlines from major US news agencies	Yes

Table 4: This table summarizes whether the literature provided their code and data. **code?** indicates if the code for preprocessing or regenerating data was released. **data?** shows if the data was provided along with their taxonomy. **sources** specifies the sources of the revisions their taxonomy and analysis are based on. **link?** indicates if the links to their code or data were provided. '/' denotes not applicable.