Characterizing Text Revisions to Better Support Collaborative

Tan Ping Ping
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Sarawk, Malaysia
pptan@unimas.my

Karin Verspoor
School of Computing Technologies
Royal Melbourne Institute of Technology
Melbourne, Australia
https://orcid.org/0000-0002-8661-1544

Timothy Miller
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
https://orcid.org/0000-0003-4908-6063

***Abstract*—Despite advancement in collaborative writing tools, the track changes capability in modern editors remains limited to highlighting syntactic changes, with authors still required to manually read through each of the revisions. We envision a collaborative authoring system where an author could accept all minor edits first and then focus on the substantial changes. To support this, we define the task of significant revision identification as the task of identifying the revisions between two versions of a text according to one of four categories, i.e. formal, meaning preserving, micro- and macro-structure. Micro-structure change corresponds to minor meaning change while macro-structure change corresponds to major meaning change. Our main contribution is to define a computational approach to this task, by framing the task as bi-directional entailment between the original and revised sentences. An existing recognition of textual entailment (RTE) system is applied to evaluate whether the revised texts entails. We evaluate the approach through a novel corpus consisting of multiple versions of drafts of academic papers written by multiple authors, which were annotated with the four revision types by both authors and non-authors of the papers. The proposed bi-directional textual entailment approach performs better than baseline edit distance approaches, which is similar to the current track changes capability built into most word processors.**

Keywords—text revision, revision identification, recognition of textual entailment

I. Introduction

Most current collaborative document and text editors such as Microsoft Word and Overleaf LaTeX provide the capability to track user edits or the history of changes made by multiple authors. Despite advances in these tools, the track changes capability remains limited to highlighting syntactic edits made by particular authors. Hence, authors are required to manually review all edits made by their co-authors, irrespective of the type or significance of the change. This may lead to major changes being overlooked as being trivial, or too much time spent reviewing trivial changes. If versioned document tools were able to automatically identify which edits co-authors should focus on, this could improve the revision efficiency in terms of attention and time by authors, especially when one draft by an author is passed to another author. Therefore, an approach to identify significant changes or revisions will hypothetically assist authors when revising in multi-author environment.

Faigley and Witte (1981) proposed a four-category taxonomy for analysing revisions: formal, meaning preserving, micro-structure and macro-structure changes. Formal changes (FC) are minor edits including copy editing operations such as corrections in spelling, tense, punctuation and format, while meaning preserving changes (MPC) are textual changes that do not alter the semantics of the text including re-phrasing. Micro-structure changes (MiSC) are meaning altering change that does not affect the original summary of the text, while macro-structure changes (MaSC) are major meaning changes that alter the original summary of the text.

Assume that we are given two versions of a text document (v_o, v_r), with each version written/edited by different authors, and a set of revised sentence pairs $(s_o, s_r)_k$, where s_o is the original sentence extracted from v_o and s_r is the revised sentence of s_o and k is the total revised sentences between (v_o, v_r). Here, we define significant revision identification (SigRevId) as the task of identifying the significance of the changes between the revised sentence pairs, according to one of the four categories proposed by Faigley and Witte (1981).

The revised sentences are characterised according to the meaning change. The definitions of micro- and macro-structure changes in Faigley and Witte (1981) are too vague to enable direct development of a computational model. Based on a review of the linguistics literature (Van Dijk, 1977a; Van Dijk, 1977b; Van Dijk, 1980) and introspective analysis of sentences revised in real-world texts, we propose the use of bi-directional textual entailment assessment for significant revision identification. The main contribution of this paper is a computational approach that can automatically detect significant revisions in a revised text and the construction of a corpus specifically for the task of significant revision identification.

II. Literature Review

Existing computational efforts to address revision in a multi-author environment typically focus on categorising various revision types such as factual edits – changes that alter the meaning – or fluency edits– changes to improve on the style and readability – (Bronner and Monz, 2012), although Daxenberger et al (2013) categorised edits to 21 predefined categories. The definitions can be related directly to the taxonomy for analysing revisions (Faigley and Witte, 1981). For instance, surface changes correspond to fluency edits while text-base changes correspond to factual edits but none of those categories look into minor and major meaning changes (i.e. micro- and macro-structure changes). Our earlier findings indicated that during the revision process in a multi-author environment, identification of micro- and macro-structure changes can support the authors better (Tan et al., 2015).

Most automated efforts that involve collaborative editors including a single author revision such as augmentative writing (Zhang and Litman, 2015) and revision analysis (Southavilay et al., 2013) not only track user edits, but also

2022 International Conference on Digital Transformation and Intelligence (ICDI) | 979-8-3503-9700-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDI57181.2022.10007395