Data & Knowledge Engineering xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Exploring alignment-classification methods in the context of professional writing assistance

Mai Duong[a,*], Minh-Quoc Nghiem[b], Ngan Luu-Thuy Nguyen[c]

[a] Ho Chi Minh City University of Information Technology, Vietnam
[b] Faculty of Information Technology, Ho Chi Minh City University of Science, Vietnam
[c] Faculty of Computer Science, Ho Chi Minh City University of Information Technology, Vietnam

ARTICLE INFO

Keywords:
Writing assistance
Alignment classification
Word alignment
Second language learning

ABSTRACT

Proofreading, the act of checking first-draft writings performed by native experts, is essential for professional writing by non-native speakers. Usually, proofreading experts return the corrected texts to the writer without reasons of correction, which makes it difficult for the writer to learn from their errors. The combination of word alignment and classification techniques can help us to analyze the original and corrected texts and use them for language learning. In this study, we explore different alignment-classification methods for this task. Our experimental results show that the best method achieved 71.8% in accuracy. We also propose a new error taxonomy for tagging learner corpora, and present our alignment-classification results on the corpus tagged with this new tagset.

1. Introduction

Composing a professional writing in English is uneasy for non-native speakers. In the field of second language learning, many works have been published on methods for helping non-native speakers to revise their first drafts [1]. In the field of computer-assisted language learning, computer systems have been built to assist non-native novice writers in writing [2–4]. There are many researches on writing-assisted techniques, but most of them are about grammatical error correction [5]. Nevertheless, the difficulty remains, and having the writing checked by a native speaker, or proofreading, has become an essential step in publishing a professional writing. While full automation of proofreading is a real demand of users, it is a challenging task. Our study focuses on an easier but still challenging goal, that is how to automatically analyze the proofread texts to make them useful for language learning. This could also be considered the first step toward automatic proofreading.

In general, proofreading is already established and many companies can do this. However, it usually provides only parallel texts and does not provide reasons for editing. For non-native speakers, providing such information is very useful. Firstly, in the process of decision making, it is easier for writers to accept or reject the correction. Secondly, they can learn from the correction to improve their writing skill by understanding the reasons behind it.

In this paper, we explores methods to solve the task that we call word-alignment classification in the context of professional writing assistance, which will be referred to as "Wa2" in this paper. This task can be decomposed into two subtasks: word alignment and alignment classification. The word-alignment subtask in Wa2 is different from the other contexts of word alignment such as for machine translation. The input of the Wa2 task is a pair of documents: an original document written by a non-native English speaker and a proofread version of that document, which has been revised by a native English expert. Output are alignments of the

[*] Corresponding author.
E-mail addresses: dt.phuong.mai@gmail.com (M. Duong), nqminh@fit.hcmus.edu.vn (M.-Q. Nghiem), ngannlt@uit.edu.vn (N.L.-T. Nguyen).

http://dx.doi.org/10.1016/j.datak.2017.08.005
Received 23 January 2017; Received in revised form 17 July 2017; Accepted 20 August 2017

correspondences between the two texts that can be used to give informative and useful feedback for the writer. Moreover, each alignment is accompanied by its alignment type, i.e. the reason why proofreader made the correction, such as “paraphrase”, “preposition”, or “determiner”. Although several research works have been carried out for similar tasks [6,7], the Wa2 task has not yet achieved enough attention and the performances so far are still far from the requirement for practical use.

The remainder of this paper is organized as follows. Section 2 compares the Wa2 task with general word alignment task and error analysis for machine translation. Section 3 introduces our new error tagset for tagging learner corpora, which is an improved version of the error tagset proposed by Nguyen and Miyao [8]. This section also provides a brief overview of the learner corpora we used in our experiments. Section 4 briefly describes the methods we used for automatic alignment and alignment classification. In Section 5, we presents our experimental results and their implications to our Wa2 task. In order to give more insights into this task, further analyses are presented in Section 6. Section 7 concludes the paper and points to avenues for future work.

2. Related works

2.1. Wa2 and the general word alignment task

Word alignment is the task of linking corresponding words that express the same meaning in a pair of texts. It has been recognized as an important component in different NLP systems such as statistical machine translation (SMT) [9,10], paraphrasing [11,12], and natural language inference (NLI) [13,14]. The word alignment problem varies significantly according to its application. MacCartney et al. [13] contrasted the characteristics of word alignment for SMT with that for NLI.

There are four characteristics why Wa2 is different from other alignment tasks. These are:

- Wa2 is a monolingual alignment task, which means the source and target languages are the same, which is English in this paper.
- Original texts are written by non-native speakers, and may contain bad writings including grammatical errors and influent uses of language.
- Source and target texts are almost semantically equivalence. We use the word *almost* because there are cases when proofreaders input additional information to make the original text clearer.
- Lengths of the source and target texts are similar. For NLI, the source text usually is longer than the target one.

2.2. Wa2 and error analysis for machine translation

The problem formulation of Wa2 is most similar to the task of *error analysis for machine translation* [15–17] (MT error analysis), thus we want to give space to discuss about the two tasks.

Both of these tasks target at classifying word alignments into error types that are meaningful to users. In MT error analysis, the machine translation output is aligned with the reference text; classifying the erroneous alignments helps to figure out how to improve the machine translation system. Thanks to this practical use, the MT error analysis task has been studied extensively in the natural language processing literature [17–19]. In a similar way, classifying the alignments between the original and proofreading texts plays an important role in helping non-native speakers to understand their weaknesses in writing in order to improve their writing skills. This is a real need of advanced English learners. However, to the extent of our knowledge, this paper is the first attempt in putting the alignment classification in the context of writing assistance.

The main difference between the Wa2 and MT error analysis tasks comes from the error typology for alignment classification. Different error typologies were proposed for the MT error analysis [15,20,18] and for the Wa2 task we find a typology in Nguyen and Miyao [8]. We find that the error types are commonly based on four general cases of alignments: exact matching, stem matching, synonyms, and paraphrases.

3. New error tagset

3.1. The SWA corpus

To the best of our knowledge, there is only one available corpus that target professional writing assistance, that is the Scientific Writing Assistance (SWA) corpus [8]. It is a collection of scientific papers written by non-native speakers. For every sentence in the corpus, there are two versions: an original and a proofread version, which was corrected by native English experts. Alignments between words and phrases in the two version were manually annotated and classified. This is the first work to exploit the SWA data for training and evaluation of word alignment for writing assistance.

The SWA corpus consists of 3485 pairs of sentences in 18 scientific documents. Out of which, there are 2516 pairs of sentences contain errors which are called inarticulations by the authors. On average, each sentence contains 30 words. In the SWA corpus, every alignment link was manually assigned one of the thirteen alignment types. There are three broad types: Inarticulation Mono-alignment, Inarticulation Bi-alignment, and Preserved. These broad types are further divided into thirteen sub-types (see Fig. 1). The most dominant type is the Preserved alignment (91.8%). Except for the Preserved type, 87.4% of the alignments are “paraphrase”, “Inarticulation mono-alignment grammar”, and “Inarticulation bi-alignment grammar”. Fig. 6 shows a pair of text with annotations in SWA corpus.

The SWA corpus has annotations to capture all types of articulation correction between original and proofread version. Each of the sentence pairs was independently annotated by two people with linguistics background, following carefully designed annotation