Example Sentence Selection for Feedback on Preposition Usage

John Lee

Department of Linguistics and Translation, City University of Hong Kong
Email: jsylee@cityu.edu.hk

***Abstract*—While many writing assistance systems can automatically correct grammatical errors, most do not provide any explanation about their suggested corrections. This paper proposes an algorithm that selects example sentences to serve as feedback on preposition usage correction. This algorithm exploits the argument/adjunct distinction to select the most relevant example sentences. Evaluation shows that the use of argumenthood information improves the quality of the selected sentences.**

***Keywords*–computer-assisted language learning; example sentence selection; grammatical error correction feedback; preposition.**

I. Introduction

A Grammatical Error Correction (GEC) system detects and corrects grammatical errors in a learner text [1]. Given the input sentence "... go shopping **to* a store" [2], for example, the system may flag the preposition "to" and propose to replace it with "at". Studies in second language acquisition have shown that feedback from language teachers can be beneficial to foreign language pedagogy [3]. Most GEC systems, however, do not provide any feedback or explanation to complement their proposed corrections.

Research on automatic feedback generation has mostly focused on *explanatory feedback*, imitating the kind of comments traditionally composed by teachers (Section II-A). For the input sentence above, the system may elaborate on its correction with a feedback message such as "To mean traveling to a place in order to take part in an activity, *go* takes *at*, *in* or *on* depending on the activity ..." [2]. To date, most algorithms generate explanatory feedback by compiling comments from experts for different error types, and assigning them to unseen learner errors [2], [4]. Significant manual effort is required to cover the large variety of learner errors.

In contrast, *example-based feedback*, which presents example sentences to illustrate correct usage for the user, requires no manual composition (Section II-B). This approach can provide wider coverage, because it can address virtually any kind of usage issues, however idiosyncratic, as long as relevant examples can be found in the corpus. In addition, it supports data-driven learning by encouraging users to discover language patterns through observation of real-world example sentences, rather than through direct comments from the system or teacher [5]. Some existing systems can already provide example-based feedback by searching for similar sentences on the web [6] or in text corpora [7]. However, we are not aware of any reported evaluation on the quality or effectiveness of the retrieved example sentences.

This paper proposes an algorithm for generating example-based feedback aimed at preposition usage errors, a common error type for students of English as a Foreign Language [8]. This algorithm exploits the argument/adjunct distinction in prepositional phrases to help determine the most relevant example sentences. Evaluation shows that argumenthood information can help select higher-quality example sentences.

The rest of the paper is organized as follows. The next section presents previous work in feedback generation and argumenthood prediction. Section III presents our approach. Section IV describes our evaluation dataset and Section V discusses the results. Finally, Section VI concludes.

II. Previous Work

The feedback generated by existing writing assistance systems tends to fall into one of two types, *explanatory feedback* (Section II-A) and *example-based feedback* (Section II-B). After summarizing current approaches for generating these two types of feedback, we describe the argument/adjunct distinction in prepositional phrases (Section II-C), which will be exploited by our algorithm.

A. Generation of Explanatory Feedback

Among GEC systems that provide explanatory feedback, most rely on experts to manually compose the feedback or explanation for each error category. In the more coarse-grained approach, a "comment bank" [9] provides generic comments for broad error types such as "wrong preposition" or "wrong article". After correcting an error in the input text, the system delivers the comments associated with that error type to the user. While these hand-crafted comments can be comprehensive, they also tend to be generic and do not directly address the specific word usage in the input sentence.

In the more fine-grained approach, the feedback is associated not to broad error types, but rather to parse tree patterns [4], [10] or error case frames [2], which facilitate more in-context feedback. Case frames for preposition errors, for example, can be specific to the particular subject, verb, direct object, preposition and prepositional object in the sentence [2]. This approach still requires a significant amount of manual annotation, since error coverage is proportional to the number of frames for which comments are available.

B. Generation of Example-based Feedback

A GEC system can also offer example sentences as feedback to illustrate correct usage, either as an alternative or a supplement to explanatory feedback. This approach requires no hand-crafted messages. Further, given the size of contemporary text corpora, it can potentially cover a wider range of errors with corpus examples that more closely address the user's errors. The *ESL Assistant*, for example, automatically performs web search to retrieve sentences containing the original and corrected phrases [6]. A CALL tool for prepositions offers a review function, where users can request fill-in-the-blank

 ISBN: 978-1-61208-738-2

items that are similar to those with which they previously experienced difficulties [7]. A sentence is considered "similar" if it contains the same preposition, prepositional object and lexical head, which can be identified in a parse tree such as the one in Figure 1.

Example sentence selection has primarily been investigated in the context of dictionary entries [11], [12], test item generation [13] and general language learning [14], typically using heuristics-based approaches. The kinds of example sentences required in these contexts share many similar characteristics with those for example-based feedback, such as well-formedness, simplicity of vocabulary, and ease of understanding. However, they target a larger variety of sentences, in order to provide a comprehensive portrait of the various aspects of the word's usage and collocational behavior. In contrast, example-based feedback aims at a narrower set of sentences that can precisely address the user's specific problem. Sentence selection for this purpose, therefore, often requires more syntactic and semantic analysis to determine the nature of the usage error. For preposition usage, this entails analyzing whether the preposition is used as an argument or adjunct.

C. Argumenthood

Arguments and adjuncts are linguistic concepts that have been intensively studied. In principle, "arguments depend on their lexical heads because they form an integral part of the phrase. Adjuncts do not." [15] An argument prepositional phrase (PP) is thus more closely related to the lexical head than an adjunct PP. For example, the phrase "to our topic" in sentence (1) in Table I is an argument PP, namely an argument of the lexical head "relevant". In contrast, "to some extent" in sentence (2) is an adjunct, serving as an adverbial to the lexical head "relevant". Argumenthood information has been shown to benefit a variety of natural language processing tasks, including PP attachment [15] and semantic role labeling [16]. It has not, however, been applied to sentence selection for feedback on grammatical errors.

There are a number of language resources that encode argument constructions, such as the verb subcategorization forms in VerbNet [17] and the grammar patterns in COBUILD [18]. Past work has attempted to distinguish between PP arguments and adjuncts with these resources, logical forms and formal grammars [19], as well as statistical models based on word embeddings and a variety of linguistic features [20].

III. Approach

Assuming that a grammatical error correction (GEC) system has corrected a preposition error in the input sentence, our task is to select the best example sentences from a corpus to explain and clarify the preposition usage. Similar to [21], we characterize preposition usage with three features: the corrected **preposition** (p'); the **prepositional object** (obj); and the **lexical head** (h). These features can be identified from a dependency parse tree. Figure 1 shows the tree for sentence (1) in Table I. Based on the dependency relations, we extract $p' =$"to", $obj =$"topic", and $h =$"relevant".

After defining the objective of the feedback (Section III-A), we discuss the types of example sentences to be considered (Section III-B), and the algorithms to be used for predicting the most suitable type (Section III-C).

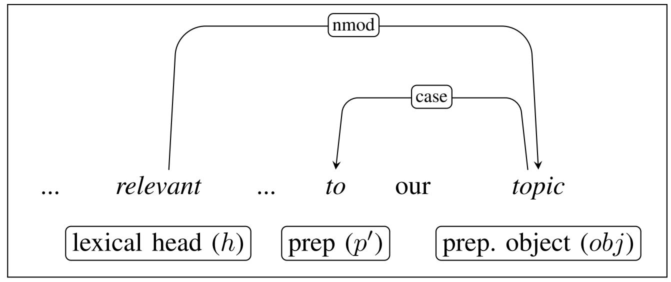

Figure 1. Extraction of the preposition, prepositional object and lexical head from a dependency parse tree, as derived by the Stanford parser [22].

A. Feedback Objective

When a GEC system corrects a preposition p to p', the user may not be able to discern the underlying reason:

> Is p' better than p because of (a) the lexical head, regardless of the choice of prepositional object? or because of (b) the prepositional object, regardless of the choice of lexical head?

For sentence (1), the answer is (a) because its PP is an argument. The ideal example sentence should make the point that the preposition "to" is required by the word "relevant", even when using other prepositional objects. In contrast, for sentence (2), the answer is (b) because its PP is an adjunct. A useful example should emphasize that "to ... extent" is the expected expression, even for lexical heads other than "relevant".

B. Types of example sentences

Table I lists some possible types of example sentences to provide feedback. An *Identical Example* is a sentence with the same p', obj and h as the input sentence. Sentences (3) and (4), for example, would serve as Identical Examples for sentences (1) and (2), respectively. An Identical Example seems useful in reinforcing the correction, since its content most closely matches the input sentence. However, by merely repeating the correction with the same h and obj, it gives no new insight and does not resolve the ambiguity noted in Section III-A: the user still would not be able to tell whether h or obj triggered the correction. We will therefore not give further considerations to Identical Examples. Instead, we focus on two kinds of example sentences:

1) Argument Example: We use the term "Argument Example" to refer to a sentence with the same p' and h as the input sentence. In Table I, sentence (5) serves as an Argument Example for (1) and (2). It gives useful feedback for sentence (1), where the *to*-PP is an argument. By using a different obj ("her"), it makes clear that the use of "to" is linked to the lexical head "relevant". In other words, it highlights the fact that *to* is an argument PP for the adjective "relevant".

In contrast, this example is less optimal for sentence (2). In reusing the lexical head "relevant", it fails to make the point that the expected expression is "to ... extent", and may even give the false impression that "*in some extent" could be appropriate with other lexical heads.

TABLE I. TYPES OF EXAMPLE SENTENCES AS FEEDBACK ON PREPOSITION USAGE.

Type	Sentence	Lexical head (h)	Prep (p')	Prep. object (obj)	Remarks
(Corrected) input	(1) This book is **relevant** ~~with~~ *to* our **topic**.	relevant	to	topic	*to*-PP is an argument
	(2) This book is **relevant** ~~in~~ *to* some **extent**.	relevant	to	extent	*to*-PP is an adjunct
Identical Example	(3) This movie is **relevant** *to* the current **topic**	relevant	to	topic	Same h, p' and obj as (1)
	(4) This movie is **relevant** *to* a large **extent**.	relevant	to	extent	Same h, p' and obj as (2)
Argument Example	(5) This movie is **relevant** *to* **her**.	relevant	to	her	Same h and p' as (1), (2)
Adjunct Example	(6) This movie is **outdated** *to* a large **extent**.	outdated	to	extent	Same p' and obj as (2)

2) Adjunct Example: We use the term "Adjunct Example" to refer to a sentence with the same p' and obj as the input sentence. In Table I, sentence (6) serves as an Adjunct Example for (2). It gives useful feedback for sentence (2), where the *to*-PP is an adjunct. By using a different lexical head, "outdated", it clarifies that the choice of "to" as preposition is not tied to "relevant"; rather, it is required for the PP "to ... extent", even when under another lexical head.

C. Algorithm for Example Sentence Selection

To evaluate the effect of the argument/adjunct distinction on the quality of example-based feedback, we implemented the following algorithms for selecting example sentences. Given h, p' and obj, the algorithm is to determine whether Argument Examples or Adjunct Examples are more suitable as example sentences.

1) Majority Baseline: Ignoring the argument/adjunct distinction, this baseline always chooses the majority type in the evaluation dataset (Section IV).

2) COBUILD Grammar Patterns Baseline: These grammar patterns consist of phrases or clauses that are used with a verb [18], adjective or noun [23]. One pattern for the adjective *relevant*, for example, is the PP "`to n`". This baseline opts for Argument Examples as feedback if p' is listed among the patterns for h. Otherwise, it chooses Adjunct Examples.

3) Association Score Difference: Recent research suggested that the phenomenon of argumenthood is not binary, but gradient [20]. The grammar patterns define a boundary between argument and adjunct, but this boundary may not be the one at which Argument Examples become more useful than Adjunct Examples, or vice versa. This algorithm uses the logDice score [24], which measures word collocation strength based on the Dice Coefficient, as a proxy to learn this boundary from user data.

Let $logDice(h, p)$ represent the logDice score for the lexical head and the preposition, and let $logDice(obj, p)$ represent the score for the prepositional object and the preposition. We compute the difference between these scores, i.e., $\Delta logDice = logDice(h, p) - logDice(obj, p)$. We choose Argument Examples if $\Delta logDice > \theta$ and Adjunct Examples otherwise, with θ to be optimized on user data.

While there are many other approaches for predicting argumenthood (Section II-C), most concentrate on verbs as lexical heads and would have required non-trivial extension for nouns and adjectives. Since our goal is not to investigate the state-of-the-art in argumenthood prediction, we chose to use the logDice score for its simplicity and availability via Sketch Engine.

IV. EVALUATION DATASET

We extracted sentences containing preposition usage errors from Release 3.3 of the National University of Singapore (NUS) Corpus of Learner English (NUCLE) [25]. To construct an evaluation dataset that is balanced in terms of the part-of-speech (POS) of the lexical head and the argument/adjunct distinction, we randomly selected 24 sentences within the following constraints:

- **Lexical head POS**: 10 sentences have verbs as lexical head, 10 have nouns, and 4 have adjectives;
- **Argument vs. Adjunct**: Among sentences with lexical heads of each POS, one half have argument PPs and the other half have adjunct PPs, according to the COBUILD Grammar Patterns (Section III-C).

Since our research focus is on example sentence selection rather than grammatical error correction (GEC), we used the gold preposition in NUCLE to ensure that GEC accuracy would not be a confounding variable. A total of 8 prepositions ("at", "for", "from", "in", "of", "on", "through", "to", and "within") are represented in the dataset.

We retrieved example sentences in Sketch Engine with the collocation (h, p) to serve as Argument Examples, and sentences with the collocation (obj, p) to serve as Adjunct Examples. For each of the 24 input sentences, we collected the first three sentences returned by Good Dictionary EXamples (GDEX) [11] to create an Argument Example Set and an Adjunct Example Set. Table II shows an example item in the evaluation dataset.

For each item, we asked five human raters to decide whether the Argument or Adjunct Example Set was more useful. All five raters were advanced non-native speakers of English with a postgraduate degree in linguistics. The argument/adjunct distinction of the sets was not disclosed to the raters.

TABLE II. EXAMPLE ITEM IN EVALUATION DATASET

(Corrected) Input	The only way to satisfy the increasing demands ~~of~~ *for* space is by achieving a better usage ...
Adjunct Example Set	Canisters aren't the best option *for* big **spaces**. It's the perfect accent lamp *for* a small **space**. So that is a very practical use *for* **space**.
Argument Example Set	The **demand** *for* processed food items have increased ... They show no sign of scaling back their **demands** *for* human rights. Thus, **demand** *for* base metals will remain very strong.

V. EVALUATION RESULTS

We applied each algorithm in Section III-C to select either the Argument or Adjunct Example Set for each item in the evaluation dataset. For the Association Score Difference algorithm (Section III-C), we obtained the logDice scores in the English Web 2015 corpus on Sketch Engine, and tuned the value of θ using leave-one-out cross-validation in the evaluation dataset (Section IV).

TABLE III. ACCURACY IN SELECTING EXAMPLE SENTENCES FOR FEEDBACK ON PREPOSITION USAGE

Algorithm	Accuracy
COBUILD Grammar Patterns baseline	67.50%
Majority baseline	72.50%
Association Score Difference	**76.67%**

Table III shows the algorithms' accuracy in selecting the example set preferred by the rater. The Majority baseline achieved an accuracy of 72.50% by always choosing the Argument Example Set. Recall that only 50% of the items in the dataset have p' listed in the COBUILD Grammar Patterns as an argument marker (Section IV). This suggests a general preference among raters for example sentences illustrating argument usage. This preference holds regardless of the POS of the lexical head. For the rater with the strongest such preference, the Argument Example Set was deemed more useful in 8 out of the 12 adjunct items.

The COBUILD Grammar Patterns yielded an accuracy of 67.50%, below the Majority baseline. When it chose Argument Examples, the raters almost always agreed. Most errors occurred when it opted for Adjunct Examples, when the raters often preferred the Argument ones. This may reflect incomplete coverage in the grammar patterns, or could be the result of the gradient effect of the argumenthood phenomenon [20].

The Association Score Difference algorithm produced the best performance, at 76.67% accuracy. The improvement over the Majority baseline, at $p < 0.074$ by McNemar's Test, approaches statistical significance. The logDice score turned out to be a close proxy of the COBUILD Grammar Patterns, generally giving higher scores to (h, p') collocations where p' is listed in the patterns. Reflecting the raters' general preference for Argument Examples, the threshold θ was tuned to a relatively large negative value. This means that the algorithm selected Adjunct Example Sets only when the logDice score for (obj, p') enjoyed a large margin over the score for (h, p'). Experimental results thus show that the Dice Coefficient was effective in making more judicious selections for Adjunct Example Sets to cater to user preference on the argument-adjunct gradient for example sentences for preposition usage.

VI. CONCLUSION

We have presented a novel approach to select example sentences as feedback on preposition usage. This algorithm exploits the argument/adjunct distinction to determine the most useful examples. Evaluation shows that it can learn user preference on the argument-adjunct gradient to improve the quality of the selected example sentences.

ACKNOWLEDGMENT

The work reported in this paper is partially funded by an Applied Research Grant (#9667151) from City University of Hong Kong.

REFERENCES

[1] A. Rozovskaya and D. Roth, "Grammatical Error Correction: Machine Translation and Classifiers," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

[2] R. Nagata, M. Vilenius, and E. Whittaker, "Correcting Preposition Errors in Learner English Using Error Case Frames and Feedback Messages," in Proc. 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014, pp. 754–764.

[3] J. Bitchener and S. Young, "The Effect of Different Types of Feedback on ESL Student Writing," Journal of Second Language Writing, vol. 14, no. 3, 2017, pp. 191–205.

[4] K. F. McCoy, C. A. Pennington, and L. Z. Suri, "English Error Correction: A Syntactic User Model based on Principled "Mal-rule" Scoring," in Proc. 5th International Conference on User Modeling, 1996.

[5] T. Johns, "Should You be Persuaded — Two Samples of Data-driven Learning Materials," in Classroom Concordancing, T. Johns and P. King, Eds. ELR Journal (4), 1991, pp. 1–16.

[6] C. Leacock, M. Gamon, and C. Brockett, "User Input and Interactions on Microsoft Research ESL Assistant," in Proc. NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications, 2009, pp. 73–81.

[7] J. Lee and M. Luo, "Personalized Exercises for Preposition Learning," in Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL) — System Demonstrations, 2016, pp. 115–120.

[8] J. Tetreault and M. Chodorow, "The Ups and Downs of Preposition Error Detection in ESL Writing," in Proc. 22nd International Conference on Computational Linguistics (COLING), 2008.

[9] D. Wible, C.-H. Kuo, F.-Y. Chien, and A. Liu, "A Web-based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers," Computers and Education, vol. 37, no. 3-4, 2001, pp. 297–315.

[10] J. Kakegawa, H. Kanda, E. Fujioka, M. Itami, and K. Itoh, "Diagnostic Processing of Japanese for Computer-Assisted Second Language Learning," in Proc. 38th Annual Meeting of the Association for Computational Linguistics (ACL), 2000.

[11] A. Kilgarriff, M. Husák, K. McAdam, M. Rundell, and P. Rychlý, "GDEX: Automatically Finding Good Dictionary Examples in a Corpus," in Proc. EURALEX, 2008.

[12] J. Didakowski, L. Lemnitzer, and A. Geyken, "Automatic Example Sentence Extraction for a Contemporary German Dictionary," in Proc. EURALEX, 2012.

[13] S. Smith, P. V. S. Avinesh, and A. Kilgarriff, "Gap-fill Tests for Language Learners: Corpus-Driven Item Generation," in Proc. 8th International Conference on Natural Language Processing (ICON), 2010.

[14] V. Baisa and V. Suchomel, "SkELL: Web Interface for English Language Learning," in Proc. Recent Advances in Slavonic Natural Language Processing, 2014.

[15] P. Merlo and E. E. Ferrer, "The Notion of Argument in Prepositional Phrase Attachment," Computational Linguistics, vol. 32, no. 3, 2006, pp. 341–378.

[16] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep Semantic Role Labeling: What Works and What's Next," in Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.

[17] M. McConville and M. O. Dzikovska, "Evaluating Complement-Modifier Distinctions in a Semantically Annotated Corpus," in Proc. LREC, 2008.

[18] G. Francis, S. Hunston, and E. Manning, Collins COBUILD Grammar Patterns 1: Verbs. London: HarperCollins, 1996.

[19] A. Villavicencio, "Learning to Distinguish PP Arguments from Adjuncts," in Proc. CoNLL, 2002.

[20] N. Kim, K. Rawlins, B. V. Durme, and P. Smolensky, "Predicting the Argumenthood of English Prepositional Phrases," in Proc. AAAI, 2018.

[21] J. Lee and S. Seneff, "Automatic generation of cloze items for prepositions," in Proc. Interspeech, 2007.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in Proc. ACL System Demonstrations, 2014, pp. 55–60.

[23] G. Francis, S. Hunston, and E. Manning, Collins COBUILD Grammar Patterns 2: Nouns and Adjectives. London: HarperCollins, 1998.

[24] P. Rychlý, "A Lexicographer-Friendly Association Score," in Proc. Recent Advances in Slavonic Natural Language Processing, 2008.

[25] D. Dahlmeier, H. T. Ng, and S. M. Wu, "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English," in Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications, 2013.