Learning Synchronous Grammar Patterns for Assisted Writing for Second Language Learners

Chi-En Wu, Jhih-Jie Chen, Jim Chang, Jason S. Chang
Department of Computer Science
National Tsing Hua University
{tony, jjc, jim, jason}@nlplab.cc

Abstract
In this paper, we present a method for extracting Synchronous Grammar Patterns (SGPs) from a given parallel corpus in order to assisted second language learners in writing. A grammar pattern consists of a head word (verb, noun, or adjective) and its syntactic environment. A synchronous grammar pattern describes a grammar pattern in the target language (e.g., English) and its counterpart in an other language (e.g., Mandarin), serving the purpose of native language support. Our method involves identifying the grammar patterns in the target language, aligning these patterns with the target language patterns, and finally filtering valid SGPs. The extracted SGPs with examples are then used to develop a prototype writing assistant system, called WriteAhead/bilingual. Evaluation on a set of randomly selected SGPs shows that our system provides satisfactory writing suggestions for English as a Second Language (ESL) learners.

1 Introduction
Lexicography is the discipline of analyzing the syntax, semantics, and pragmatics of the language to compile a dictionary, with a description of vocabulary and grammar. The compiling process involves time-consuming delineating word senses, analyzing grammatical information, and providing example sentences. Since 1970s, computational approach of statistical analysis of large-scale corpora was widely adopted in lexicography, which originates from the COBUILD project, led by John Sinclair, aiming at building a large-scale electronic corpus. The COBUILD project lead to dictionaries and grammar books, including Collins COBUILD Grammar Patterns 1: Verbs (Patterns, 1996) and Collins COBUILD Grammar Patterns 2: Nouns and Adjectives (Patterns, 1998). These two books describe grammar patterns of common verbs, nouns and adjectives in English, with the concept that most English words tend to follow only a limited set of patterns, which relates to the structure, usage, and the meaning of a word.

Later, Hunston and Francis (2000) propose Pattern Grammar with rules describing the intricate relation between word and grammar in one simple representational scheme, which explores the local regularities such as complementation structure, consisting of a headword with a sequence of preposition, noun phrase, verb phrase, clause (e.g., apologize for n), or a limited set of special words and phrases.

In this paper, we describe a method for automatically identifying the Chinese counterpart (e.g., “與 n 接觸”) of a given English grammar pattern (e.g., “contact with n”), along with the bilingual examples. Such pair of extracted patterns is call a Synchronous Grammar Pattern (SGP). SGPs can be used to support the compilation process of bilingual dictionary reducing the construction time and to improve the learning efficiency of ESL learners. With this in mind, we develop a prototype system, WriteAway, to assist writing for Chinese EFL learners.

2 Translation Pattern Assistant
We have implemented a prototype system as a web application, aimed at assisting second language learner in writing with native language support. At run time, WriteAway obtains the last content word the user just types in and displays relevant SGPs instantly as the user writes away. The prototype system, WriteAway, is accessible at https://spg-write.herokuapp.com
3 Extracting Synchronous Grammar Patterns

The extracting process involves recognizing the grammar patterns in the target language, aligning these patterns with their native language counterpart, and finally filtering valid SGPs with bilingual examples. We use a much simpler approach than previous work (Yen et al., 2015). We rely on a list of English grammar patterns from the HTML version of **COLLINS COBUILD GRAMMAR PATTERNS 1: VERBS** available at (http://arts-ccr-002.bham.ac.uk/ccr/patgram/). Therefore, the main focus is to identify the instances of these verb patterns and their counter part and to convert the counterpart instances into patterns.

3.1 Identifying English Grammar Patterns

In the identification process, we first use the GENIA Tagger (Tsuruoka et al., 2005) to shallow parse English sentences to obtain part of speech (POS) and chunk information (“B”,“I”,“O” symbols respectively indicate words at the beginning of a chunk, inside a chunk, and not part of NP, VP, ADJP, and ADVP).

Then, we identify head context words and elements of possible grammar patterns in the given sentences. Considering the input sentence “I apologize for my behavior.”, we identify the verb “apologize” as a headword “V” followed by the preposition “for” and a noun phrase “my behavior” with ‘n’ based on the simple relation between the parse results and the notation of Pattern Grammar. In so doing, we identify an instance of the pattern “V for n” for headword “apologize”, after we verify that this pattern can be found in **COLLINS COBUILD GRAMMAR PATTERNS 1: VERBS**. The phrase “apologize for my behavior” is retained for further processing (See Table 1).

3.2 Align English Pattern to Chinese

After obtaining the target language grammar patterns and instances for each headword, we then proceed to extract the corresponding native language grammar pattern and its example instances.

For that, we use a Chinese word segment system, CKIP (Ma and Chen, 2003), to tokenize and tag Chinese sentence with POS information. We also use a word aligner, **fast_align** (Dyer et al., 2013) to explore the crossing-lingual relationship between the target language and native language words (e.g., English and Mandarin words). Finally, we convert the aligned native counterpart instances into grammar patterns.

![Figure 1: The prototype system, WriteAway](image)

![Figure 2: SGP and example phrase extraction according to alignment](image)

| Word   | POS | B-I-O | Annotation | Pattern  |
|--------|-----|-------|------------|----------|
| I      | PRP | B-NP  |            |          |
| apologize | VBP | B-VP  | V          | (V for n) |
| for    | IN  | B-PP  |            |          |
| my     | PRP$| B-NP  | NP         |          |
| behavior | NN  | I-NP  | NP         |          |
| .      | .   | O     |            |          |

Table 1: Anchor ‘apologize for n’ to a sentence

See Figure 2 for an example of aligning “apologize: V for n” in a English sentence with its Chinese counterpart. When word alignment is 100% accurate, aligning and deriving synchronous patterns is straightforward. As shown in Figure 2, the headword “apologize” is aligned to “道歉”, the preposition “for” to “為” and the noun phrase “my behavior” to “我的行為” converted to the same phrase label “n”. Consequently, we can derive the SGP pair (e.g., “apologize for n”, <“為 n 道歉”>) from the aligned bilingual instance (e.g., “apologize for my behaviour”, <“為自己的行為道歉”>).

However, word alignment is prone to errors, causing the SGP extraction process to derive erroneous results. Typically, a target-language word may be aligned incorrectly leading to incorrect
links, missing links, or unnecessary links leading to an incorrectly identified counterpart instance and pattern.

Figure 3: A SGP extraction failure

An example of word alignment error is shown in Figure 3. The pattern “exchange n1 with n2” is retrieved from sentence “I should like to exchange a few words with you in confidence.”. However, we derive an incorrect grammar pattern “和 n2 交换 n1”, caused by the incorrect alignments (“words”, “私下”). If “words” is only related with “意見”, we obtain the correct grammar pattern “和 n2 交换 n1” and bilingual example (“exchange a few words with you” “和你交換點意見”).

To cope with word alignment errors, we stipulate that content phrase alignments are one-to-one. For one-to-two alignment, we select the longest consecutive Chinese segment (e.g., “點 意見”), and ignore the remaining disjoint segment (e.g., “私下”) aligned to an English phrase chunk, because longer segments tend to be correct. With this method, we can let the noun phrase “a few words with you” “和你交換點意見”.

3.3 Re-rank the Chinese Pattern

Table 2: The potential Chinese pattern templates for English pattern template ‘V for n’

| Rank | Ch Template | Frequency | Instance |
|------|-------------|-----------|-----------|
| 1    | V n         | 15900     | run for n, 競選 n |
| 2    | N n         | 2000      | run for n, 競選 n |
| 3    | P n         | 950       | 說 for n, 向 n 道歉 |
| 4    | n V         | 1900      | 說 for n, 向 n 道歉 |
| 5    | D V n       | 1850      | care for n, 報 規 照顧 n |
| 6    | V V n       | 1670      | care for n, 負 責 照顧 n |
| 7    | P V n       | 1400      | run for n, 為了 競選 n |
| 8    | P n V       | 1390      | apologize for n, 為 n 道歉 |

Table 3: rerank the Chinese patterns of ‘run for n’

3.4 Selecting Good Example Phrases

In order to give concrete examples of these rather abstract synchronous grammar patterns, we extend the method described in (Kilgarriff et al., 2008) to select bilingual examples from the parallel corpus. The principles are as follows:

1. Correctness (English). The length of English pattern example multiplied by \( r \) must be similar with the length of the Chinese pattern example. Note that \( r \) is the average sentence length ratio between English and Chinese. This is to avoid selecting examples with word alignment errors.

2. Readability. Let \( l_E \) and \( l_C \) be the aver-
4 Evaluation

Our evaluation focused on verifying the correctness of extracted SGPs. First, we grouped SGPs by their corresponding English pattern templates. Next, we randomly sampled 10 English grammar patterns from each group along with top 5 corresponding Chinese grammar patterns. Then, we asked two linguists to assess the appropriateness and quality of using the SGP for translation. In the assessment, each Chinese pattern is given a label of (C)orrect, (A)cceptable or (I)ncorrect. We evaluated a set of 1,497 Chinese grammar patterns for 31 different types of English patterns. Table 4 lists the counts and the proportion of the annotation results. There are 44% SGPs tagged with CC, 5.5% with CA, and 20% with AA. Overall, there are approximately 70% sampled SGPs are correct or acceptable.

In addition, we calculated the average score of the evaluation while assessing the scores of C = 2, A = 1 and I = 0. The average score is 1.2, which indicates that the results are only slightly better than acceptable, and obvious there is much room for improvement.

5 Conclusion and Future Work

In this paper, we have presented a method for automatically extracting Synchronous Grammar Patterns from a parallel corpus. The procedure involves extracting English patterns from parallel corpora, performing alignment of pattern sequences to Chinese sequences, generating and re-ranking counterpart Chinese patterns. The evaluation results show that our approach provides mostly correct or acceptable translation patterns that can be effectively exploited in assisted writing for second language learners. For that, we have also developed a prototype system so that ESL learners can write more confidently and frequently based on the synchronous grammar patterns displayed by the system.

We also conducted a preliminary investigation into the origins of incorrect SPGs and found that these errors were mainly due to alignment errors and segmentation errors. Moreover, idioms are usually hard to aligned and generalized into an SPG (e.g., “樂不思蜀” to “reluctant to leave”). Overall, common patterns with literal translation tend to lead to correct and useful SPGs for learner-writers, implying that a larger corpus can help producing more accurate SPGs. We will continue to work on the cases of SPG for nouns and adjectives.

References

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Susan Hunston and Gill Francis. 2000. Pattern grammar: A corpus-driven approach to the lexical grammar of english. Computational Linguistics, 27(2).

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In Proc. Euralex.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to ckip chinese word segmentation system for the first international chinese word segmentation bakeoff. In Proceedings of the second SIGIHAN workshop on Chinese language processing-Volume 17, pages 168–171. Association for Computational Linguistics.

Collins COBUILD Grammar Patterns. 1996. 1: Verbs. Collins COBUILD, the University of Birmingham.

Collins Cobuild Grammar Patterns. 1998. 2: Nouns and adjectives.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun’ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In Panhellenic Conference on Informatics, pages 382–392. Springer.

Tzu-Hsi Yen, Jian-Cheng Wu, Jim Chang, Joanne Boisson, and Jason S Chang. 2015. Writeahead: Mining grammar patterns in corpora for assisted writing. In ACL (System Demonstrations), pages 139–144.