

An Information Retrieval Approach to Facilitating Second Language Learning

Jyi-Shane Liu^a, Ching-Ying Lee^b, Pei-Chung Hung^a

^aDepartment of Computer Science, National Chengchi University, Taiwan, R. O. C.

^bDepartment of English, National Taiwan Normal University, Taiwan, R. O. C.

jshliu@cs.nccu.edu.tw

Abstract: Data driven language learning promotes learner autonomy and discovery learning by providing learners with authentic foreign language data for self-directed or guided exploration. However, the effective use of corpora data requires a certain level of linguistic knowledge. We propose an information retrieval augmentation to concordance for adapting to self-directed context of independent learners. The approach involves an expression element model and a retrieving mechanism so as to reduce linguistic threshold and enhance learner empowerment. Simulation results with English proficiency tests and students' writing samples support the effectiveness of the approach.

Keywords: Data driven language learning, information retrieval, learner empowerment.

Introduction

Data Driven (language) Learning (DDL) emphasizes the idea of viewing and transforming language data as direct knowledge sources for learners to acquire [3]. The learning process is learner-initiated and emphasizes the learners' own effort to form and test hypotheses about the target language so as to gain linguistic insight and promote language proficiency. While DDL methodology and benefits have been discussed in much of the work in the literature, prerequisites and cautions in practical educational contexts have also been noted [6]. Corpora and concordance alone cannot ensure the realization of DDL advantages. Both teachers and learners have to be trained and adjust to pedagogical changes in roles, knowledge, skills, and materials [1]. Students' frustration in using concordance had also been observed due to query formulation error and considerable time consumption [8].

The notion of "learners as researchers" in DDL seems to be arguably over-stretched for less advanced learners with no assistance, especially in the second language context. Formulation of productive queries is a crucial first step in using concordance and is particularly difficult for language learners without proper assistance [6]. Misspellings which spoil productive queries are common. Successful use of substitutes requires near-native competence in anticipating word alternatives. It is also difficult for learners to independently phrase queries that will expose solutions to their language use problems. DDL employs corpus and concordance to confirm or reject *a priori* assumptions about language use. Independent learners may quickly run out of ideas in the explicit language awareness task of elaborating assumptions and fail to drive any useful conclusions.

In this paper, we argue that concordance programs must be augmented for language learners working in an independent context. We propose an information retrieval augmentation to concordance that provides flexible query specifications and relevance ranking of language use examples. Hypothesis formation becomes adaptable to the

language capacity of users in the form of adjustable query constraints. The approach empowers language learners to conduct self-directed corpus reference and learning activities with less demanding level of language capacity. Simulation results with English proficiency tests and students' writing samples verify the utility of the approach.

1. Limitations of Concordance in Independent Learning

The rationale of using concordancing as a data driven learning tool rests on the view that second language proficiency is primarily promoted by learners' own effort to form and test hypothesis about the target language. Previous language learning concepts, such as language awareness [2] and consciousness raising [5], sustain that second language acquisition processes may be enhanced by the conscious and systematic exploration of the target language. Our position on exploiting language resources for language learning is in the same line with the current trend of corpus tools development. However, we argue that some augmentation effort must be made to enhance concordance programs for the purpose of independent language learning. Successful inductive learning of second language from corpus exploration comes from well-prepared pedagogical activities and specific instructions in a classroom setting [7]. Concordance programs, in most of their current forms, pose several limitations in supporting learners' self-directed exploration and reference for second language learning.

- *Linguistic Threshold*: The difficulty of forming a productive hypothesis entails a linguistic threshold to make the best use of concordance and imposes an obstacle for learners with insufficient language knowledge. Less advanced independent learners will not be able to draw much benefit from concordance exploration with frustration emerging as a foreseeable effect.
- *Insufficient Empowerment*: The current approach to concordancing is intended to be used as an overall observation tool on language phenomena of specific target words. Learners with vague ideas and incorrect assumptions of language use are not sufficiently empowered to conduct experimental searches and gradually narrow down to answers with increasing evidences.
- *Lack of Adaptability*: Most of the current forms of concordancing lack adaptability to learners' different language levels and learning purposes. From the standpoint of information service provision, current concordancing mechanisms are not context dependent and are not useful for a larger proportion of second language learners.

2. Information Retrieval Augmentation to Concordance

We present an information retrieval augmentation to concordance for providing more flexible language information services and adapting to self-directed context of independent learners. The approach is designed to anticipate users with various levels of language knowledge and provide flexible query forms to cope with language knowledge gaps or misconceptions of second language learners. The language information retrieval method is a process that involves receiving learners' language information need with flexible query representation, retrieving sentences from data sources that potentially match learners' inquiry, and recommending a set of relevant sentences in a ranked order.

2.1 Expression Element Module

The expression element module is designed to offer query flexibility for learners to represent their various forms of language knowledge.

- **Exact words:** Exact words are used when learners are able to provide correct and complete spelling of a word or multiple words.
- **Prefix and Suffix:** Prefix (or suffix) (%) of a word allows learners to specify partial spelling of a word to represent their incomplete word memory.
- **Wildcard:** A wildcard (#) allows learners to include an uncertain or unspecified single word as part of the query.
- **Part-of-Speech (POS):** POS tags provide an additional constraint when learners possess word class knowledge about the target word, for example, preposition (P), adjective (J), noun (N), adverb (D), and verb (V).
- **Subsequence:** Subsequence (*) represents a sequence of zero to multiple unspecified words. It is designed to allow convenient inquiry for some phrasal structures and word combination that include various contexts in the middle, such as “rather ... than”.

The set of expression elements constitute a space of language use constraints in which learners can select appropriate expression elements to represent a particular language information need.

2.2 Retrieval Module

Retrieval module performs the task of matching query with data items (sentences) in the corpus and selecting proper candidates that may provide useful examples for learners' language information need. We adopt both search rules of exact match and partial match. Exact match requires that all the occurrence conditions specified in the query are satisfied by the candidate sentences in the corpus. Partial match allows the selection of a candidate sentence in which some of the constraints in the query are not met. Consider a scenario in which an ESL/EFL (English as Second/Foreign Language) learner has only a partial idea of the two words “only” and “also” in the complete phrasal structure of “not only ... but also”. The learner can specify a query in the form of “only also” and partial match will retrieve sentences that contain the phrasal structure of “not only ... but also”. This will offer the learner an opportunity to recognize the complete phrase and learn its proper usage.

2.3 Ranking Module

The purpose of ranking module is to improve the effectiveness of language use consultation so that learners may obtain necessary information in a short list. The ranking module evaluates the selected candidates and sorts the reference list in a decreasing relevance order. We adopt the Multiple Sequence Alignment (MSA) technique to perform the relevance evaluation between query and sentence candidates. MSA technique [4], previously developed in bioinformatics, has been used to evaluate the similarity between biological (such as DNA or protein) sequences.

2.4 Language Information Inquiry

The approach was implemented in a system called SAW. We use some query cases to demonstrate actual result of SAW with BNC (British National Corpus). Four queries, “deter% by”, “native P”, “responsible #”, and “either * or”, are used as test examples. They cover the expression elements of prefix, exact word, POS, wildcard, and subsequence. Both search options are tested to compare their effects. In particular, partial search is used in the query of “either * or”, while exact search is used in the other three cases. Figure 1 shows the actual image of SAW with the test query “deter% by”. A total of 721 sentences were

retrieved from BNC and ranked by relevance. A unit of ten sentences was displayed in a page for quick reference. The list of matched words and number of use cases in the corpus are: determined (676), deterred (27), determination (12), deterrence (2), determine (2), determinations (1), determining (1). Table 1 summarizes reference results of the other three queries. The current match score setting favors shorter sentences.

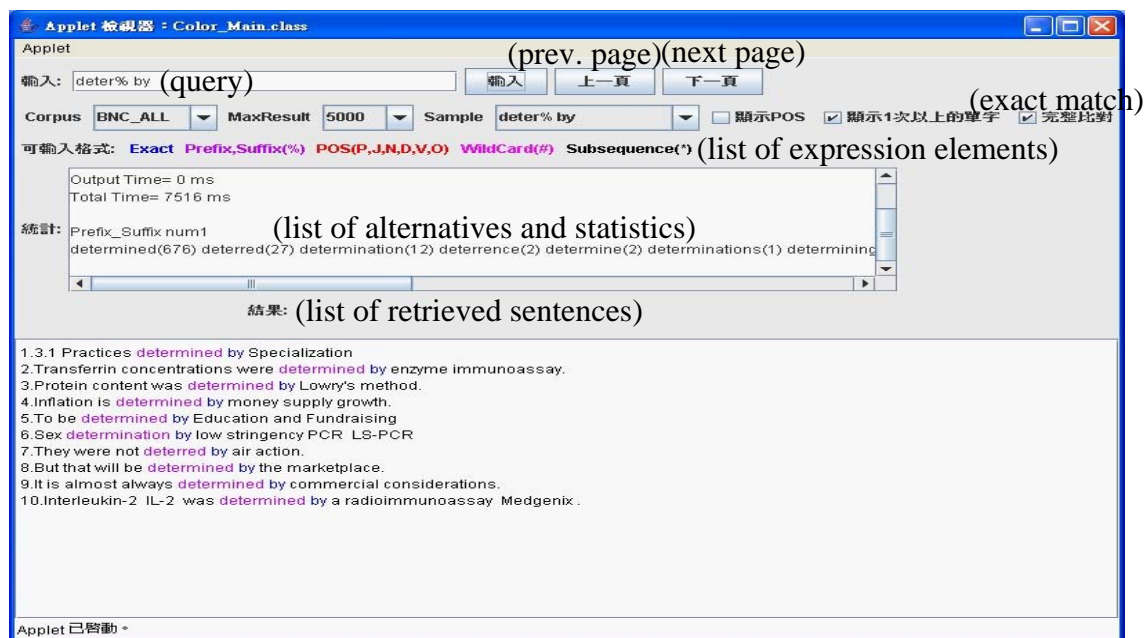


Figure 1. SAW's Use Result of Query "deter% by"

3 Experiment and Evaluation

We used question items in English test to simulate language use problems of ESL/EFL learners on collocations. The English test conducted in the experiment consists of question items concerning 10 two-word collocations (specialized in, familiarity with, guard against, etc.). Four types of expression elements, exact word, prefix, POS, and wildcard, were used to form the group of testing queries for each question item. Except queries by exact words, all other types of queries also contain combinational variations. Retrieved results were evaluated by the performance measure of precision in information retrieval and were averaged over the same type of query variations.

The first type of query was composed of exact words of the target collocation, such as "specialize in", and contains no variation. The second type of query was formed by the combination of an exact word and a prefix. One of the two words in the collocation was given as an exact word, and the first letter of the other word was given as prefix. For the collocation of "specialize in", two variations, "s% in" and "specialize i%", were tested. The third type of query assigned a POS tag on one of the two words in the collocation. For the same collocation "specialize in", two variations, "V in" and "specialize P", were formed. The fourth type of query used a wildcard in one of the two words in the collocation, such as "# in" and "specialize #". The precision of the retrieved result was computed over the selected ten sentences. As expected, query type of exact words has the highest precision, which is 1. The average precision of prefix, POS, and wildcard over the ten collocations are 0.45, 0.305, and 0.14, respectively. The results indicate that, other than exact words, prefix imposes stronger constraints than POS and wildcard. Yet, they can all provide referential answers for learners' language use problems.

Table 1. SAW's Use Result of Other Test Queries

Query: native P	Statistics: of (76), to (40), on (2), in (1), by (1), with (1), at(1)
Top 5 retrieved sentences:	<ol style="list-style-type: none"> 1. Jacklin goes native <i>at</i> Dalmahoy. 2. Both are beers native <i>to</i> Munich. 3. It is not native <i>to</i> my disposition. 4. A native <i>of</i> Tralee, Mairead is single. 5. The pie is native <i>to</i> both Cheshire and Lancashire.
Query: responsible #	Statistics: for (2596), to (104), and (45), in (19), are (15), government (15), if (13), people (12), person (11), attitude (9), ...
Top 5 retrieved sentences:	<ol style="list-style-type: none"> 1. Am I responsible <i>for</i> that? 2. Is he responsible <i>about</i> money? 3. The agency responsible <i>was</i> GGK. 4. I am responsible <i>for</i> security. 5. We are responsible <i>for</i> legislation.
Query: either * or	Statistics: N/A
Top 5 retrieved sentences:	<ol style="list-style-type: none"> 1. She either <i>barked</i> or <i>shouted</i>. 2. He either <i>knows</i> or <i>not</i>. 3. Choose either <i>mild</i> or <i>low</i>. 4. You either <i>love</i> or <i>hate</i> it. 5. Mills were either <i>demolished</i> or <i>converted</i>.

4. Conclusion

We observe that concordancing, as a tool for data-driven language learning, pose obstacles for independent learners. This has prohibited a large portion of second language learners to exploit corpus as a self-accessed language resource and achieve better learning result. We propose to augment concordance programs with an information retrieval approach such that language use information embedded in corpus can be conveniently revealed with flexible query and retrieval. The approach was implemented and tested with simulated language use problems in English capability tests and students' writing samples. Experimental results show the approach was successful in supporting self-directed language learning.

Acknowledgements

This research was partially supported by Taiwan's National Science Counsel under grants NSC-95-2422-H-004-003 and NSC 96-2422-H-004-001.

References

- [1] Gabrielatos, C. (2005). Corpora and language teaching: just a fling or wedding bells? *Teaching English as a Second or Foreign Language*, 8(4), 1-37.
- [2] Hawkins, E. (1987). *Awareness of Language: an Introduction*. Cambridge: Cambridge University Press.
- [3] Johns, T. (1991). From printout to handout: grammar and vocabulary teaching in the context of data driven learning. In T. Johns & P. King (eds.) *Classroom Concordancing*, special issue of *English Language Research Journal*, 4, 27-45.
- [4] Needleman, S., & Wunsh, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, 48, 443-453.
- [5] Rutherford, W., & Smith, M. S. (1985). Consciousness-raising and universal grammar. *Applied Linguistics*, 6(3), 274-28.
- [6] Stevens, V. (1995). Concordancing with language learners: why? when? what? *Computer-Assisted English Language Learning*, 6(2), 2-10.
- [7] Tribble, C., & Johns, G. (1997). *Concordances in the Classroom*. Athelstan.
- [8] Whistle, J. (1999). Concordancing with students using an 'off-the-web' corpus. *ReCALL*, 11(2), 74-80.