行政院國家科學委員會專題研究計畫　期中進度報告

總計畫(1/3)

計畫類別：整合型計畫
計畫編號：NSC93-2524-S-032-004-
執行期間：93年05月01日至94年04月30日
執行單位：淡江大學英文學系

計畫主持人：衛友賢
共同主持人：柯華葳

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94年4月22日

分散式認知, 電腦運算與隨處的數位語言學習-運用自然語言處理於建置網路英語學習環境(1/3)

計畫編號：NSC 93－2524－S－032－004
執行期限：九十三年五月一日至九十四年四月三十日

主持人： 衛友賢　 淡江大學英文系

Introduction

The purposes of this overall grand project are to develop digital tools and architecture for implementing browser—based ubiquitous online English vocabulary learning as well as a theoretical framework for motivating the design of these tools and this architecture and for investigating their effects on users' vocabulary learning. Given the uniqueness of the project's research contribution, a further goal is to disseminate the emerging results especially through international collaborations. The first year has witnessed substantial progress in each of these main areas: theory development, novel digital tool design and implementation, and international collaboration with academic institutions in Europe (Louvain University, Belgium), the US (Columbia University, NY) and Canada (University of Toronto). The following provides an overview of the first year's progress, covering our established international collaborations and summaries of the individual project's first-year results.

International Collaborations

We have established international collaborations with institutions in Canada, the United States, and the EU, Belgium. The following describes these three briefly.

University of Toronto, Canada:

The Knowledge Forum platform has been developed at the Ontario Institute for Studies in Education under the direction of Professors Marlene Scardamalia and Carl Bereiter and is one of the world's most well-known and respected digital learning platforms. During a visit by Wible to OISE sponsored by Scardamalia's research project, a collaboration was established whereby the ubiquitous language learning tools developed by our team would be integrated into the Knowledge Forum platform. The basic mode of integration has been arranged and details will be worked out during second visit to Toronto in the summer of 2005, again sponsored by Scardamalia's research project.

University of Louvain, Belgium, EU:

Professor Sylviane Granger, former director of the Centre for English Corpus Research at the University of Louvain, Belgium, and the current director of that University's Language Institute (10,000 language students and 70 instructors) has asked to collaborate with us in implementing our system in their Language Institute as the main language learning platform. The ubiquitous tools will be an integral part of the platform we offer, and their Language Institute's use of the tools will provide valuable data on the tools effects on learning.

Columbia University, New York, USA:

Our research group led a visit to Columbia University, Columbia Teachers College (TC) in June 2004 for exchanges with Professor Jo Anne Kleifgen and her colleagues and graduate students on topics of digital language learning. The Columbia University Ph.D. students also engaged in online discussions and demonstrations with English teachers from Taiwan using the IWiLL platform for these discussions. The participants then met face to face when the Taiwan group traveled to New York for a visit to Kleifgen's group at Columbia.

The trip was a remarkable success, and our two groups are now finalizing plans for a two-day research roundtable on digital language learning at Columbia University for the summer of 2005. This roundtable is to include more senior Columbia researchers and more researchers from our group in Taiwan. The groups are currently proposing projects for longer term collaborative research on digital language learning, and we hope this summer's roundtable will be the first of many.

Individual Project 1: Second Language Learning and its Interfaces (NSC 93-2524-S-032-005)

Introduction

The purpose of this subproject is to provide the theoretical framework and methodologies for investigating the effects that the digital tools and resources designed in the other two subprojects wield on the language learning of the target users and to supply criteria for the design of these tools, criteria motivated by research in instructed second language acquisition and distributed cognition. The targeted domain of language learning is narrowed to English as a second language and to vocabulary learning. Within this first year of funding, the project has produced substantive results in the construction of such a theoretical framework and in the design requirements that have led to breakthroughs in the design of novel digital tools for ubiquitous language learning. These results have attracted the attention of international scholars and led to concrete international collaborations.

Distributed Cognition

We have taken the notion of distributed cognition as central to our way of conceptualizing the relationship between the language learner and the digital tools that learner uses. Our contribution is to apply this construct to the framing of investigations of digital learning. The role for distributed cognition within a research framework for digital language learning has been described and motivated in detail in Wible (2005) "Distributed Cognition, Learning Events, and Research on Digital Language Learning" (submitted to Computers and Education, Nov 2004, ms 54 pp.).

Distributed cognition plays two sorts of roles in our project: (1) as a construct in the research framework that we are developing, and (2) as a construct in the empirical investigations we are currently designing and conducting. One of our empirical investigations will be reported at EuroCALL 2005 in Wible, Kuo, Wang, and Tsao (2005). In that investigation we tracked the learners use of and attentiveness to feedback on collocation errors they made in both translation exercises and essay writing. Moreover, three groups received corrections on these errors of different levels of explicitness. The results show that the explicitness of digital feedback as well as the attentiveness of subjects to this feedback had statistically significant differential effects on the learners' improvements in their use of the targeted collocations over a one month period.

Ubiquitous Online Language Learning

This notion of ubiquitous online language learning is a novel conceptualization for implementing language learning support in noisy digital environments. It contrasts sharply with both current digital learning approaches as well as traditional classroom learning approaches. The conceit of our approach is that it creates a natural fit between the structure and content of the World Wide Web as a source of language input for language learners on the one hand and second language acquisition theory which gives a primary role to target language input for learners within authentic communicative contexts that fit the learners' own purposes and interests on the other.

One of our two full papers presented at IEEE's ICALT2004, entitled "Contextualized Language Learning in the Digital Wild: Tools and a Framework," was a presentation of the philosophy of ubiquitous online language learning along with a description of two novel tools we developed in cooperation with the second individual project (Kuo) which implement this philosophy: Collocator and Word Spider.

At EuroCALL2004 (Vienna) we presented a full paper on a novel design for learner corpora which we call Dynamic Event-Driven (DyED) Learner Corpora , one which we have already implemented in our 3-million word corpus of Taiwan leaners' English (EnglishTLC). The design captures not only the texts written by these learners in English, but the contexts of the writing. In other words, while current learner corpora a taken as collection of texts, the DyED learner corpus design takes these texts as writing events that occurred in situated contexts. The corpus indexes the learners' texts to the assignments which evoked the texts, to all of the teacher's feedback on each text, to the revisions that each learner made in response to these instances of teacher feedback, and so on. Thus one piece of student writing in our current version of EnglishTLC is indexed to an entire mesh of related digital events that may exert influences on this writing. Thus the DyED corpus design supports a much richer array of research on student language learning and student writing than other current learner corpora.

Domain Knowledge Results

An important feature of the research project is a clearly defined focus on a narrow scope of the learning domain. Specifically, our work is limited to the learning of lexical knowledge (vocabulary acquisition). Within this domain, a crucial research niche is how digital tools (specifically computational linguistic techniques) can be applied in novel ways to extract the sorts of linguistic knowledge that is intended for the learners. In this case, it is how to extract lexical knowledge. On this front, we have this year made novel contributions in computational lexicography. In fact, we could say that we have seen a breakthrough in our work which sets it apart from other work being done internationally within the same domain. This work is a close collaboration between this individual project and the NL2P individual project—a combination of linguistic and pedagogical expertise on the one hand and computational techniques on the other. What follows is a brief progress report on this work.

Learnability and Collocations

Two of the digital tools developed in Individual Project 2 (Kuo) are designed to detect collocations. The Collocator in particular is designed to do this in noisy unrestricted environments where learners browse freely. Here we motivate the tool from the perspective of language learning and learnability.

An apparent source of difficulty for learners in acquiring collocations is that they are idiosyncratic. For example, there would appear to be nothing in the meaning of the words involved which would predict that *make a conclusion* odd whereas *draw a conclusion* is acceptable, or that we can intensify the noun *respect* with the adjective *great* (They have *great respect* for her) but not with the near synonymous adjective *big* (*They have *big respect* for her). These restrictions illustrate the phenomenon of collocation: many words are unpredictably picky about the other words with which they can co-occur. The central motivation for our Collocator tool is that this pickiness (or 'collocability'), which learners must master, is not detectable from their direct encounters with target language input.

Lexical Clustering

Our lexical research has extended beyond collocations this year. Collocations are typically word pairings (though some may consist of 3 words). A less well-defined concept of word combinations has become popular in second language teaching recently called lexical chunks. Lexical chunks are combinations of words regularly used together, such as '…with respect to…' or '…in the event of…' These are extremely important components of a proficient English user's lexical competence. They also represent a wide-open research niche which we have identified and are currently working within. While numerous books and articles now urge language teachers to teach lexical chunks, there are no resources providing these chunks. There are no reference books of 'English lexical chunks' (though there are numerous collocation dictionaries). There are also no computational tools that extract chunks (though there are many that extract collocations).

Again, in collaboration with Individual Project 2 (NLP), we have developed a prototype of an automatic lexical chunk extractor. It is based on greedy algorithm we have designed which identifies strings of words of any length (including strings of non-contiguous words) that constitute lexical chunks worth learning. For example, given the word 'fact' as input and no other guidance, our current prototype automatically discovers the chunk: 'despite the fact that' as well as collocations like 'In fact' among others. It does this by running our algorithm on 20 million words of the British National Corpus. Given the target word 'point' as input, it detects chunks such as 'from the point of view of…' We are currently refining the algorithm to improve precision and recall and designing an interface for learners. The current interface is for internal research use only.)

Individual Project 2: NL2P (NSC 93-2524-S-032-006)

Problem formulation

NL2P denotes Natural Second Language Processing. It is different from conventional Natural Language Processing in the sense that these language inputs and outputs are highly related with target language learners. Not only learner language output such as writing and speaking, but also learner language input such as what they see, hear and acquire are of concern in this subproject. There are twofold needs to be further addressed.

1. Handling noisy data

The language inputs that learners acquire in real world are not always like hand-crafted lessons of specific language curricula, well-formatted and flawless. Since we want to extract all features of language inputs, e.g., what learners browse on the web, NL2P tools that are able to handle language inputs in a noisy environment is desired for networked language learning.

2. No intrusion

One of this project's goals is providing learning assistant or drawing learners' attention on something worth to learn while learners roam in digital English environment. No matter what kind of effects these assistants produce (the effects are discussed in learning theory and personalization sub-project), these NL2P tools shouldn't make any noise which would interfere with learners' behavior and cognitive process of learning.

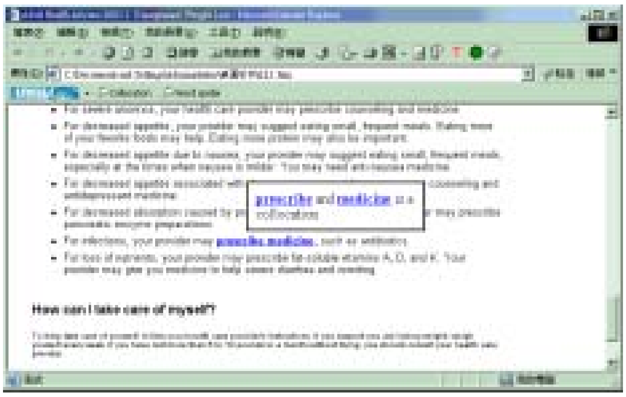
NL2P prototyped tools

Based on the needs described above, we have developed two NL2P tools in the first year of the project. Both of them are embedded in Web Browser. Here are the detailed descriptions about these two tools:

1. Collocator

The purpose of Collocator is to offer the learners collocational knowledge from a single reading experience. The approach of Collocator is to enhance reading texts that are freely selected by learners on line, highlighting for them in real time precisely those word combinations in the text that are collocations. The tool detects such collocations in the learner's text in real time by exploiting statistical word association measures on a 30-million-word portion of the British National Corpus (BNC). Combinations of words that achieve a sufficiently high association score (calculated on BNC) to constitute collocations and which co-occur within a specified window of proximity to each other in the targeted text are highlighted there as potential collocations.

Figure 1 shows the collocation *prescribe medicine* highlighted by Collocator and a pop-up text indicating its status as a collocation. A link from each detected collocation is added in real time which lists additional examples of this same collocation in order to provide users with richer and more intensive exposure to the same collocation.



prescribe and medicine is a collocation

For infections, your provider may prescribe medicine, such as antibiotics.

How can I take care of myself?

Figure 1. Collocator

2. Similar Pages Provider

The purpose of Similar Pages Provider is to offer similar information on the web and raise the related keywords exposure for users. Like Collocator, the reading texts are selected by users on line and a list of similar pages is then provided in another browser window, shown as Figure 2. The approach to finding similar pages is based on the hyperlink-structure of the WWW. The approach assumes that the content of two pages are similar if they have the highly overlapped backward and forward links. For example, if two pages are both pointed by CALICO, we assume that both pages are about CALL. This tool also can provide high repeated vocabulary exposure about a single event, which is suggested by many language learning researchers.

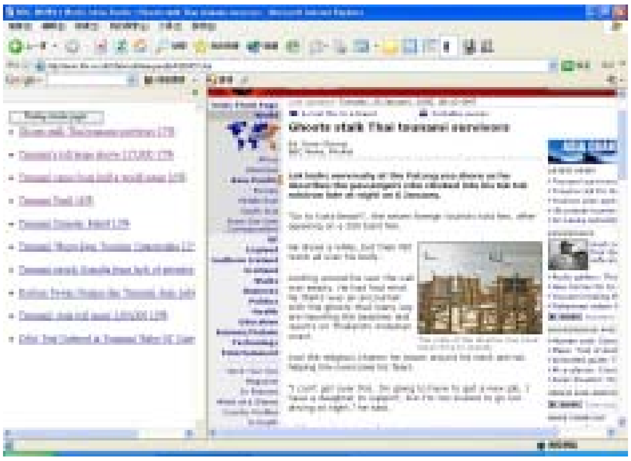


Figure 2. Similar Pages Provider

Individual Project 3: Knowledge Management (93-2524-S-001-001)

Introduction

In the report, we summarize the research results in the period from May 2004 till now. There are three research problems studied in this project, namely (1) progressive analysis scheme for web text classification, (2) life profile based event detection from text stream, and (3) auto-quiz for ubiquitous English networked learning.

Progressive Analysis Scheme

The first problem studied is aiming for managing web documents based on their structural features. In writing convention, the authors usually emphasize representative terms by surrounding them with intense-emphasized tag-pairs. Thus, during classification, we can identify terms' classification contributions based the cues from surrounding tag-pairs, and only extract contributive terms for classification.

In the study we propose a novel region-based web document classification approach, Progressive Analysis Scheme (PAS). PAS analyzes tag-regions progressively in descending order of their sound emphasized degrees, until there are sufficient aggregated features for category confirmation. The *enhanced surroundings modeling* and *evaluation* strategies produce the sound emphasis degrees of surroundings models as well as the emphasized degrees of tag-regions surrounded. To avoid falling into emphasis traps again, the *emphasized degree adjustment strategy* revises dynamically emphasized degrees of related noisy-tendentious tag-regions while the classifier falls into a trap. And thus, the tag-region analysis sequence constantly adapts to the real contributions of tag-regions during classification. Because the quality and quantity of analyzed tag-regions are controlled carefully, the proposed approach can achieve high classification performance and low computation cost. The experimental performance verifies the expected merits of PAS.

Life Profile Based Event Detection

The second research topic is aiming for managing web documents based on their relationship in time sequence. The publishing activities on the Internet nowadays are prevailing that when an event occurs, autonomous reporters may publish related documents along the event's life span. These documents of an event are sequenced according to their chronological order. For major events, different and even changing contexts make traditional text clustering algorithms incapable of detecting and tracking the events effectively. To detect events in such a dynamic environment, we have developed an adaptive method to detect and track diverse, changing, heterogeneous documents.

In this study, we propose the concept of life profiles which utilizes hidden Markov models to model the dynamic activeness of events. Each life profile has a distinct development of activeness status and can be learned from training events with similar activeness status developments. During the event detection process, each cluster is associated with appropriate life profiles and together with the activeness status to determine a dynamic clustering similarity threshold. We propose an event detection method called LIPED (LIfe Profile based Event Detection). It incorporates the proposed life profile framework with an incremental clustering algorithm for effective event detection. The experiments show LIPED achieves at least 5% improvement to other known approaches

Auto-Quiz for Ubiquitous English Learning

The third topic is aiming for a mechanism and for automatic quiz giving and grading for ubiquitous English learning. When a learner browses a English web document from Internet, there is a learning activity going on and the idea of this study is to construct a semantic network for the browsed text and give quiz according to some quiz-form templates and learner status. The core of this mechanism is to "understand" the text and to prepare quiz template broad enough to cover all the needs. Currently, we have exploited a grammar analyzer and

Wordnet to construct semantic network for text and underway to find a representation for the quiz template and work on some simple examples. The study is still in its early stage as proposed in the project proposal.