TransAhead: A Writing Assistant for CAT and CALL

[*]Chung-chi Huang [++]Ping-che Yang [*]Mei-hua Chen [*]Hung-ting Hsieh [+]Ting-hui Kao [+]Jason S. Chang

[*]ISA, NTHU, HsinChu, Taiwan, R.O.C.
[++]III, Taipei, Taiwan, R.O.C.
[+]CS, NTHU, HsinChu, Taiwan, R.O.C.
{u901571,maciaclark,chen.meihua,vincent732,maxis1718,jason.jschang}gmail.com

Abstract

We introduce a method for learning to predict the following grammar and text of the ongoing translation given a source text. In our approach, predictions are offered aimed at reducing users' burden on lexical and grammar choices, and improving productivity. The method involves learning syntactic phraseology and translation equivalents. At run-time, the source and its translation prefix are sliced into ngrams to generate subsequent grammar and translation predictions. We present a prototype writing assistant, TransAhead[1], that applies the method to where computer-assisted translation and language learning meet. The preliminary results show that the method has great potentials in CAT and CALL (significant boost in translation quality is observed).

1. Introduction

More and more language learners use the MT systems on the Web for language understanding or learning. However, web translation systems typically suggest a, usually far from perfect, one-best translation and hardly interact with the user.

Language learning/sentence translation could be achieved more interactively and appropriately if a system recognized translation as a collaborative sequence of the user's learning and choosing from the machine-generated predictions of the next-in-line grammar and text and the machine's adapting to the user's accepting /overriding the suggestions.

Consider the source sentence "我們在結束這個交易上扮演重要角色" (We play an important role in closing this deal). The best learning environment is probably not the one solely providing the automated translation. A good learning environment might comprise a writing assistant that gives the user direct control over the target text and offers text and grammar predictions following the ongoing translations.

We present a new system, TransAhead, that automatically learns to predict/suggest the grammatical constructs and lexical translations expected to immediately follow the current translation given a source text, and adapts to the user's choices. Example TransAhead responses to the source "我們在結束這個交易上扮演重要角色" and the ongoing translation "we" and "we play an important role" are shown in Figure 1[2](a) and (b) respectively. TransAhead has determined the probable subsequent grammatical constructions with constituents lexically translated, shown in pop-up menus (e.g., Figure 1(b) shows a prediction "IN[*in*] VBG[*close*, *end*, …]" due to the history "play role" where lexical items in square brackets are lemmas of potential translations). TransAhead learns these constructs and translations during training.

At run-time, TransAhead starts with a source sentence, and iteratively collaborates with the user: by making predictions on the successive grammar patterns and lexical translations, and by adapting to the user's translation choices to reduce source ambiguities (e.g., word segmentation and senses). In our prototype, TransAhead mediates between users and automatic modules to boost users' writing/ translation performance (e.g., productivity).

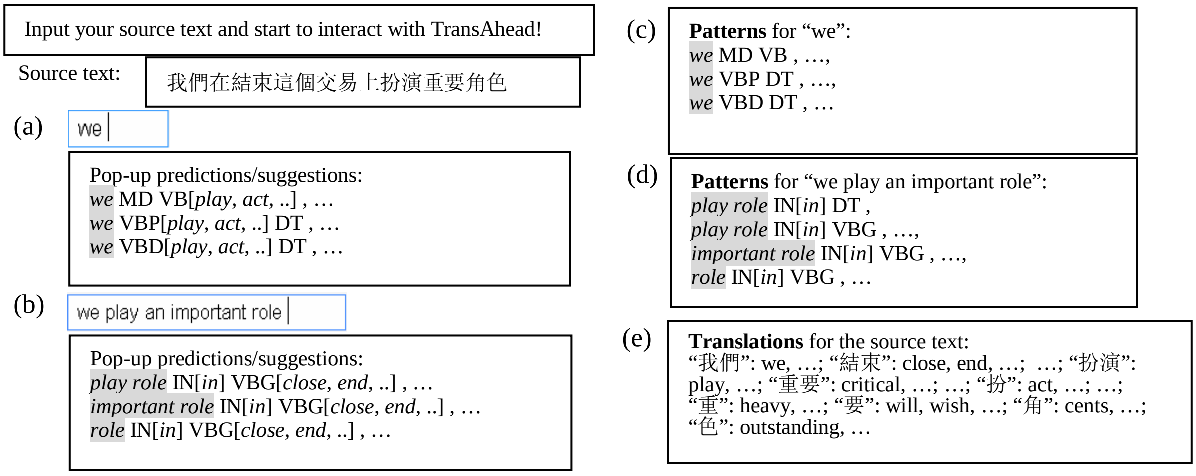
2. Related Work

CAT has been an area of active research. Our work addresses an aspect of CAT focusing on language learning. Specifically, our goal is to build a human-computer collaborative writing assistant: helping the language learner with in-text grammar and translation and at the same

[1] Available at http://140.114.214.80/theSite/TransAhead/ which, for the time being, only supports Chrome browsers.

[2] Note that grammatical constituents (in all-capitalized words) are represented using Penn parts-of-speech and the history based on the user input is shown in shades.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 16–19,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics



Input your source text and start to interact with TransAhead!

Source text: 我們在結束這個交易上扮演重要角色

(a) we

Pop-up predictions/suggestions:
we MD VB[*play*, *act*, ..] , …
we VBP[*play*, *act*, ..] DT , …
we VBD[*play*, *act*, ..] DT , …

(b) we play an important role

Pop-up predictions/suggestions:
play role IN[*in*] VBG[*close*, *end*, ..] , …
important role IN[*in*] VBG[*close*, *end*, ..] , …
role IN[*in*] VBG[*close*, *end*, ..] , …

(c) **Patterns** for "we":
we MD VB , …,
we VBP DT , …,
we VBD DT , …

(d) **Patterns** for "we play an important role":
play role IN[*in*] DT ,
play role IN[*in*] VBG , …,
important role IN[*in*] VBG , …,
role IN[*in*] VBG , …

(e) **Translations** for the source text:
"我們": we, …; "結束": close, end, …; …; "扮演": play, …; "重要": critical, …; …; "扮": act, …; …; "重": heavy, …; "要": will, wish, …; "角": cents, …; "色": outstanding, …

Figure 1. Example TransAhead responses to a source text under the translation (a) "we" and (b) "we play an important role". Note that the grammar/text predictions of (a) and (b) are not placed directly under the current input focus for space limit. (c) and (d) depict predominant grammar constructs which follow and (e) summarizes the translations for the source's character-based ngrams.

time updating the system's segmentation /translation options through the user's word choices. Our intended users are different from those of the previous research focusing on what professional translator can bring for MT systems (e.g., Brown and Nirenburg, 1990).

More recently, interactive MT (IMT) systems have begun to shift the user's role from analyses of the source text to the formation of the target translation. TransType project (Foster et al., 2002) describes such pioneering system that supports next word predictions. Koehn (2009) develops caitra which displays one phrase translation at a time and offers alternative translation options. Both systems are similar in spirit to our work. The main difference is that we do not expect the user to be a professional translator and we provide translation hints along with grammar predictions to avoid the generalization issue facing phrase-based system.

Recent work has been done on using fully-fledged statistical MT systems to produce target hypotheses completing user-validated translation prefix in IMT paradigm. Barrachina et al. (2008) investigate the applicability of different MT kernels within IMT framework. Nepveu et al. (2004) and Ortiz-Martinez et al. (2011) further exploit user feedbacks for better IMT systems and user experience. Instead of trigged by user correction, our method is triggered by word delimiter and assists in target language learning.

In contrast to the previous CAT research, we present a writing assistant that suggests subsequent grammar constructs with translations and interactively collaborates with learners, in view of reducing users' burden on grammar and word choice and enhancing their writing quality.

3. The TransAhead System

3.1 Problem Statement

For CAT and CALL, we focus on predicting a set of grammar patterns with lexical translations likely to follow the current target translation given a source text. The predictions will be examined by a human user directly. Not to overwhelm the user, our goal is to return a reasonable-sized set of predictions that contain suitable word choices and correct grammar to choose and learn from. Formally speaking,

Problem Statement: We are given a target-language reference corpus C_t, a parallel corpus C_{st}, a source-language text S, and its target translation prefix T_p. Our goal is to provide a set of predictions based on C_t and C_{st} likely to further translate S in terms of grammar and text. For this, we transform S and T_p into sets of ngrams such that the predominant grammar constructs with suitable translation options following T_p are likely to be acquired.

3.2 Learning to Find Pattern and Translation

We attempt to find syntax-based phraseology and translation equivalents beforehand (four-staged) so that a real-time system is achievable.

Firstly, we syntactically analyze the corpus C_t. In light of the phrases in grammar book (e.g., *one's* in "make up *one's* mind"), we resort to parts-of-speech for syntactic generalization. Secondly, we build up inverted files of the words in C_t for the next stage (i.e., pattern grammar generation). Apart from sentence and position information, a word's lemma and part-of-speech (POS) are also recorded.

We then leverage the procedure in Figure 2 to generate grammar patterns for any given sequence of words (e.g., contiguous or not).

```
procedure PatternFinding($query$,$N$,$C_t$)
(1)  $interInvList$=findInvertedFile($w_1$ of $query$)
     for each word $w_i$ in $query$ except for $w_1$
(2)     $InvList$=findInvertedFile($w_i$)
(3a)    $newInterInvList= \phi$ ; $i$=1; $j$=1
(3b)    while $i$<=length($interInvList$) and $j$<=lengh($InvList$)
(3c)       if $interInvList[i]$.SentNo==$InvList[j]$.SentNo
(3d)          Insert($newInterInvList$, $interInvList[i]$,$InvList[j]$)
           else
(3e)          Move $i$,$j$ accordingly
(3f)    $interInvList$=$newInterInvList$
(4) $Usage= \phi$
    for each $element$ in $interInvList$
(5)     $Usage$+={PatternGrammarGeneration($element$,$C_t$)}
(6) Sort patterns in $Usage$ in descending order of frequency
(7) return the $N$ patterns in $Usage$ with highest frequency
```

Figure 2. Automatically generating pattern grammar.

The algorithm first identifies the sentences containing the given sequence of words, *query*. Iteratively, Step (3) performs an AND operation on the inverted file, *InvList*, of the current word w_i and *interInvList*, a previous intersected results.

Afterwards, we analyze *query*'s syntax-based phraseology (Step (5)). For each *element* of the form ([wordPosi(w_1),…,wordPosi(w_n)], *sentence number*) denoting the positions of *query*'s words in the *sentence*, we generate grammar pattern involving replacing words with POS tags and words in wordPosi(w_i) with lemmas, and extracting fixed-window[3] segments surrounding *query* from the transformed sentence. The result is a set of grammatical, contextual patterns.

The procedure finally returns top *N* predominant syntactic patterns associated with the query. Such patterns characterizing the query's word usages follow the notion of pattern grammar in (Hunston and Francis, 2000) and are collected across the target language.

In the fourth and final stage, we exploit C_{st} for bilingual phrase acquisition, rather than a manual dictionary, to achieve better translation coverage and variety. We obtain phrase pairs through leveraging IBM models to word-align the bitexts, "smoothing" the directional word alignments via grow-diagonal-final, and extracting translation equivalents using (Koehn et al., 2003).

3.3 Run-Time Grammar and Text Prediction

Once translation equivalents and phraseological tendencies are learned, TransAhead then predicts/suggests the following grammar and text of a translation prefix given the source text using the procedure in Figure 3.

We first slice the source text *S* and its translation prefix T_p into character-level and word-level ngrams respectively. Step (3) and (4) retrieve the translations and patterns learned from Section 3.2. Step (3) acquires the active target-language vocabulary that may be used to translate the source text. To alleviate the word boundary issue in MT raised by Ma et al. (2007), TransAhead non-deterministically segments the source text using character ngrams and proceeds with collaborations with the user to obtain the segmentation for MT and to complete the translation. Note that a user vocabulary of preference (due to users' domain of knowledge or errors of the system) may be exploited for better system performance. On the other hand, Step (4) extracts patterns preceding with the history ngrams of $\{t_j\}$.

```
procedure MakePrediction($S$,$T_p$)
(1) Assign sliceNgram($S$) to $\{s_i\}$
(2) Assign sliceNgram($T_p$) to $\{t_j\}$
(3) $TransOptions$=findTranslation($\{s_i\}$,$T_p$)
(4) $GramOptions$=findPattern($\{t_j\}$)
(5) Evaluate translation options in $TransOptions$
        and incorporate them into $GramOptions$
(6) Return $GramOptions$
```

Figure 3. Predicting pattern grammar and translations.

In Step (5), we first evaluate and rank the translation candidates using linear combination:

$$\lambda_1 \times \left(P_1\left(t \mid s_i \right) + P_1\left(s_i \mid t \right) \right) + \lambda_2 \times P_2\left(t \mid T_p \right)$$

where λ_i is combination weight, P_1 and P_2 are translation and language model respectively, and *t* is one of the translation candidates under *S* and T_p. Subsequently, we incorporate the lemmatized translation candidates into grammar constituents in *GramOptions*. For example, we would include "close" in pattern "*play role* IN[*in*] VBG" as "*play role* IN[*in*] VBG[*close*]".

At last, the algorithm returns the representative grammar patterns with confident translations expected to follow the ongoing translation and further translate the source. This algorithm will be triggered by word delimiter to provide an interactive environment where CAT and CALL meet.

4. Preliminary Results

To train TransAhead, we used British National Corpus and Hong Kong Parallel Text and deployed GENIA tagger for POS analyses.

To evaluate TransAhead in CAT and CALL, we introduced it to a class of 34 (Chinese) first-year college students learning English as foreign language. Designed to be intuitive to the general public, esp. language learners, presentational tutorial lasted only for a minute. After the tutorial, the participants were asked to translate 15

[3] Inspired by (Gamon and Leacock, 2010).

Chinese texts from (Huang et al., 2011a) one by one (half with TransAhead assistance, and the other without). Encouragingly, the experimental group (i.e., with the help of our system) achieved *much* better translation quality than the control group in BLEU (Papineni et al., 2002) (i.e., 35.49 vs. 26.46) and *significantly* reduced the performance gap between language learners and automatic decoder of Google Translate (44.82). We noticed that, for the source “我們在結束這個交易上扮演重要角色”, 90% of the participants in the experimental group finished with more grammatical and fluent translations (see Figure 4) than (less interactive) Google Translate (“We conclude this transaction plays an important role”). In comparison, 50% of the translations of *the* source from the control group were erroneous.

1. we play(ed) a critical role in closing/sealing this/the deal.
2. we play(ed) an important role in ending/closing this/the deal.

Figure 4. Example translations with TransAhead assistance.

Post-experiment surveys indicate that a) the participants found TransAhead intuitive enough to collaborate with in writing/translation; b) the participants found TransAhead suggestions satisfying, accepted, and learned from them; c) interactivity made translation and language learning more fun and the participants found TransAhead very recommendable and would like to use the system again in future translation tasks.

5. Future Work and Summary

Many avenues exist for future research and improvement. For example, in the linear combination, the patterns’ frequencies could be considered and the feature weight could be better tuned. Furthermore, interesting directions to explore include leveraging user input such as (Nepveu et al., 2004) and (Ortiz-Martinez et al., 2010) and serially combining a grammar checker (Huang et al., 2011b). Yet another direction would be to investigate the possibility of using human-computer collaborated translation pairs to re-train word boundaries suitable for MT.

In summary, we have introduced a method for learning to offer grammar and text predictions expected to assist the user in translation and writing (or even language learning). We have implemented and evaluated the method. The preliminary results are encouragingly promising, prompting us to further qualitatively and quantitatively evaluate our system in the near future (i.e., learners’ productivity, typing speed and keystroke ratios of “del” and “backspace” (possibly hesitating on the grammar and lexical choices), and human-computer interaction, among others).

Acknowledgement

This study is conducted under the “Project Digital Convergence Service Open Platform” of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomas, E. Vidal, and J.-M. Vilar. 2008. Statistical approaches to computer-assisted translation. *Computer Linguistics*, 35(1): 3-28.

R. D. Brown and S. Nirenburg. 1990. Human-computer interaction for semantic disambiguation. In *Proceedings of COLING*, pages 42-47.

G. Foster, P. Langlais, E. Macklovitch, and G. Lapalme. 2002. TransType: text prediction for translators. In *Proceedings of ACL Demonstrations*, pages 93-94.

M. Gamon and C. Leacock. 2010. Search right and thou shalt find … using web queries for learner error detection. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37-44.

C.-C. Huang, M.-H. Chen, S.-T. Huang, H.-C. Liou, and J. S. Chang. 2011a. GRASP: grammar- and syntax-based pattern-finder in CALL. In *Proceedings of ACL*.

C.-C. Huang, M.-H. Chen, S.-T. Huang, and J. S. Chang. 2011b. EdIt: a broad-coverage grammar checker using pattern grammar. In *Proceedings of ACL*.

S. Hunston and G. Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

P. Koehn. 2009. A web-based interactive computer aided translation tool. In *Proceedings of ACL*.

Y. Ma, N. Stroppa, and A. Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of ACL*.

L. Nepveu, G. Lapalme, P. Langlais, and G. Foster. 2004. Adaptive language and translation models for interactive machine translation. In *Proceedings of EMNLP*.

Franz Josef Och and Hermann Ney. 2003. A systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

D. Ortiz-Martinez, L. A. Leiva, V. Alabau, I. Garcia-Varea, and F. Casacuberta. 2011. An interactive machine translation system with online learning. In *Proceedings of ACL System Demonstrations*, pages 68-73.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311-318.