	A			B			C			D			total		
	time (s)	keystrokes	quality (BLEU)	time (s)	keystrokes	quality (BLEU)	time (s)	keystrokes	quality (BLEU)	time (s)	keystrokes	quality (BLEU)	time (s)	keystrokes	quality (BLEU)
Google	114.68	209.83	68.17	110.67	236.78	72.25	80.39	168.65	75.96	100.30	184.30	71.57	102.38	204.26	72.12
CoCat	89.61** (21.86%↓)	138.41** (34.04%↓)	76.31** (8.14↑)	98.05** (11.40%↓)	168.13** (28.99%↓)	80.42** (8.17↑)	68.05** (15.35%↓)	93.94** (44.30%↓)	86.06** (10.10↑)	71.56** (28.65%↓)	124.33** (32.50%↓)	82.84** (11.27↑)	**84.03** (17.89%↓)**	**134.85** (33.98%↓)**	**81.29** (9.17↑)**
PE+Google	64.7	100.66	78.49	52.93	92.24	80.74	83.25	158.13	77.02	71.78	121.81	77.72	66.59	115.75	78.79
PE+CoCat	52.03** (19.58%↓)	59.36** (41.03%↓)	81.53** (3.04↑)	48.34** (8.68%↓)	63.44** (31.22%↓)	85.32** (4.58↑)	65.43** (21.41%↓)	80.77** (48.92%↓)	84.05** (7.03↑)	66.90** (6.80%↓)	82.11** (32.60%↓)	72.76 (4.96↓)	**56.63** (14.97%↓)**	**69.865** (39.64%↓)**	**81.98** (3.19↑)**

Table 3: Translation time, keystrokes and translation quality. The numbers in parentheses represent the improvement over the corresponding previous line. Individual results vary. "**" means the scores are significantly better than the corresponding previous line with $p < 0.01$.

with CoCat input method ("PE+CoCat"). Naturally, for each human translator, he/she should translate different sentences when using different assistant tools. Thus, we splitted the test data into four subsets randomly and evenly. Table 1 shows the details about the statistics of the 4 groups of test subset data. Table 2 shows the details about the permutation of assignments inspired by the previous works (Koehn, 2009a; Green et al., 2014).

In the real world, there are many factors which may influence our experimental results, such as the different difficulties of the sentences to be translated, the tolerance of the long period of translation test and different levels of translators. To eliminate the irrelevant effects, we use the permutation of assignments in Table 2 based on the following assumptions: (1) the minor discrepancy of difficulty degrees of four test subsets can be negligible; (2) the fatigue degree difference of a particular translator in different time in one day can be negligible.

3.2 Data Cleaning

To exclude the translation irrelevant factors, such as the time spent on searching for terms and the moments of rest, we process the user interaction log as follows:

(1) Remove all the interactions which are irrelevant to the assistant tools from the timeline, such as looking up the dictionary online and searching information online.

(2) Exclude all of the time intervals lasting longer than 10 seconds between two adjacent interactions.

(3) Select the best four from the 12 human translations as references for each source sentence, and average the scores of human translations using BLEU-4 evaluation metric (Papineni et al., 2002).

3.3 Results and Analysis

We analyze the human productivity in terms of translation time, keystrokes and translation quality. To improve the robustness, we average the result values of repeated measurements. Let's take the "translation time" for example. According to the permutation of assignments in Table 2, the sentence s_i in subset M_1 has been translated by three translators in group A under the assistance of "Google". For the instance s_i, we average the three values of "translation time" given by the system and get the value $time_{s_i}^{Google}$. We compute the average translation time of a subset under the assistance of "Google" as follows:

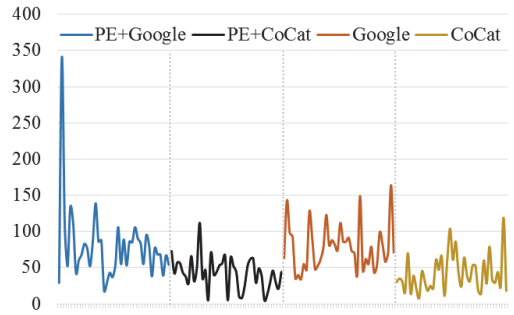

Figure 4: One translator's records on translation time. The graph plots the time spent on translation (in seconds, y-axis) against the sentence ID (x-axis).

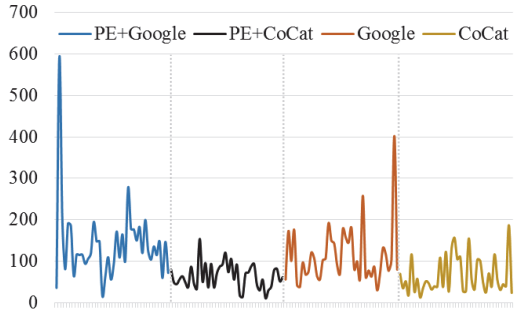

Figure 5: One translator's records on keystrokes. The graph plots the number of keystrokes spent on translation (y-axis) against the sentence ID (x-axis).

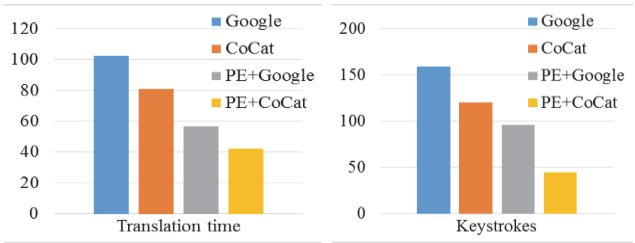

Figure 6: The comparisons of translation time and keystrokes of the four assistances applied to the sentence "CPC's discipline agency announced on Jan. 16 that Huo has been placed under investigation for suspected serious violation of party disciplines and laws".

$$time_{M_j}^{Google} = \frac{\sum_{s_i \in M_j} time_{s_i}^{Google}}{|M_j|}, j = 1,2,3,4$$

Then we calculate the average translation time of all sentences under the assistance of "Google" using the following formula:

$$time^{Google} = \frac{\sum_{i=1}^{160} time_{s_i}^{google}}{160}.$$

For keystrokes and translation quality, they are calculated in the same way.

For example, translation time and keystroke consumption on each sentence of a specific translator in group C are reported in Figure 4 and Figure 5. As we can see in the figures, CoCat helps her save about 46% time and about 41% keystrokes in the scratch mode, and save about 45% time and about 54% keystrokes in the post-editing mode.

The detailed results of all the human translators are reported in Table 3. On average, all human translators are faster and also achieve better translation quality using any of types of assistance offered. What's more, human translators are faster and also achieve better translation quality using CoCat (translating from scratch or post-editing).

For translation time and keystrokes, the figures in Table 3 show that our proposed CoCat always helps human translators significantly (with $p < 0.01$), saving more than 14% time and over 33% keystrokes compared with the strong baseline, i.e., post-editing using Google Pinyin (line 4 vs. line 3 and line 6 vs. line 5).

For translation quality, the figures in Table 3 show that CoCat can help human translators improve the translation quality significantly as well (with $p < 0.01$) by more than 3 absolute BLEU scores over the strong baseline.

Take a specific sentence as an example, such as "CPC's discipline agency announced on Jan. 16 that Huo has been placed under investigation for suspected serious violation of party discip-lines and laws", the comparison statistics of translation time and keystrokes are reported in Figure 6. CoCat can save about 21% time and about 24% keystrokes in the scratch mode, and save about 26% time and about 53% keystrokes in the post-editing mode.

Overall, the results in Table 3 indicate that post-editing consistently outperform unassisted translation. It is in line with the findings reported by Koehn (2012). Meanwhile, the post-editing well integrated with our proposed CoCat input method further improves the translation productivity.

What's more, if we focus on the comparison between "CoCat" and "PE+Google", we can find that the difference of the translation quality is very small. In the industrial world, the poor performance of the automatic translation engine is often a headache for human translators to edit the MT results. The comparison between "CoCat" and "PE+Google" tells us that we can make human translators generate better translation in less time with the aid of MT without headache.

In summary, we can draw the conclusion that the proposed new input method makes it easier for human translators to interact with MT systems effectively and imperceptibly.

4 Related Work

The goal of this paper is to improve the productivity and efficiency of human translators by fully exploiting the MT technology. The core idea is to provide human translators translation candidates effectively and friendly. There are two kinds of related work focusing on offering translation suggestions.

Koehn (2009a; Koehn et al., 2014) developed the tool Caitra which aims at providing translation suggestions to complete the target language sentence. Based on MT post-editing, their method can offer word and phrase translation candidates through interative machine translation. Green et al. (2014) made extensive modifications for the MT system and designed a new CAT interface. Their methods are tightly coupled with statistical machine translation in which only left-to-right decoding is allowed and dynamic decoding in interactive machine translation is usually time-consuming. In contrast, we integrate most of the useful knowledge of the MT system into a well designed CoCat input method that provides the translation suggestions more friendly and imperceptibly without forcing the human translators to take a view of the MT outputs. Besides using MT outputs, we are the first to exploit depth information used by MT, such as translation rules and decoding hypotheses.

Recently, Li (2012) and Fang (2013) also attempted to incorporate the SMT information into the Chinese Pinyin input method. In their approaches, when they developed their input methods, only the MT model scores and the fuzzy word alignment between the MT output and the human translation output are employed. However, there are two disadvantages in their approaches. On the one hand, the dynamic MT model scores are difficult to calculate and these model scores are not compatible with other features in input methods. On the other hand, the fuzzy word alignment contains much noise which would not benefit much to the input method. Instead, we design the log-linear model for the input method CoCat and integrate the translation rules, decoding hypotheses and the n-best translation list of the MT system. In addition, we propose the n-gram prediction model to further improve the efficiency of human translators.

5 Conclusion

In this paper, we have presented a novel input method CoCat which deeply integrates MT into CAT effectively and imperceptibly. This well-designed input method is modeled with a log-linear framework, and takes as features most of the useful knowledge of the MT system, such as translation rules, decoding hypotheses and n-best translation lists. Furthermore, we have proposed an n-gram prediction model that further speeds up the translation typing process.

The human translation experiments on English-to-Chinese have shown that the proposed approach can not only help human translators significantly save the time and keystrokes, but also substantially improve the final translation quality. The experiments have also shown that post-editing well integrated with our proposed approach further improves the translation productivity.

Acknowledgement

We thank anonymous reviewers for their valuable comments. The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61333018 and No. 61403379 and supported by the West Light Foundation of Chinese Academy of Sciences under Grant No. LHXZ201301.

References

[Barrachina *et al*., 2009] Statistical Approaches to Computer-Assisted Translation. *Computational Linguistics*, 35(1): 3-28, 2009.

[Carl *et al*., 2011] The Process of Post-editing: A Pilot Study. *Copenhagen Studies in Language*, 41:131-142, 2011.

[Fang, 2013] Research and Implementation on Aided Translation Tools Based on Input Method. *Xiamen University*. 2013.

[Foster and Lapalme, 2002] Text Prediction for Translators. *Université de Montréal*, *2002*.

[Garay-Vitoria and Abascal, 2006] Text Prediction Systems: a Survey. *Universal Access in the Information Society*, 4(3): 188-203, 2006.

[Green *et al*., 2014] Human Effort and Machine Learnability in Computer Aided Translation. In *Proceedings of the EMNLP 2014*.

[Li, 2012] A Pinyin Input Method Editor with English-Chinese Aided Translation Function. In *2012 International Conference on Computer Science and Services System*.

[Kasami, 1965] An Efficient Recognition and Syntax Analysis Algorithm for Context-free Languages. *Technical Report AFCRL-65-758*. Air Force Cambridge Research Laboratory, Bedford, MA, 1965.

[Koehn, 2004] Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the EMNLP 2004.*

[Koehn, 2009a] A Process Study of Computer-Aided Transltion. *Machine Translation Journal*, 23(4):241-263, 2009.

[Koehn, 2009b] A Web-Based Interactive Computer Aided Translation Tool. In *Proceedings of the ACL-IJCNLP 2009*.

[Koehn, 2012] Computer-added Trasnlation. Machine Translation Marathon. 2012.

[Koehn *et al*., 2014] Refinements to Interactive Translation Prediction Based on Search Graphs. In *Proceedings of the ACL 2014*.

[Papineni *et al*., 2002] BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL 2002*.

[Snover *et al*., 2006] A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas 2006*.

[Xiong *et al*., 2006] Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *proceedings of COLING-ACL 2006*.

[Younger, 1967] Recognition and Parsing of Context-free Languages in Time n^3. *Information and Control*, 10(2):189-208, 1967.

[Zaidan, 2009] Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*. 91:79-88, 2009.

[Zhechev *et al*., 2012] Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice*.