Chinese Character Decomposition for Neural MT with Multi-Word Expressions

Lifeng Han[1], **Gareth J. F. Jones**[1], **Alan F. Smeaton**[2] and **Paolo Bolzoni**
[1] ADAPT Research Centre
[2] Insight Centre for Data Analytics
School of Computing, Dublin City University, Dublin, Ireland
lifeng.han@adaptcentre.ie, paolo.bolzoni.brown@gmail.com

Abstract

Chinese character decomposition has been used as a feature to enhance Machine Translation (MT) models, combining radicals into character and word level models. Recent work has investigated ideograph or stroke level embedding. However, questions remain about different decomposition levels of Chinese character representations, radical and strokes, best suited for MT. To investigate the impact of Chinese decomposition embedding in detail, i.e., radical, stroke, and intermediate levels, and how well these decompositions represent the meaning of the original character sequences, we carry out analysis with both automated and human evaluation of MT. Furthermore, we investigate if the combination of decomposed Multiword Expressions (MWEs) can enhance the model learning. MWE integration into MT has seen more than a decade of exploration. However, decomposed MWEs has not previously been explored.

1 Introduction

Despite Neural Machine Translation (NMT) (Cho et al., 2014; Johnson et al., 2016; Vaswani et al., 2017; Lample and Conneau, 2019) having recently replaced Statistical Machine Translation (SMT) (Brown et al., 1993; Och and Ney, 2003; Chiang, 2005; Koehn, 2010) as the state-of-the-art, research questions still remain, such as how to deal with *out-of-vocabulary* (OOV) words, how best to integrate *linguistic knowledge* and how best to correctly translate *multi-word expressions* (MWEs) (Sag et al., 2002; Moreau et al., 2018; Han et al., 2020a). For OOV word translation for European languages, substantial improvements have been made in terms of rare and unseen words by incorporating sub-word knowledge using Byte Pair Encoding (BPE) (Sennrich et al., 2016). However, such methods cannot be directly applied to Chinese, Japanese and other ideographic languages.

Integrating sub-character level information, such as Chinese ideograph and radicals as learning knowledge has been used to enhance features in NMT systems (Han and Kuang, 2018; Zhang and Matsumoto, 2018; Zhang and Komachi, 2018). Han and Kuang (2018), for example, explain that the meaning of some unseen or low frequency Chinese characters can be estimated and translated using *radicals* decomposed from the Chinese characters, as long as the learning model can acquire knowledge of these radicals within the training corpus.

Chinese characters often include two pieces of information, with *semantics* encoded within radicals and a *phonetic* part. The phonetic part is related to the pronunciation of the overall character, either the same or similar. For instance, Chinese characters with this two-stroke radical, 刂 (tí dāo páng), ordinarily relate to *knife* in meaning, such as the Chinese character 劍 (jiàn, *sword*) and multi-character expression 鋒利 (fēnglì, *sharp*). The radical 刂 (tí dāo páng) preserves the meaning of knife because it is a variation of a drawing of a knife evolving from the original bronze inscription (Fig. 4 in Appendices).

Not only can the radical part of a character be decomposed into smaller fragments of strokes but the phonetic part can also be decomposed. Thus there are often several levels of decomposition that can be applied to Chinese characters by combining different levels of decomposition of each part of the Chinese character. As one example, Figure 1 shows the three decomposition levels from our model and the full stroke form of the above mentioned characters 劍(jiàn) and 鋒(fēng) . To date, little work has been carried out to investigate the full potential of these alternative levels of decomposition of Chinese characters for the purpose of Machine Translation (MT).

In this work, we investigate Chinese character decomposition, and additionally we investigate another area related to Chinese characters, namely Chinese MWEs. We firstly investigate translation at increasing levels of decomposition of Chinese characters using underlying radicals, as well as the additional Chinese character strokes (corresponding to ever-smaller units), breaking down characters into component parts as this is likely to reduce the number of unknown words. Then, in order to better deal with MWEs which have a common occurrence in the general context (Sag et al., 2002), and working in the opposing direction in terms of meaning representation, we investigate translating larger units of Chinese text, with the aim of restricting translation of larger groups of Chinese characters that should be translated together as one unit. In addition to investigating the effects of decomposing characters we simultaneously apply methods of incorporating MWEs into translation. MWEs can appear in Chinese in a range of ways, such as fixed (or semi-fixed) expressions, metaphor, idiomatic phrases, and institutional, personal or location names, amongst others.

In summary, in this paper, we investigate (i) the degree to which Chinese radical and stroke sequences represent the original word and character sequences that they are composed of; (ii) the difference in performance achieved by each decomposition level; (iii) the effect of radical and stroke representations in MWEs for MT. Furthermore, we offer (available at radical4mt[1]):

- an open-source suite of Chinese character decomposition extraction tools;
- a Chinese ⇔ English MWE corpus where Chinese characters have been decomposed

The rest of this paper is organized as follows: Section 2 provides some related work in character and radical related MT; Section 3 and 4 introduce our Chinese decomposition procedure into radical and strokes, and the experimental design; Section 5 provides our evaluations from both automatic and human perspectives; Section 6 includes conclusions and plans for future work.

[1] `https://github.com/poethan/MWE4MT`

2 Related Work

Chinese character decomposition has been explored recently for MT. For instance, Han and Kuang (2018) and Zhang and Matsumoto (2018), considered radical embeddings as additional features for Chinese → English and Japanese ⇔ Chinese NMT. In Han and Kuang (2018), a range of encoding models including word+character, word+radical, and word+character+radical were tested. The final setting with word+character+radical achieved the best performance on a standard NIST [2] MT evaluation data set for Chinese → English. Furthermore, Zhang and Matsumoto (2018) applied radical embeddings as additional features to character level LSTM-based NMT on Japanese → Chinese translation. None of the aforementioned work has however investigated the performance of decomposed character sequences and the effects of varied decomposition degrees in combination with MWEs. Subsequently, Zhang and Komachi (2018) developed bidirectional English ⇔ Japanese, English ⇔ Chinese and Chinese ⇔ Japanese NMT with word, character, ideograph (the phonetics and semantics parts of characters are separated) and stroke levels, with experiments showing that the *ideograph* level was best for ZH→EN MT, while the stroke level was best for JP→EN MT. Although their ideograph and stroke level setting replaced the original character and word sequences, there was no investigation of *intermediate decomposition* performance, and they only used BLEU score as the automated evaluation with no human assessment involved. This gives us inspiration to explore the performance of intermediate level embedding between ideograph and strokes for the MT task.

3 Chinese Character Decomposition

We introduce the character decomposition approach and the extraction tools which we apply in this work (code will be publicly available). We utilize the open source IDS dictionary [3] which was derived from the CHISE (CHaracter Information Service Environment) project[4]. It is comprised of 88,940 Chinese characters from CJK (Chinese, Japanese, Korean script) Unified Ideographs

[2] `https://www.nist.gov/programs-projects/machine-translation`
[3] `https://github.com/cjkvi/cjkvi-ids`
[4] `http://www.chise.org/`

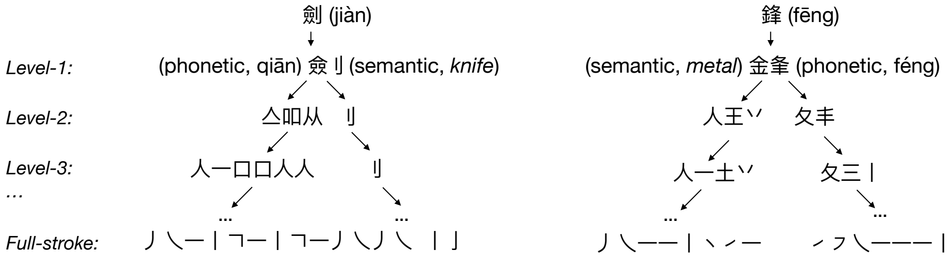

Figure 1: Examples of the decomposition of Chinese characters.

and the corresponding decomposition sequences of each character. Most characters are decomposed as a single sequence, but characters can have up to four possible decomposed representations. The reason for this is that the character can come from different resources, such as Chinese Hanzi (G, H, T for Mainland, Hong Kong, and Taiwan), Japanese Kanji (J), Korean Hanja (K), and Vietnamese ChuNom (V), etc.[5] Even though they have the same root of Hanzi, the historical development of languages and writing systems in different territories has resulted in their certain degree of variations in the appearance and stroke order, for instances, (且, qiě) vs (目, mù) from the second character example in Figure 2.

Character	Decomposition	Decomposition
丽 (lì)	⿱一⿰⿵冂丶⿵冂丶[G]	⿰⿱一⿵冂丶⿱一⿵冂丶[T]
具 (jù)	⿱⿴且一八[GTKV]	⿳目一八[J]
函 (hán)	⿶凵⿻了⿱丷八[GTV]	⿶凵⿻丂⿱丷八[JK]
勇 (yǒng)	⿱甬力[GTV]	⿱⿱龴田力[JK]
Character construction: ⿱: up-down, ⿰: left-right, ⿵⿶⿴: inside-outside, ⿻: embedded		

Figure 2: Character examples from IDS dictionary; the grey parts of decomposition graphs represent the construction structure of the character.

Figure 2 shows example characters that have two different decomposition sequences. In our experiments, when there is more than one decomposed representation of a given character, we choose the Chinese mainland decomposition standard (G) for the model, since the corpora we use correspond best to simplified Chinese as used in mainland China. The examples in Figure 2 also show the general construction and corresponding decomposition styles of Chinese characters, such as *left-right*, *up-down*, *inside-outside*, and *embedded* amongst others. To obtain a decomposition level L representation of Chinese character α, we go through the IDS file L times. Each time, we search the IDS file character list to match the newly generated smaller sized characters and replace them with decomposed representation recursively.

4 NMT Experiments

We test the various levels of decomposed Chinese and Chinese MWEs using publicly available data from the WMT-2018 shared tasks Chinese to English, using the preprocessed (word segmented) data as training data (Bojar et al., 2018). The original word boundaries were preserved in decomposition sequences. To get better generalizability of our decomposition model, we used a large size, the first 5 million parallel sentences for training across all learning steps. The corpora "newsdev2017" used for development and "newstest2017" for testing are from WMT-2017 MT shared task (Bojar et al., 2017). These include 2002 and 2001 parallel Chinese ⇔ English respectively. We use the THUMT (Zhang et al., 2017) toolkit which is an implementation of several attention-based Transformer architectures (Vaswani et al., 2017) for NMT and set up the encoder-decoder as 7+7 layers. Batch size is set as 6250. For sub-word encoding BPE technology, we use *32K BPE operations* that are learned from the bilingual training set. We use Google's Colab platform to run our experiments[6]. We name the baseline model using character sequences (with word boundary) as *character sequence model*. For MWE integrated models, we apply the same bilingual MWE extraction pipeline from our work (Han et al., 2020b), similar to (Rikters and Bojar, 2017), which is an

[5] Universal Coded Character Set (10646:2017) `standards.iso.org/ittf/PubliclyAvailableStandards`

[6] `https://colab.research.google.com`

	20k	100k	120k	140k	160k	180k
baseline	18.39	**21.56**	21.45	21.31	21.29	21.42
base+MWE	**18.49**	21.39	**21.67**	**21.83**	**21.42**	**21.86**
RXD3	16.48	20.75	20.73	20.93	20.98	21.14
RXD3+MWE	17.82	21.36	21.50	21.31	**21.42**	21.47
RXD2	11.84	13.26	12.88	13.02	13.38	12.86
RXD1/ideograph	15.52	20.67	20.61	21.26	20.76	21.00

Figure 3: Chinese→English BLEU scores for increasing learning steps; RXD1/2/3 represents the decomposition level of Chinese characters. RXD1 indicates *ideograph* from (Zhang and Komachi, 2018)

automated pre-defined PoS pattern-based extraction procedure with filtering threshold set to 0.85 to remove lower quality translation pairs. We integrate these extracted bilingual MWEs back into the training set to investigate if they help the MT learning. In the decomposed models, we replace the original Chinese character sequences from the corpus with decomposed character-piece sequence inputs for training, development and testing (with original word boundary kept).

5 Evaluation

In order to assess the performance of each model employing a different meaning representation in terms of decomposition and MWEs, we carried out both automatic, BLEU (Papineni et al., 2002) in Fig. 3, and human evaluation (Direct Assessment) of the outputs of the system. Since decomposition level 3 yields generally higher scores than the other two levels, we also applied decomposition of MWEs to level 3 and concatenated the bilingual glossaries to the training.

We used the models with the most learning steps, 180K, and run human evaluation on the Amazon Mechanical Turk crowd-sourcing platform,[7] including the strict quality control measures of Graham et al. (2016). Direct Assessment scores for systems were calculated as in Graham et al. (2019) by firstly computing an average score per translation before calculating the overall average for a system from its average scores for translations. Significance tests in the form of Wilcoxon Rank-Sum test are then applied to score distributions of the latter to identify systems that significantly outperform other systems in the human evaluation.

Results of the Direct Assessment human evaluation are shown in Table 1 where similarly performing systems are clustered together (denoted by horizontal lines in the table). Systems in a given lower ranked cluster are significantly outperformed by all systems in a higher ranked cluster. Amongst the six models included in the human evaluation, the first five form a cluster with very similar performance according to human assessors, including the baseline, MWE, RXD1, RXD3MWE, and RXD3 which do not outperform each other with any significance. RXD2, on the other hand, is far behind the other models in terms of performance according to human judges (also the automated BLEU score) performing significantly worse than all other runs (at $p < 0.05$). As the tradition of WMT shared task workshop, we cluster the first five models into one group, while the RXD2 into a second group. Furthermore, human evaluation results in Table 1 show that the top five models all achieve high performance on-par with state-of-the-art in Chinese to English MT.

We also discovered that the decomposed models generated fewer system parameters for the neural nets to learn, which potentially reduces computational complexity. For instance, the total trainable variable size of the character sequence baseline model is 89,456,896, while this number decreased to 80,288,000 and 80,591,104 respectively for the RXD3 and RXD2 models (a 10.25% drop for RXD3). As mentioned by Goodfellow et al. (2016), in NLP tasks the total number of possible words is so large that the word sequence models have to operate on an extremely high-dimensional and sparse discrete space. The decomposition model reduced the overall size of possible tokens for the model to learn, which is more space efficient.

For the automatic and human evaluation results, where the decomposition level 2 achieved surprisingly lower score than the other levels, error analysis revealed an important insight. While level-1 decomposition encoded the original character sequences into radical representations, and this typically contains semantic and phonetic parts of the character, level-3 gives a deeper decomposition of the character such as the stroke level pieces with sequence order. In contrast, however, level-2 decomposition appears to introduce some intermediate characters that mislead model learning. These intermediate level characters are usually constructed from fewer strokes than the orig-

[7] `https://www.mturk.com`

Ave. raw	Ave. z	n	N	
73.2	0.161	1,232	1,639	BASE
71.6	0.125	1,262	1,659	MWE
71.6	0.113	1,257	1,672	RXD1
71.3	0.109	1,214	1,593	RXD3MWE
70.2	0.073	1,260	1,626	RXD3
53.9	−0.533	1,227	1,620	RXD2

Table 1: Human evaluation results for systems using Direct Assessment, where Ave. raw = the average score for translations calculated from raw Direct Assessment scores for translations, Ave. z = the average score for translations after score standardization per human assessor mean and standard deviation score, n is the number of distinct translations included in the human evaluation (the sample size used in significance testing), N is the number of human assessments (including repeat assessment).

inal root character, but can be decomposed from it. As in Figure 1, from decomposition level-2, we get new characters 从 (cóng) and 王 (wáng) respectively from 劍 (Jiàn, *sword*) and 鋒 (fēng, *edge/sharp point*), but they have no direct meaning from their father characters, instead meaning "from" and "king" respectively. In summary, decomposition level-2 tends to generate some intermediate characters that do not preserve the meaning of the original root character's radical, nor those of the strokes, but rather smaller sized independent characters with fewer strokes that result in other meanings.

6 Conclusions and Future Work

In this work, we tested the varying degrees of Chinese character decomposition and their effect on Chinese to English NMT with attention architecture. To the best of our knowledge, this is the first work on detailed decomposition level of Chinese characters for NMT, and decomposition representation for MWEs. We conducted experiments for decomposition levels 1 to 3; we had a look at level 4 decomposition and it appears similar to level 3 sequences. We publish our extraction toolkit free for academic usage. We conducted automated evaluation with the BLEU metric, and crowd sourced human evaluation with the direct assessment (DA) methodology. Our conclusion is that the Chinese character decomposition levels 1 and 3 can be used to represent or replace the original character sequence in an MT task, and that this achieves similar performance to the original character sequence model in our NMT setting. However, decomposition level 2 is not suitable to represent the original character sequence in meaning at least for MT. We leave it to future work to explore the performance of different decomposition levels in other NLP tasks.

Another finding from our experiments is that while adding bilingual MWE terms can both increase character and decomposed level MT score according to the automatic metric BLEU, the human evaluation shows no statistical significance between them. Significance testing using automated evaluation metrics will be carried out in our future work, such as METEOR (Banerjee and Lavie, 2005), and LEPOR (Han et al., 2012; Han, 2014), in addition to BLEU.

We will consider different MWE integration methods in future and reduce the training set to investigate the differences in low-resource scenarios (5 million sentence pairs for training set were used in this work). We will also sample a set of the testing results and conduct a human analysis regarding the MWE translation accuracy from different representation models. We will further investigate different strategies of *combining* several level of decompositions together and their corresponding performances in semantic representation, such as MT task. The IDS file we applied to this work limited the performance of full stroke level capability, and we will look for alternative methods to achieve full-stroke level character sequence extraction for NLP tasks investigation.

Acknowledgments

We thank Yvette Graham for helping with human evaluation, Eoin Brophy for helps with Colab, and thank the anonymous reviewers for their thorough reviews and insightful feedback. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The input of Alan Smeaton is part-funded by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight Centre).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Lifeng Han. 2014. *LEPOR: An Augmented Machine Translation Evaluation Metric*. University of Macau.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020a. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020b. MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2970–2979, Marseille, France. European Language Resources Association.

Lifeng Han and Shaohui Kuang. 2018. Incorporating chinese radicals into neural machine translation: Deeper than character level. In *Proceedings of ESSLLI-2018 Student Session*, pages 54–65. Association for Logic, Language and Information (FoLLI).

Lifeng Han, Derek F. Wong, and Lidia S. Chao. 2012. Lepor: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, page 441–450. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel, and Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 177 – 207. Language Science Press.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Matīss Rikters and Ondřej Bojar. 2017. Paying Attention to Multi-Word Expressions in Neural Machine Translation. In *Proceedings of the 16th Machine Translation Summit (MT Summit 2017)*, Nagoya, Japan.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing System*, pages 6000–6010.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. 2017. Thumt: An open source toolkit for neural machine translation. *ArXiv*, abs/1706.06415.

Jinyi Zhang and Tadahiro Matsumoto. 2018. Improving character-level japanese-chinese neural machine translation with radicals as an additional input feature. *CoRR*, abs/1805.02937.

Longtu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 17–25, Brussels, Belgium. Association for Computational Linguistics.

Appendices

Appendix A: Chinese Character Knowledge

Figure 4 demonstrates the meaning preservation root of Chinese radicals, where the evolution of the Chinese character 刀 (Dāo), meaning *knife*, evolved from bronze inscription form to contemporary character and radical form, 刂 (named as: tí dāo páng).

NMT for Asian languages has included translation at the level of phrase, word, and character sequences (see Figure 5).

Appendix B: More Details of Evaluation

The evaluation scores of character sequence baseline NMT, character decomposed NMT and MWE-NMT according to the BLEU metric are presented in Fig. 3. The RXD1 model, decomposition level 1, is the *ideograph* model Zhang and Komachi (2018) used for their experiments where the phonetics (声旁 shēng páng) and semantics (形旁 xíng páng) parts of character are separated initially.

From the automated evaluation results, we see that decomposition model RXD3 has very close BLEU scores to the baseline character sequence (both with word boundary) model. This is very interesting since the level 3 Chinese decomposition is typically impossible (or too difficult) for even native language human speakers to read and understand. Furthermore, by adding the decomposed MWEs back into the learning corpus, "rxd3+MWE" (RXD3MWE) yields higher BLEU scores in some learning steps than the baseline model. To gain further insight, we provide the learning curve with the learning steps and corresponding automated-scores in Figure 6.

The BLEU score increasing ratio in decomposed models (from RXD3 to RXD3MWE) is larger than the ratio in original character sequence models (from BASE to BASEMWE) by adding MWEs in general. Furthermore, the increase in performance is very consistent by adding MWEs from the decomposed model, compared to the conventional character sequence model. For instance, the performance has a surprisingly drop at 100K learning steps for BASEMWE.

Appendix C: Looking into MT Examples

From the learning curves in Fig. 6, we suggest that with 5 million training sentences and 7+7 layers of encoder-decoder neural nets, the Transformer model becomes too flat in its learning rate curve with 100K learning steps, and this applies to both original character sequence model and decomposition models.

In light of this, we look at the MT outputs from head sentences of testing file at 100K learning steps models, and provide some insight into errors made by each model. Even though the automated BLEU metric gives the baseline model a higher score 21.56 than the RXD3 model (20.75) the translation of some Chinese MWE terms is better with the RXD3 model. For instance, in Figure 7, the Chinese MWE 商场 (Shāngchǎng) in the first sentence is correctly translated as *mall* by RXD3 model but translated as *shop* by the baseline character sequence model; the MWE 楼梯间 (lóutījiān) in the second sentence is correctly translated as *stairwell* by the RXD3 model while translated as *stairs* by baseline. Furthermore, the MWE 近日 (Jìnrì) meaning *recently* is totally missed out by the original character sequence model, which results in a misleading am-

Chinese radical 刂 (Dāo, knife) evolution from Pictogram to Regular script					
商 Shang Dynasty (1600-1046BC)		西周 Western-Zhou Dynasty (1045-771BC)	戰國 Warring States period (476-221BC)	漢 Han Dynasty (202BC-220)	東漢 Eastern Han (from 57AD on)
Bronze inscriptions	Oracle bone script	Bronze Inscription	Silk	篆 (on Seal)	Regular script

Figure 4: Example Chinese radical, 刂 (Dāo), where the character evolved from leftmost pictogram to present day regular script (rightmost) containing only two strokes. The two strokes are called as 竪 (Shù, vertical) + 竪 (Shù gōu, vertical with hook). The corresponding character representation is 刀 (Dāo).

Word level	28 / 歲 / 廚師 / 被 / 發現 / 死 / 於 / 舊金山 / 一家 / 商場
Character	28 歲 廚 師 被 發 現 死 於 舊 金 山 一 家 商 場
Pronunciation	èr shí bā Suì chú shī bèi fā xiàn sǐ yú jiù jīn shān yī jiā shāng chǎng
Radical	28 止戌少 广尌 自帀 衤皮 癶弢 王見 歹匕 方仒 雈臼 人王丷 山 一 宀豕 亠丷冏 土昜
English Ref.	28-Year-Old Chef Found Dead at San Francisco Mall

Figure 5: Example of Chinese word to character level changes for MT. Pronunciation is Mandarin in Pinyin. The English reference here is taken from the corpus we used for our experiments.

biguous translation of an even larger content, i.e., did the chief moved to San Francisco (SF) *recently* or *this week*. We will not get this clearly from the character base sequence model, however, the MWE 近日 (Jìnrì) is correctly translated by the RXD3 model and the overall meaning of the sentence is clear that the chef moved to SF *recently* and was found dead *this week*.

We also attach the translations of these two sentences by four other models. With regard to the first sentence MWEs, all the four models translate San Francisco mall correctly as REF and RXD3 beating BASE model. In terms of the second sentence MWEs, BASEMWE and RXD2 drop out the MWE 近日 (Jìnrì, *recently*) as BASE model, and all the four models drop out the translation of MWE 楼梯间 (lóutījiān, *stairwell*).

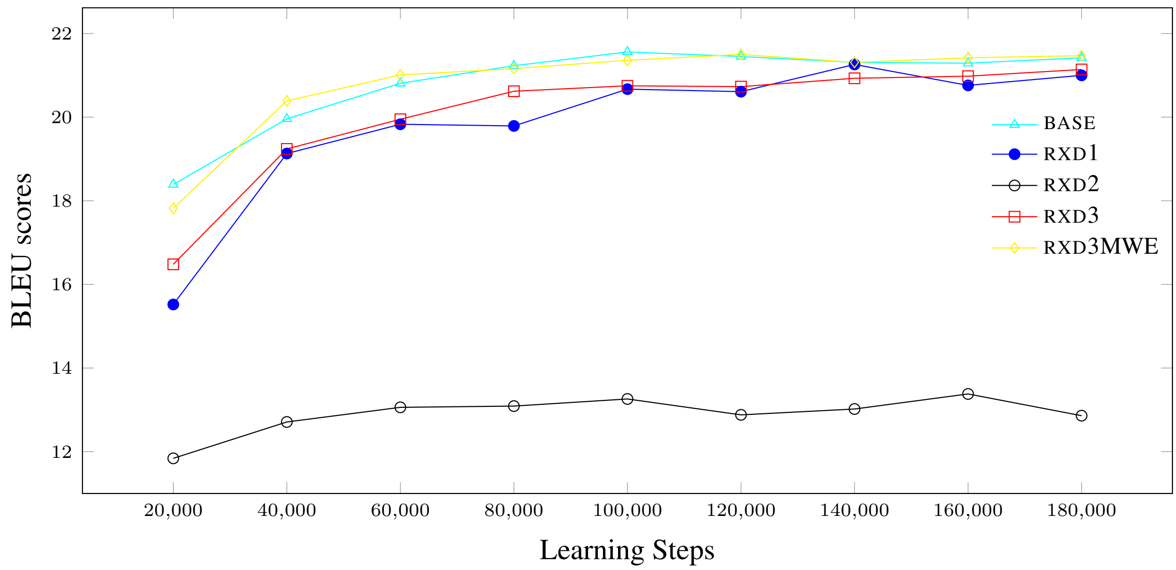

Figure 6: Learning curves from different models with BLEU metric

src	28 岁 厨师 被 发现 死 于 **旧金山** 一家 **商场** 近日 刚 搬 至 旧金山 的 一位 28 岁 厨师 本周 被 发现 死 于 当地 一家 **商场** 的 **楼梯间** 。
ref	28 @-@ Year @-@ Old Chef Found Dead at **San Francisco Mall** a 28 @-@ year @-@ old chef who had **recently** moved to San Francisco was found dead in the **stairwell** of a local **mall** this week .
rxd3	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who **recently** moved to San Francisco has been found dead on a **stairwell** in a local **mall** this week .
base	the 28 @-@ year @-@ old chef was found dead in a **shop** in **San Francisco** a 28 @-@ year @-@ old chef who has moved to San Francisco this week was found dead on the **stairs** of a local **mall** .
base MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef who **recently** moved to San Francisco was found dead this week at a local mall .
rxd3 MWE	28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef **recently** moved to San Francisco was found dead this week at a local mall .
rxd1	the 28 @-@ year @-@ old chef was found dead at a **San Francisco mall** a 28 @-@ year @-@ old chef **recently** moved to San Francisco was found dead in a local shopping mall this week .
rxd2	the 28 @-@ year @-@ old chef was found dead in a **San Francisco mall** a 28 @-@ year @-@ old San Francisco chef was found dead in a local mall this week .

Figure 7: Samples of the English MT output at 100K learning steps: RXD1, RXD2 and RXD3 are the Chinese decomposition with level 1 to 3, BASE is the character sequence model, BASEMWE and RXD3MWE are character sequence model with MWEs and decomposition level 3 model with decomposed MWEs, and src/ref represents source/reference.