# Recurrent Neural Network-based Tuple Sequence Model for Machine Translation

**Youzheng Wu, Taro Watanabe, Chiori Hori**
National Institute of Information and Communications Technology (NICT), Japan
erzhengcn@gmail.com
{taro.watanabe, chiori.hori}@nict.go.jp

## Abstract

In this paper, we propose a recurrent neural network-based tuple sequence model (RNNTSM) that can help phrase-based translation model overcome the phrasal independence assumption. Our RNNTSM can potentially capture arbitrary long contextual information during estimating probabilities of tuples in continuous space. It, however, has severe data sparsity problem due to the large tuple vocabulary coupled with the limited bilingual training data. To tackle this problem, we propose two improvements. The first is to factorize bilingual tuples of RNNTSM into source and target sides, we call factorized RNNTSM. The second is to decompose phrasal bilingual tuples to word bilingual tuples for providing fine-grained tuple model. Our extensive experimental results on the IWSLT2012 test sets[1] showed that the proposed approach essentially improved the translation quality over state-of-the-art phrase-based translation systems (baselines) and recurrent neural network language models (RNNLMs). Compared with the baselines, the BLEU scores on English-French and English-German tasks were greatly enhanced by 2.1%-2.6% and 1.8%-2.1%, respectively.

## 1 Introduction

The phrase-based translation systems (Koehn et al., 2003) rely on language model and lexicalized reordering model to capture lexical dependencies that span phrase boundaries. Their translation models, however, do not explicitly model context dependencies between translation units. To address this limitation, Marino et al. (2006) and Crego and Yvon (2010) proposed n-gram-based translation systems to capture dependencies across phrasal boundaries. The n-gram translation models have been shown to be effective in helping the phrase-based translation models overcome the phrasal independence assumption (Durrani et al., 2013; Zhang et al., 2013). Most of the n-gram translation models (Marino et al., 2006; Durrani et al., 2013; Zhang et al., 2013) employed Markov (n-gram) model over sequence of bilingual tuples also known as minimal translation units (MTUs).

Recently, some pioneer studies (Schwenk et al., 2007; Son et al., 2012) proposed feed-forward neural networks with factorizations to model bilingual tuples in a continuous space. Although the authors reported some gains over the n-gram model in machine translation tasks, these models can only capture a limited amount of context and remain a kind of n-gram model. In language modeling, experimental results in (Mikolov et al., 2011; Arisoy et al., 2012; Sundermeyer et al., 2013) showed that recurrent neural networks (RNNs) outperform feed-forward neural networks in both perplexity and word error rate in speech recognition even though it is harder to train properly.

Therefore, in this paper we take the advantages of RNN and tuple sequence model and propose recurrent neural network-based tuple sequence models (RNNTSMs) to improve phrase-based translation system. Our RNNTSMs are capable of modeling long-span context and have better generalization. Compared with such related studies as (Schwenk et al., 2006; Son et al., 2012), our main contributions can

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/

[1]The IWSLT workshop aims at translating TED speeches (http://www.ted.com), a collection of public lectures covering a variety of topics.



*Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1908–1917, Dublin, Ireland, August 23-29 2014.

be summarized as: (i) our models can be regarded as deep neural network translation models because they can capture arbitrary-length context potentially, which are proven to estimate more accurate probabilities of bilingual tuples; (ii) we extend the conventional RNNTSM to factorized RNNTSMs that can significantly overcome the data sparseness problem caused by the large vocabularies of bilingual tuples by incorporating the factors from the source and the target sides in addition to bilingual tuples; (iii) we investigate heuristic rules to decompose phrasal bilingual tuples to word bilingual tuples for reducing the out-of-tuple-vocabulary rate and providing fine-grained tuple sequence model; (iv) we integrate the proposed models into the state-of-the-art phrase-based translation system (MOSES) as a supplement of the work in (Son et al., 2012) that is a complete n-gram translation system.

## 2 Related Work

The n-gram translation model (Marino et al., 2006) is a Markov model over phrasal bilingual tuples and can improve the phrase-based translation system (Koehn et al., 2003) by providing contextual dependencies between phrase pairs. To further improve the n-gram translation model, Crego and Yvon (2010) explored factored bilingual n-gram language models. Durrani et al. (2011) proposed a joint sequence model for the translation and reordering probabilities. Zhang et al. (2013) explored multiple decomposition structures as well as dynamic bidirectional decomposition. Since neural networks advance the state of the art in the fields of image processing, acoustic modeling (Seide et al., 2011), language modeling (Bengio et al., 2003), natural language processing (Collobert et al., 2011; Socher et al., 2013), machine transliteration (Deselaers et al., 2009), etc, some prior studies have been done on neural network-based translation models (NNTMs).

One kind of the NNTMs relies on word-to-word alignment information or phrasal bilingual tuples. For example, Schwenk et al. (2007) investigated feed-forward neural networks to model bilingual tuples in continuous space. Son et al. (2012) improved this idea by decomposing tuple units, i.e., distinguishing the source and target sides of the tuple units, to address data sparsity issues. Although the authors reported some gains over the n-gram model in the BLEU scores on some tasks, these models can only capture a limited amount of context and remain a kind of n-gram model. In addition, a feed-forward neural network independent from bilingual tuples was proposed (Schwenk, 2012), which can infer meaningful translation probabilities for phrase pairs not seen in the training data.

Another kind of the NNTMs do not rely on alignment. Auli et al. (2013) and Kalchbrenner and Blunsom (2013) proposed joint language and translation model with recurrent neural networks, in which latent semantic analysis and convolutional sentence model were used to model source-side sentence. Potentially, they can exploit an unbounded history of both source and target words thanks to recurrent connections. However, they only modestly observed gains over the recurrent neural network language model. Previous studies (Wu and Wang, 2007; Yang et al., 2013) showed that the performance of word alignment (alignment error rate) is nearly 80%. That means explicit word alignment may be more reliable as a way to represent the corresponding bilingual sentences compared with an implicit compressed vector representation (Auli et al., 2013).

Our RNNTSM takes the advantages of the above NNTMs, that is, RNN enables our model to capture long-span contextual information, while tuple sequence model uses word alignment without much information loss. Furthermore, factorized RNN and word bilingual tuples are proposed to address data sparsity issue. To the best of our knowledge, few studies have been done on this aspect.

## 3 Tuple Sequence Model

In tuple sequence model, bilingual tuples are translation units extracted from word-to-word alignment. They are composed of source phrases and their aligned target phrases that are also known as minimal translation units (MTUs) and thus cannot be broken down any further without violating the constrains of the translation rules. This condition results in a unique segmentation of the bilingual sentence pair given its alignment. In our implementation, GIZA++ with `grow-diag-final-and` setting is used to conduct word-to-word alignments in both directions, source-to-target and target-to-source (Och and

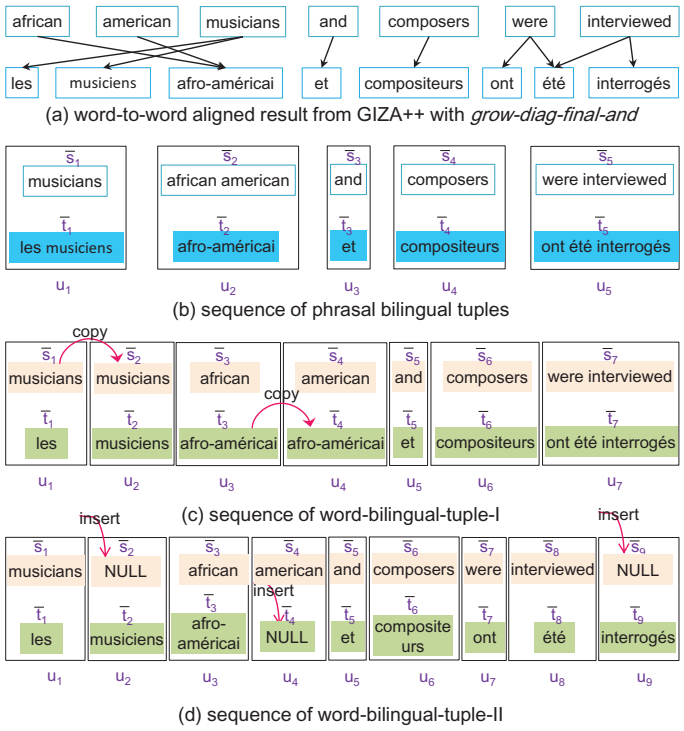


Figure 1: An example of generating basic bilingual tuples from word alignment information.

Ney, 2003). Ncode toolkit[2] is used to generate a unique bilingual segmentation of word-to-word aligned sentence. Figure 1(a)-(b) illustrates the process of generating bilingual tuple. As can be seen in Figure 1, bilingual tuple $u_1$ is composed of source phrase $\bar{s}_1$ (musicians) and target phrase $\bar{t}_1$ (les musiciens) linked to $\bar{s}_1$. Because this type of bilingual tuples are composed of one or more words from the source side and zero or more words from the target side, we call them phrasal bilingual tuples.

The phrasal bilingual tuple is not able to provide translations for individual words that appear tied to other words unless they occur alone in some other tuple. For example, if target phrase $\bar{t}_k$="les musiciens" is always aligned to source phrase $\bar{s}_k$="musicians" in the training corpus, then no word-to-word translation probability for "musicians:musiciens" will exist. This becomes a serious drawback when a large number of phrasal bilingual tuples are extracted from one-to-many, many-to-one, and many-to-many alignments. To tackle the issue, we propose to decompose phrasal bilingual tuples into word bilingual tuples for providing fine-grained tuple sequence model. Suppose source phrase $\bar{s}_k$, a sequence of source word $s_{k1}, s_{k2}, ..., s_{kI}$, is aligned to target phrase $\bar{t}_k$, a sequence of target word $t_{k1}, t_{k2}, ..., t_{kJ}$, in which $I$ and $J$ refer to the number of words in source phrase and that in target phrase. The following two types of heuristic rules are considered.

(**word-bilingual-tuple-I**): For one-to-many alignments, we copy $s_{kI}$ $J-1$ times to fill the short phrase $\bar{s}_k$. For many-to-one alignments, we copy $t_{kJ}$ $I-1$ times to fill the phrase $\bar{t}_k$. For many-to-many alignment, a maximum phrase length, we set it to 5, is used to avoid vocabulary explosion. That means, if $I > 5$; then $\bar{s}_k$=⟨unk⟩, if $J > 5$; then $\bar{t}_k$=⟨unk⟩.

(**word-bilingual-tuple-II**): For one-to-many, many-to-one, and many-to-many alignments, we insert a special token "NULL" $\mid J - I \mid$ times to fill the short phrase, and map each word in the extended phrase monotonically to generate a word-wise tuple sequence.

The Figure 1(c)-(d) demonstrate the decomposition results. As shown in Figure 1(c), the translation probability of "musicians" being aligned to "musiciens" can be learned in the word bilingual tuples. The word bilingual tuples enable our model use information from source-side of the tuples for computing translation probabilities of some tuples. For example, translating "musicians:musiciens" benefits from its source word "musicians". Table 2 in Section 4 shows the sizes of the tuple vocabularies. We can see

[2] http://ncode.limsi.fr/

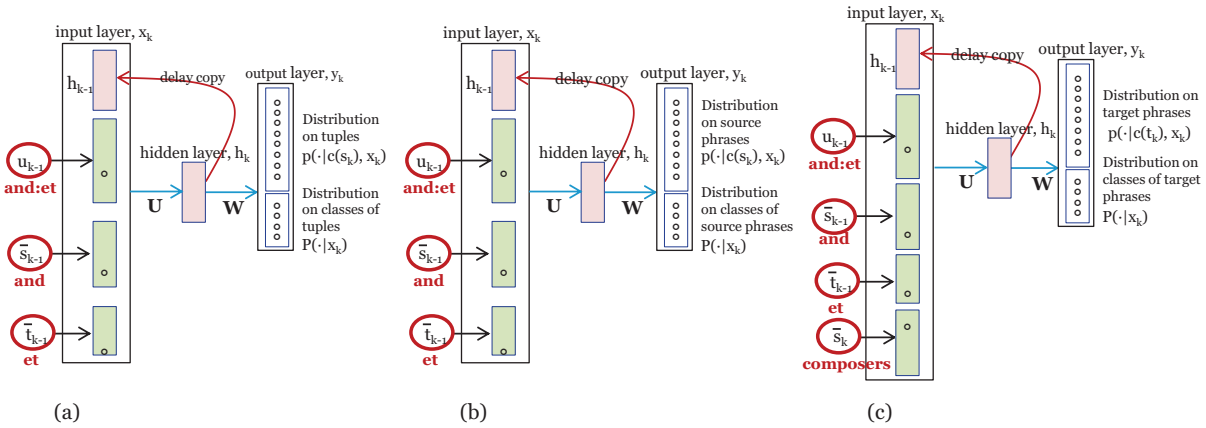


Figure 2: (a): factorized RNNTSM, called fRNNTSM for short, which will go back to the RNNTSM model when $\bar{s}_{k-1}$ and $\bar{t}_{k-1}$ are dropped. (b): fRNNTSM$_{\texttt{source}}$. (c) fRNNTSM$_{\texttt{target}}$.

that the word-bilingual-tuple-I has lower out-of-tuple-vocabulary (OOTV) rate, though it increases the tuple vocabulary. The word-bilingual-tuple-II greatly reduces the tuple vocabulary and the OOTV rate. Note that some words may not be aligned correctly, like "NULL-musiciens". However, generating these tuples can be viewed as a language model process that exploits previous source and target words, and current source word contained in previous tuples like "les-musiciens".

Thus, given a target sentence **t**, a source sentence **s**, and its alignment **a**, the tuple sequence model can be defined over the sequence of bilingual tuples $(u_1, u_2, ..., u_m)$ as follows.

$$p(\mathbf{t}, \mathbf{s}, \mathbf{a}) = \prod_{k=1}^{m} p(u_k|u_{k-1}, u_{k-2}, ..., u_1) = \prod_{k=1}^{m} p(u_k|u_{k-1}, u_{k-2}, ..., u_{k-n+1}) \quad (1)$$

where $u_k$ denotes the $k$-th bilingual tuple of a given bilingual sentence pair. Each bilingual tuple $u_k$ contains a source phrase $\bar{s}_k$ and its aligned target phrase $\bar{t}_k$[3]. Formally, $u_k$=$\bar{s}_k$:$\bar{t}_k$. The tuple sequence model does not make any phrasal independence assumption and generates a tuple by looking at a context of previous tuples. The n-gram translation models are Markov models over sequences of tuples, they generate a tuple by looking at previous n-1 tuples.

## 4 Recurrent Neural Network-based Tuple Sequence Model

In order to use long-span context, this paper presents a recurrent neural network-based tuple sequence model (RNNTSM) to approximate the probability $p(u_i|u_{i-1}, ..., u_1)$. Our RNNTSM can potentially capture arbitrary long context rather than n-1 previous tuples. The input layer encodes bilingual tuples by using 1-of-$n$ coding, and the output layer produces a probability distribution over all bilingual tuples. The hidden layer maintains a representation of the sentence history. This RNNTSM, however, has severe data sparsity problem due to the large tuple vocabulary coupled with the limited bilingual training data.

### 4.1 Factorized RNNTSM

To solve the problem, we extend the RNNTSM model with factorizing tuples in input layer, as shown in Figure 2(a). Specifically, it consists of an input layer $x$, a hidden layer $h$ (state layer), and an output layer $y$. The connection weights among layers are denoted by matrixes **U** and **W**. Unlike the RNNTSM, which predicts probability $p(u_k|u_{k-1}, h_{k-1})$, the factorized RNNTSM predicts probability $p(u_k|u_{k-1}, \bar{s}_{k-1}, \bar{t}_{k-1}, h_{k-1})$ of generating following tuple $u_k$ and is explicitly conditioned on the preceding tuple $u_{k-1}$, source-side of the tuple $\bar{s}_{k-1}$, and target-side of the tuple $\bar{t}_{k-1}$. It is implicitly conditioned on the entire history by the delay copy of hidden layer $h_{k-1}$. For those tuples (approximately 20% as shown in Table 2) that are not contained in the training data, i.e., co-occurrence $(s_{i-1}, t_{i-1})$

[3]Phrases turn to words in the word bilingual tuples. For convenience, we do not distinguish them in our paper.

non-exist while either $s_{i-1}$ or $t_{i-1}$ exists, the factorized RNNTSM backs off to the source- ($s_{i-1}$) or target-side ($t_{i-1}$). This process resembles factored $n$-gram language model (Duh and Kirchhoff, 2004). However, the RNNTSM, computing $p(u_i|u_{i-1}, h_{i-1})$, cannot estimate the probabilities for those tuples. In the special case that $\bar{s}_{k-1}$ and $\bar{t}_{k-1}$ are dropped, the factorized RNNTSM goes back to the RNNTSM. For convenience, $u_{k-1}$, $\bar{s}_{k-1}$ and $\bar{t}_{k-1}$ are called features. In the input layer, each feature is encoded into a feature vector using the 1-of-$n$ coding. The tuple $u_{k-1}$, the source phrase $\bar{s}_{k-1}$ and the target phrase $\bar{t}_{k-1}$ are encoded into $|u|$-dimension feature vector $v^{u}_{k-1}$, $|\bar{s}|$-dimension feature vector $v^{\bar{s}}_{k-1}$ and $|\bar{t}|$-dimension feature vector $v^{\bar{t}}_{k-1}$, respectively. Here, $|u|$, $|\bar{s}|$ and $|\bar{t}|$ stand for the sizes of the tuple, the source phrase, and the target phrase vocabularies. Finally, the input layer $x_k$ is formed by concatenating feature vectors and hidden layer $h_{k-1}$ at the preceding time step, as shown in the following equation.

$$x_k = [v^{u}_{k-1},\ v^{\bar{s}}_{k-1},\ v^{\bar{t}}_{k-1},\ h_{k-1}] \tag{2}$$

The neurons in the hidden and output layers are computed as follows:

$$\begin{aligned} h_k &= f(\mathbf{U} \times x_k), \ \ y_k = g(\mathbf{W} \times h_k) \\ f(z) &= \frac{1}{1+e^{-z}}, \quad g(z) = \frac{e^{z_m}}{\sum_k e^{z_k}} \end{aligned} \tag{3}$$

To speed-up both in the training and testing processes, we map bilingual tuples into classes with frequency binning and divide the output layer into two parts following (Mikolov et al., 2010). The first part estimates the posterior probability distribution over all classes. The second computes the posterior probability distribution over the tuples that belong to class $c(u_k)$, the one that contains predicted tuple $u_k$. Finally, translation probability $p(u_k|u_{k-1}, \bar{s}_{k-1}, \bar{t}_{k-1}, h_{k-1})$ is calculated by,

$$p(u_k|u_{k-1}, \bar{s}_{k-1}, \bar{t}_{k-1}, h_{k-1}) = p(c(u_k)|x_k) \times p(u_k|c(u_k), x_k) \tag{4}$$

### 4.2 Factorized RNNTSM on source and target phrases

The above factorized RNNTSM is conditioned on the previous context during computing the probability for tuple $u_k$. It does not exploit its source side $\bar{s}_k$. For example, tuple "composers:compositeurs" does not benefit from "composers". To address this limitation, we rewrite the probability in Equation 1.

$$\begin{aligned} p(u_k|u_{k-1}, u_{k-2}, ..., u_1) &= p(s_k, t_k|u_{k-1}, u_{k-2}, ..., u_1) \\ &= p(s_k|u_{k-1}, u_{k-2}, ..., u_1) \times p(t_k|s_k, u_{k-1}, u_{k-2}, ..., u_1) \end{aligned} \tag{5}$$

The first sub-model $p(s_k|u_{k-1}, u_{k-2}, ..., u_1)$ computes the probability distribution over source phrases. This model, called fRNNTSM$_{\texttt{source}}$ for short, can be regarded as a reordering model. The second sub-model $p(t_k|s_k, u_{k-1}, u_{k-2}, ..., u_1)$ is a translation model, abbreviated as fRNNTSM$_{\texttt{target}}$, which computes the probability distribution over $\bar{t}_k$ that are translated from $\bar{s}_k$. The two sub-models are computed with the recurrent neural networks shown in Figure 2(b)-(c). Another advantage of using the factorized RNNTSM on source and target phrases separately is that their training become faster because the vocabulary sizes of the source and target phrases are much smaller than that of the tuples.

### 4.3 Training

Training can be performed by the back-propagation through time (BPTT) algorithm (Boden, 2002) by minimizing an error function defined in the following equations.

$$L = \frac{1}{2} \times \sum_{i=1}^{N} (o_i - p_i)^2 + \gamma \times (\sum_{lk} u^2_{lk} + \sum_{tl} w^2_{tl}) \tag{6}$$

where $N$ is the number of training instances, $o_i$ denotes the desired output; i.e., the probability should be 1.0 for the predicted tuple in the training sentence and 0.0 for all others. $\gamma$ is the regularization term's weight, which is determined experimentally using a validation set. The training algorithm randomly initializes the matrixes and updates them with Equation 7 over all the training instances in several iterations.

| | English-French | | | English-German | | |
|---|---|---|---|---|---|---|
| | tst2010 | tst2011 | tst2012 | tst2010 | tst2011 | tst2012 |
| Baseline | 30.15 | 35.97 | 35.48 | 20.29 | 21.48 | 19.30 |
| +RNNTSM | $30.51_{(0.3)}$ | $36.11_{(0.1)}$ | $36.44_{(0.9)}$ | $20.67_{(0.4)}$ | $21.85_{(0.4)}$ | $19.56_{(0.3)}$ |
| +fRNNTSM (1) | $31.83_{(1.6)}$ | $37.58_{(1.6)}$ | $37.74_{(2.2)}$ | $21.67_{(1.4)}$ | $22.89_{(1.4)}$ | $20.60_{(1.3)}$ |
| +fRNNTSM$_{\texttt{source}}$(2) +fRNNTSM$_{\texttt{target}}$(3) | $31.89_{(1.7)}$ | $38.23_{(2.2)}$ | $37.82_{(2.3)}$ | $21.49_{(1.2)}$ | $22.94_{(1.4)}$ | $20.41_{(1.1)}$ |
| +(1) +(2) +(3) | $32.26_{(2.1)}$ | $38.36_{(2.4)}$ | $38.11_{(2.6)}$ | $21.80_{(1.5)}$ | $22.88_{(1.4)}$ | $20.76_{(1.5)}$ |

Table 1: BLEU scores of the RNNTSMs, the factorized RNNTSM (fRNNTSM), the fRNNTSM$_{\texttt{source}}$ (sfRNNTSM), the fRNNTSM$_{\texttt{target}}$ with the word-bilingual-tuple-I and their combination. The numbers in the parentheses are the absolute improvements over the Baseline.

In Equation 7, $\psi$ stands for one of the connection weights in the neural networks and $\eta$ is the learning rate. After each iteration, it uses validation data for stopping and controlling the learning rate. Usually, our RNNs needs 10 to 20 iterations.

$$\psi^{new} = \psi^{previous} - \eta \times \frac{\partial L}{\partial \psi} \tag{7}$$

## 5 Experiments

We experiment with two language pairs on the IWSLT2012 data sets (Federico et al., 2012), with English as source and French, German as target. The IWSLT data comes from TED speecheds, given by leaders in various fields and covering an open set of topics in technology, entertainment, design, and many others. In the following experiments, the IWSLT dev2010 set is used as the tuning set, the tst2010, tst2011, and tst2012 as the test sets.

Phrase-based translation systems are constructed as baselines using standard settings (GIZA++ alignment, grow-diag-final-and, lexical reordering models, SRILM, and MERT optimizer) in the MOSES toolkit (Koehn et al., 2007). The proposed models are used to re-score n-best lists produced by the baseline systems. The n-best size is set to at most 1000 for each test sentence. During the n-best re-scoring, the weights are re-tuned on the dev2010 data set with MERT optimizer[4]. The proposed RNN-based models are evaluated on a small task and a large task. For the parameters of all the RNN-based models, we set the number of hidden neurons in the hidden layer to 480 and classes in the output layer to 300.

### 5.1 Small Task

In the small task, the training data only contains the speech-style bi-text, i.e., the human translation of TED speeches. Specially, the corpora for the English-French and English-German pairs contain 139K and 128K parallel sentences. The language model is a standard 4-gram language model with the Kneser-Ney discounting. Both the $n$-gram LM and the RNNLM are trained on the target side of the bi-text corpus. As the first experiment, we compare the proposed RNNTSMs with the word-bilingual-tuple-I. Table 1 summarizes the results. The main findings from this experiment are: (1) The RNNTSM yields modest improvements of 0.3%-0.4% over the baseline system on most the test sets. (2) The factorized RNNTSMs essentially outperform the baseline and the RNNTSM on all the test sets. Specially, the improvements of the factorized RNNTSM and the combination of the fRNNTSM$_{\texttt{source}}$ and the fRNNTSM$_{\texttt{target}}$ over the baseline for the English-French task range 1.6%-2.2% and 1.7%-2.3%. For the English-German pair, these improvements are between 1.3%-1.4% and 1.1%-1.4%. The results indicate that the factorized RNNTSMs can well address the data sparsity problem of the RNNTSM. (3) The improvements for the English-German pair are comparatively smaller than that for the English-French pair. This is because German is a morphologically rich language (Fraser et al., 2013), its vocabulary is larger and the sparsity

[4]To get statistically reliable comparison (Clark et al., 2011), replication of the MERT optimizer and test set evaluation are performed five times. We finally report the average BLEU scores in the following experiments.

| | English-French | | | English-German | | |
|---|---|---|---|---|---|---|
| | #Tuple | #Source/#Target | OOTV | #Tuple | #Source/#Target | OOTV |
| Phrasal bilingual tuple | 308K | 130K/175K | 26.0% | 315K | 148K/196K | 23.4% |
| word-bilingual-tuple-I | 332K | 100K/111K | 23.7% | 351K | 104K/135K | 22.2% |
| word-bilingual-tuple-II | 293K | 44K/56K | 14.9% | 327K | 43K/86K | 14.8% |

Table 2: Vocabulary sizes. OOTV refers to the out-of-tuple-vocabulary rate on the dev2010 set. K stands for thousands.

| | English-French | | | English-German | | |
|---|---|---|---|---|---|---|
| | tst2010 | tst2011 | tst2012 | tst2010 | tst2011 | tst2012 |
| +fRNNTSM$_{\mathrm{p}}$ | 31.44 | 37.68 | 37.34 | 20.76 | 21.80 | 19.57 |
| +fRNNTSM$_{\mathrm{I}}$ | $31.83_{(0.4)}$ | $37.58_{(-0.1)}$ | $37.74_{(0.4)}$ | $21.67_{(0.9)}$ | $22.89_{(1.1)}$ | $20.60_{(1.0)}$ |
| +fRNNTSM$_{\mathrm{II}}$ | $31.73_{(0.3)}$ | 37.66 | $37.78_{(0.4)}$ | $22.00_{(1.2)}$ | $23.24_{(1.4)}$ | $21.09_{(1.5)}$ |
| +fRNNTSM$_{\mathrm{I}}$ +fRNNTSM$_{\mathrm{II}}$ | $31.98_{(0.6)}$ | $37.97_{(0.3)}$ | $38.14_{(0.6)}$ | $22.19_{(1.4)}$ | $23.25_{(1.4)}$ | $21.17_{(1.7)}$ |

Table 3: BLEU scores of the factorized RNNTSM with various types of bilingual tuples. fRNNTSM$_{\mathrm{p}}$ refers to the fRNNTSM with phrasal bilingual tuples, fRNNTSM$_{\mathrm{I}}$ to the fRNNTSM with the word-bilingual-tuple-I, etc. + means these models are used with the baseline systems. The numbers in the parentheses are the absolute improvements over the +fRNNTSM$_{\mathrm{p}}$.

problem is more serious. (4) There is no significant difference between the factorized RNNTSM and the combination of the fRNNTSM$_{\mathtt{source}}$ and the fRNNTSM$_{\mathtt{target}}$ on most of the test sets except for the tst2011 set of the English-French task. However, the BLEU scores are modestly improved by combining the three factorized RNNTSMs.

The second experiment is to compare the phrasal bilingual tuples and the word bilingual tuples. Table 2 lists the vocabulary sizes of the tuples, source and target phrases. For the word-bilingual-tuple-I, the bilingual tuple vocabulary size increases by 10% in both the English-French and the English-German pairs. Compared with the phrasal bilingual tuples, the bilingual tuple vocabulary size in the word-bilingual-tuple-II slightly changes. In addition, decomposing the tuples is capable to greatly reduce the out-of-tuple-vocabulary rate by approximately 50% in the word-bilingual-tuple-II. Table 3 compares bilingual tuples in terms of BLEU scores. It can be clearly seen that both the word-bilingual-tuple-I and the word-bilingual-tuple-II achieve better performance than the phrasal bilingual tuple on most of the test sets. The BLEU improvements of the word-bilingual-tuple-II over the phrasal tuple range 1.2-1.5 points on the English-German task. The main reason may be lie in: the decomposition can provide word-to-word translation probabilities (such as "musicians:musiciens" in the example of Section 2) for those non-one-to-one alignments. Thus the translation system will have a translation option for an isolated occurrence of such words. Another important observation is that the decomposition performs differently on the English-French and English-German tasks. For example, there exists slight difference between the word-bilingual-tuple-I and the word-bilingual-tuple-II for the English-French task. However, for the English-German task, the word-bilingual-tuple-II significantly outperforms the word-bilingual-tuple-I by 0.4 BLEU scores. Lastly, we achieve modest improvements by combining the two types of word bilingual tuples.

This paper proposes three factorized RNNTSMs and two types of word bilingual tuples. In this experiment, we combine all of them (+Combination contains 6 models) and compare with RNN-based language model (Mikolov et al., 2010). Table 4 summarizes the results. As shown in Table 4 and Table 1, the combination can further enhance the performance on the English-German task. For example, the combination improves the factorized RNNTSM with the word-bilingual-tuple-I from 20.76 to 21.29 on the tst2012 set of the English-German task. Moreover, the combination significantly outperforms the RNNLM. The improvements over the RNNLMs on all test sets range 0.7-1.2 BLEU scores. The

| | English-French | | | English-German | | |
|---|---|---|---|---|---|---|
| | tst2010 | tst2010 | tst2012 | tst2010 | tst2010 | tst2012 |
| Baseline | $30.15_{(-1.2)}$ | $35.97_{(-1.2)}$ | $35.48_{(-1.5)}$ | $20.29_{(-0.8)}$ | $21.48_{(-0.9)}$ | $19.30_{(-0.8)}$ |
| +RNNLM | 31.43 | 37.23 | 37.04 | 21.14 | 22.39 | 20.08 |
| +Combination | $32.10_{(0.7)}$ | $38.04_{(0.8)}$ | $37.75_{(0.8)}$ | $22.13_{(1.0)}$ | $23.64_{(1.2)}$ | $21.29_{(1.2)}$ |

Table 4: BLEU scores of the combination of our proposed models and RNNLM in the small task. The numbers in the parentheses are the absolute improvements over the RNNLM.

| | English-French | | | English-German | | |
|---|---|---|---|---|---|---|
| | tst2010 | tst2010 | tst2012 | tst2010 | tst2010 | tst2012 |
| Baseline | $32.92_{(-1.0)}$ | $38.67_{(-1.2)}$ | $39.41_{(-1.4)}$ | $22.29_{(-0.5)}$ | $23.67_{(-0.4)}$ | $20.83_{(-0.7)}$ |
| +RNNLM | 33.93 | 39.90 | 40.82 | 22.80 | 24.12 | 21.49 |
| +Combination | $34.24_{(0.3)}$ | $40.37_{(0.5)}$ | $40.92_{(0.1)}$ | $23.61_{(0.8)}$ | $25.18_{(1.1)}$ | $22.64_{(1.1)}$ |

Table 5: BLEU scores of the combination of our proposed models and RNNLM. The numbers in the parentheses are the absolute improvements over the RNNLM.

improvement over the baseline are between 1.9-2.3 BLEU points.

### 5.2 Large Task

In the large task, the training data includes both speech-style and text-style bi-text corpora. The text-style bi-text corpora are collected from the WMT2012 campaign[5], including CommonCrawl, NewsCommentary, and Europarl. Totally, the numbers of the parallel sentences are 4.35M for the English-French task and 3.85M for the English-German task. The language model is obtained by linear interpolation of several 4-gram models trained on the target side of bi-text corpora and the LDC French Gigaword corpus.

Table 5 reports the results. +Combination means the combination of six models, as described in Table 4. We can observe that: (1) The combination of the proposed RNNTSMs only trained on the speech-style data can essentially enhance the baselines by 1.2-1.8 BLEU points. (2) The improvements over the RNNLMs are significant on the English-German task but these improvements are modest on the English-French task. Note that the factorized RNNTSMs and the RNNLMs in the large task are also only trained the speech-style parallel corpus. In future work, we will train them on a bigger corpus, which can be expected to further increase the performance (Auli et al., 2013; Wu et al., 2012).

## 6 Conclusion

Most prior neural network-based translation models either employ feed-forward neural networks to explicitly integrate source information via word-to-word alignment, or use recurrent neural networks in which source information is implicitly represented with a compressed vector. In this paper, we present recurrent neural network-based tuple sequence models (RNNTSMs) to compute probabilities of bilingual tuples in continuous space. One of major advantages is their potential to capture long-span history compared with feed-forward neural networks. In addition, our models can well address the data sparsity problem thanks to the fine-grained word bilingual tuples and the factorized recurrent neural networks. As can be concluded from the experimental results on the IWSLT2012 test sets, our factorized RNNTSMs with the proposed bilingual tuples can essentially improve the BLEU scores for the English-French and English-German tasks.

We plan to incorporate re-ordering and syntactic features into RNNTSMs and evaluate them on distant language pairs, such as English-Chinese (Japanese) tasks in the future. Moreover, we will prune large tuple vocabulary and speed up the training on bigger data.

[5] http://www.statmt.org/wmt12/translation-task.html

## References


Ebru Arisoy, Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramabhadran. 2012. Deep neural network language models. In *Proceedings of NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28.

Michael Auli, Galley Michel, Quirk Chris, and Zweig Geoffrey. 2013. Joint language and translation modeling with recurrent neural networks. In *Proceedings of EMNLP2013*, pages 1044–1054.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. In *Journal of Machine Learning Research*, pages 1137–1155.

Mikael Boden. 2002. A guide to recurrent neural networks and backpropagation. Technical report.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of ACL 2011*, pages 176–181.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(3):2493–2537.

Josep M. Crego and Francois Yvon. 2010. Factored bilingual n-gram language models for statistical machine translation. *Machine Translation, Special Issue: Pushing the frontiers of SMT*, 24(2):159–175.

Thomas Deselaers, Sasa Hasan, Oliver Bender, and Hermann Ney. 2009. A deep learning approach to machine transliteration. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 233–241.

Kevin Duh and Katrin Kirchhoff. 2004. Automatic learning of language model structure. In *Proceedings of COLING 2004*, pages 148–154.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. *Proceedings of ACL 2011*, pages 1045–1054.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can markov models over minimal translation units help phrase-based smt? *Proceedings of ACL 2013*, pages 399–405.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stuker. 2012. Overview of the iwslt 2012 evaluation campaign. In *Proceedings of IWSLT 2012*.

Alexander Fraser, Helmut Schmid, Richard Farkas, Renjing Wang, and Hinrich Schutze. 2013. Knowledge sources for constituent parsing of german, a morphologically rich and less-configurational language. *Computational Linguistics*, 39(1):57–85.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous transaltion models. In *Proceedings of EMNLP2013*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *Proceedings of NAACL 2003*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, and etc. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL2007 on Interactive Poster and Demonstration Sessions*, pages 177–180.

Jose B. Marino, Rafael E. Banchs, Josep M. Crego, Adria de Gispert, Patrik Lambert, Jose A.R. Fonollosa, and Marta R. Costa-jussa. 2006. N-gram-based machine translation. In *Computational Linguistics*, volume Volume 32 Issue 4, pages 527–549.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan. H. Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of INTERSPEECH 2010*, pages 1045–1048.

Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Honza Cernocky. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Proceedings of INTERSPEECH 2011*, pages 605–608.

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

Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of COLING/ACL 2006*, pages 723–730.

Holger Schwenk, Marta R. Costa-jussa, and Jose A. R. Fonollosa. 2007. Smooth bilingual n-gram translation. In *Proceedings of EMNLP/HLT 2007*, pages 430–438.

Holger Schwenk. 2012. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012*, pages 1071–1080.

Frank Seide, Gang Li, Xie Chen, and Dong Yu. 2011. Feature engineering in context-dependentdeep neural networks for conversational speech transcription. In *Proceedings of ASRU 2011*, pages 24–29.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng, and Chris Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP 2013*.

Le Hai Son, Alexandre Allauzen, and Francois Yvon. 2012. Continuous space translation models with neural networks. In *Proceedings of HLT-NAACL 2012*, pages 39–48.

Martin Sundermeyer, Ilya Oparin, Jean-Luc Gauvain, Ben Freiberg, Ralf Schlter, and Hermann Ney. 2013. Comparison of feedforward and recurrent neural network language models. In *Proceedings of ICASSP 2013*, pages 8430–8433.

Hua Wu and Haifeng Wang. 2007. Comparative study of word alignment heuristics and phrase-based smt. In *Proceedings of MT SUMMIT XI*, pages 305–312.

Youzheng Wu, Xugang Lu, Hitoshi Yamamoto, Shigeki Matsuda, Chiori Hori, and Hideki Kashioka. 2012. Factored language model based on recurrent neural network. In *Proceedings of COLING 2012*, pages 2835–2850.

Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word alignment modeling with context dependent deep neural network. In *Proceedings of ACL2013*, pages 166–175.

Hui Zhang, Kristina Toutanova, Chris Quirk, and Jianfeng Gao. 2013. Beyond left-to-right: Multiple decomposition structures for smt. *Proceedings of NAACL-HLT 2013*, pages 12–21.