Rule Refinement for Spoken Language Translation by Retrieving the Missing Translation of Content Words

Linfeng Song†, Jun Xie†, Xing Wang‡, Yajuan Lü† and Qun Liu†§
†*Key Laboratory of Intelligent Information Processing*
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
‡*School of Computer Science & Technology*
Soochow University, Suzhou, China
§*Centre for Next Generation Localisation*
School of Computing, Dublin City University, Dublin, Ireland
{*songlinfeng,xiejun,wangxing,lvyajuan,liuqun*}*@ict.ac.cn*

***Abstract*—Spoken language translation usually suffers from the missing translation of content words, failing to generate the appropriate translation. In this paper we propose a novel Mutual Information based method to improve spoken language translation by retrieving the missing translation of content words. We exploit several features that indicate how well the inner content words are translated for each rule to let MT systems select better translation rules. Experimental results show that our method can improve translation performance significantly ranging from 1.95 to 4.47 BLEU points on different test sets.**

***Keywords*-Spoken Language Translation; Content Words; Mutual Information; Rule Refinement;**

I. Introduction

The last several years have seen a land rush in research on spoken language translation (SLT). One commonly-cited weaknesses of SLT is the missing translation of content words[1]. For example, our experimental results on IWSLT 2010[2] data sets show that 14.6% of the sentences in the testsets suffer from this problem. Since content words usually contain very meaningful information, the missing translation of content words affects the quality of SLT in profound ways.

However previous works mainly focus on discarding redundant or incorrect machine translation (MT) rules rather than refining them. For example, [3]–[5] modify the rule extraction process to filter bad rules during extraction, while [6]–[14] directly filter redundant rules or bad rules after the ruletable has been generated. Although some of their works take the missing translation of content words as a criterion when recognizing incorrect translation rules, they suffer the high risk of filtering correct rules as well.

On the other hand, Huck and Ney [15] propose several Insertion and Deletion models that can be used to avoid the missing translation of content words and experimental results show significant improvement. However their work is not specific on the problem of missing content words (such as making use of relevant linguistic knowledge to help recognize content words) and they do not give relevant evidence to prove their improvement comes from the retrieving of content words.

In this paper, we propose a Mutual Information (MI) [1] based model that considers the bilingual correspondences between the content words and their translations. Our approach is inspired by the observation that there is a significant correspondence between each content word s and its proper translation t in word aligned bilingual corpus. Further more, in spoken language, there are usually some idioms or slangs that can not be translated at the word level. For example, the idiom "爱屋及乌" which means "love me love my dog" should be translated as a whole phrase. Therefore, we exploit the bilingual correspondences at different granularities (i.e., word-level, phrase-level). We exploit MI based features to indicate how well the content words or phrases in a MT rule are translated, giving prior to the translation of content words. Experimental results on IWSLT data sets show that our approach achieves significant improvement in translation performance significantly ranging from 1.95 to 4.47 BLEU points on different test sets.

II. The Model for Rule Refinement

Generally a *content word* refers to a content word in the source language. Any continuous content word sequence makes up a *content phrase*. Any content phrase not contained in other content phrases is called a *maximal content phrase*. We call a content word or content phrase and its translation a *content pair* and call the target side phrase of a content pair a *target content phrase*.

A. Model Description

For each translation rule $< S, T, \sim >$, we define its model score through formula (1):

$$\begin{aligned} Score(S,T) &= \frac{\sum_{s_i \in S} score(s_i,T)}{count(s_i)} \\ &= \frac{\sum_{s_i \in S} \arg\max_j MI(s_i,t_j)}{count(s_i)} \end{aligned} \quad (1)$$

We suppose the source phrase S to be a bag-of-words, that is each source word s_i is independent from others. We define the overall score $Score(S,T)$ to be the average of the model score between each content word s_i and the

[1]In linguistics content words are words such as nouns, verbs, adjectives that refer to some object, action, or other non-linguistic meaning.

[2]http://iwslt2010.fbk.eu/

978-0-7695-5063-3/13 $26.00 © 2013 IEEE
DOI 10.1109/IALP.2013.23

Algorithm 1 Scoring algorithm for Maximal Content Phrase

1: *Input*: Maximal Content Phrase *mcp*
2: $map < Span, double > \ beam \ = \ \{\}$
3: **for** $length \leftarrow 1...len(mcp)$ **do**
4: **for** $i \leftarrow 0...len(mcp) - length$ **do**
5: $j = i + length - 1$
6: $phr = words[i:j]$
7: $score\_w = calc\_model\_score(phr)$
8: $score\_s = DOUBLE\_MIN$
9: **for** $k \leftarrow i...j-1$ **do**
10: $sl = beam[i:k]$
11: $sr = beam[k+1:j]$
12: $temp = (sl + sr)/2.0$
13: $score\_s = \max(score\_s, temp)$
14: $beam[i,j] = \max(score\_w, score\_s)$
15: *return* $beam[0, len(mcp) - 1]$

target phrase T. The reason for using the average score is that the number of content words varies from rule to rule, so summing up would lead our model bias to the rules containing more content words.

Directly calculating $score(s_i, T)$ is inappropriate because considering the whole phrase T will suffers data sparseness problem. Here we assume T to be a bag-of-words too and define $score(s_i, T)$ to be the maximum Mutual Information score of any content pair (s_i, t_j) containing s_i. This is intuitive since we only want to capture the co-relation between each s_i and its current best translation t_j under our model, such as the co-relation between word " 地铁" and "subway" in translation rule $< 坐\ 地铁, take\ a\ subway, \sim >$.

B. Training

We make statistics on bilingual corpus C with word alignment information. For each content pair (s, t), we define $P(s, t)$ to be the probability of co-appearance and alignment agreement between s and t in the corpus, and then define $P(s)$ and $P(t)$ as the probability of s and t's appearance respectively. So the Mutual Information between s and t can be calculated through formula (2):

$$MI(s,t) = log\frac{P(s,t)}{P(s)P(t)} \tag{2}$$

Similar to the classic phrase extracting process of SMT [17], the training process enumerates every bilingual correspondences of different granularities (both word-level and phrase-level) in corpus C with maximal length limitation M, and then extract every content pair (s, t) that satisfies alignment agreement. Suppose we have extracted N content pairs from the training corpus, then we further define $P(s, t)$, $P(s)$ and $P(t)$ through formula (3) which α can be either s or t or both. Finally we train our model based on formula (2) and formula (3).

$$P(\alpha) = \frac{\#count(\alpha)}{N} \tag{3}$$

C. Rule Scoring

For each MT rule *r*, we first obtain all the maximal content phrases in it. And then we calculate the model score for each maximal content phrase *mcp*. Finally we take the average of these scores as the score for the whole rule:

$$P(r) = \frac{\sum_{mcp \in r} score(mcp)}{count(mcp)} \tag{4}$$

We define our scoring algorithm for each maximal content phrase as a binary bottom-up dynamic programming procedure. We adopt the classic CYK algorithm which has been used in many NLP areas such as parsing and SMT. Shown in Algorithm 1, we deal with each sub content phrase phr in bottom-up order (line 3 to 6). For each content phrase phr, we first calculate $score\_w$ which represents the maximal Mutual Information score between phr and any target phrase in the rule (line 7). In addition, we enumerate each possible binary division of phr (line 9), calculate the average of division (line 10 to 12) and save the maximum as $score\_s$ (line 13). We finally take the maximum between $score\_w$ and $score\_s$ as the final score for phr (line 14).

D. Content Words Recognition

Content words recognition can be achieved by either POS-tag based method or stoplist based method. Generally POS-tag based methods consider any word with certain tags (such as Verbs and Nouns) as content word, while stoplist based methods consider any word not being included in the stoplist as content word. In this work we use a stoplist based method and our stoplist is composed by top 50 high-frequency words plus manually collected common functional words.

Basically any source word $s \in S$ in a translation rule $< S, T, \sim >$ that is not included in our stoplist is a content word. However in some qualified rule, some content words may be wrongly separated into several parts. For example, in rule $< 坐\ 地\ 铁, take\ a\ subway, \sim >$, word "地铁" which means "subway" is wrongly separated into words " 地" and " 铁". In order to handle this situation, we also consider another type of words as content words too. That is any word that can compose a linguistic word with its siblings and that linguistic word is not in our stoplist. We consider any word linguistic if it is in the HowNet[3].

In the example above, we consider word " 地" as content word, because it can compose linguistic word "地铁" with its right sibling word "铁" and word " 地铁" is not included in our stoplist. Similarly we consider word "铁" as content word too.

III. Incorporate into SMT

Generally we integrate our model score as a feature into the log-linear model of SMT. However we find that some special rules may be overestimated by our model because of the deficiency of Mutual Information. We observe that most of these rules contain many unaligned content words in both sides. So we introduce two penalties to punish those situations.

[3]http://www.keenage.com/

System	DEVSET4	DEVSET5	DEVSET6
baseline	61.26	60.39	49.25
+SU	61.68	61.02	49.65
+TU	62.16	61.39	49.98
+SU+TU	62.06	61.32	49.84
phr_MI	63.48	64.86	51.20
+SU	63.80	64.91	51.45
+TU	63.89	64.97	51.97
+SU+TU	64.12	64.92	51.96

Table I: Experimental results on IWSLT 2010 test sets. *baseline* is the hierarchical phrase-based system, *phr_MI* means our model, *+SU* means our model plus source unaligned penalty, *+TU* means our model plus target unaligned penalty and *+SU+TU* means our model plus both penalties.

Source Unaligned Penalty

This feature represents the number of unaligned source content words in a rule. We only consider content words because it is reasonable for functional words to be unaligned while it is not for content words. Incorrect rules with many unaligned content words should be punished.

Target Unaligned Penalty

This feature records the number of unaligned target content words in a rule. Similarly bad rules with many unaligned target content words should be punished too.

We integrate them into the log-linear model of SMT as different features and the overall framework is shown in formula (5):

$$\underbrace{\sum_i \lambda_i h_i(e,f)}_{\text{classic features}} + \underbrace{\sum_p \lambda_p f_p(e,f)}_{\text{our features}} \quad (5)$$

IV. Experiment

A. Setup

Our training data consists of 280K Chinese-English oral sentence pairs, 40K of which are the training data for IWSLT 2010 CH-EN travel dialogue translation task, the rests come from our in-house data. We choose DEVSET2 of the same task as our development set and choose data sets from DEVSET4 to DEVSET6 of that task as our test sets. There is no overlap among training set, development set and test set.

We obtained the word alignments by running GIZA++ [18] on training data in both directions and applying "grow-diag-final" refinement. Then we extracted translation rules for Hierarchical Phrase-based(HPB) model [20]. We applied SRI Language Model Toolkit [16] to train a 5-gram language model with Kneser-Ney smoothing on the target side of the training corpus.

We took our in-house hierarchical phrase-based system as baseline. In our experiments, parameters were tuned by minimum error rate training (MERT) [19]. The translation quality was evaluated by the case insensitive NIST BLEU-4 metric[4]. We trained our model based on the GIZA++ alignment result on the training data. We extracted content pairs with the maximal length limitation $M = 4$.

[4]ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

B. Results

We did our experiments with the configuration just mentioned above, and the results are shown in table I. We observe that our model improves the translation quality by +2.22, +4.47 and +1.95 BLEU points respectively on each data set which proves the effectiveness of our model.

We also tested the effectiveness of the penalties which are mentioned in Section 3. Firstly, we can see that slight improvement is achieved by simply incorporating either *+SU* or *+TU* or both. Besides, the experimental results indicate that *+TU* is slightly more effective than *+SU* which quite fits our forecast. It's because the source content words are more accurately recognized by our content word recognition method while the target content words are automatically recognized by mapping from the source content words. So there are less room for improvement by introducing *+SU*. Finally, experimental results show that the combination of *+SU* and *+TU* is not more effective than simply using *+TU*. From this we conclude that single *+TU* is enough and adding *+SU* may not help. Additionally the divergence between *+SU* and *+TU* may be harmful for the performance.

C. Case Study

We compare some actual translations of IWSLT 2010 data sets generated by the baseline system and our *phr_MI* model. Shown in table II, we mark the different parts with boldface.

In the first example, there is a colloquial Chinese phrase "十 分钟 一 趟" which means "every ten minutes". Although it has the same meaning with the Chinese phrase "每 十 分钟", the former is much harder to translate since there is no lexical correspondence between "一 趟" and "every". In this case, although "一 趟" is not well translated in the baseline system, our model can captures the phrasal correspondence between "十 分钟 一 趟" and "every ten minutes".

As for the second case, we argue that it is difficult to translate for two reasons: firstly this Segment is composed by two sub-sentences and there is no obvious division (such as a comma or full stop) between them; secondly the subject of the first sub-sentence is omitted. Due to these reasons, both "身体" and "尽量" are lost in the results of baseline system. However, since both (身体, health) and (尽量, try to) are easily captured lexicalized correspondences, our model can encourage SMT systems to correctly translate them.

V. Conclusion

In this paper, we present a Mutual Information based model for rule refinement by retrieving the missing content words. Additionally we introduce several penalties to punish those rules containing too many unaligned content words in both sides. Finally we integrate our model and these penalties into the log-linear model of SMT as separate features. Our Experimental results show that

SEGMENT 155 OF DEVSET5	
Source	没有 你 需要 到 主街 上去 十 分钟 **一 趟** 。
Baseline	no , you need to go over to the main street ten minutes .
phr_MI	no , you need to go over to the main street **every** ten minutes .
Reference	no . you need to go over to the main street . the bus comes every ten minutes .
SEGMENT 118 OF DEVSET6	
Source	只要 是 **对 身体** 好 我 **尽量** 什么 都 吃 。
Baseline	as long as it 's good , i 'll eat anything .
phr_MI	as long as it 's good **for health** , i'll **try to** eat anything .
Reference	as long as it 's good for me , i 'll eat just about anything .

Table II: Some actual translations of IWSLT 2010 data sets produced by the baseline system and our *phr_MI* model. The differences between the two systems are marked with boldface.

our model can significantly improve the performance of Spoken Language Translation.

ACKNOWLEDGMENT

The authors were supported by 863 State Key Project No. 2011AA01A207. Qun Liu was also partially supported by Science Foundation Ireland (Grant No.07/CE/I1142) as part of the CNGL at Dublin City University. We thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] Robert M. Fano 1961. *Transmission of Information*. New York, New York: MIT-Press.

[2] Richard Zens, Daisy Stanton and Peng Xu. 2012. A Systematic Comparison of Phrase Table Pruning Techniques. *In Proceedings of EMNLP-CoNLL*, pages 972 – 983, Jeju Island, Korea, July.

[3] Nan Duan, Mu Li, Ming Zhou and Lei Cui. 2011. Improving Phrase Extraction via MBR Phrase Scoring and Pruning. *In Proceedings of MT Summit XIII*, pages 189 – 197, Xiamen, China, September.

[4] German Sanchis-Trilles, Daniel Ortiz-Martinez, Jesus Gonzalez-Rubio, Jorge Gonzalez and Francisco Casacuberta. 2011. Bilingual Segmentation for Phrasetable Pruning in Statistical Machine Translation. *In Proceedings of EAMT*, pages 257 – 264, Leuven, Belgium, May.

[5] Nadi Tomeh, Marco Turchi, Guillaume Wisinewski, Alexandre Allauzen, Francois Yvon. 2011. How Good Are Your Phrases? Assessing Phrase Quality with Single Class Classification. *In Proceedings of IWSLT*, pages 261 – 268, San Francisco, California, December.

[6] Howard Johnson, Joel Martin, George Foster and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. *In Proceedings of EMNLP-CoNLL*, pages 967 – 975, Prague, Czech Republic, June.

[7] Andreas Zollmann, Ashish Venugopal, Franz Och and Jay Ponte. 2008. A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. *In Proceedings of Coling*, pages 1145-1152, Manchester, August.

[8] Libin Shen, Jinxi Xu and Ralph Weischedel. 2008. A New String-to-Dependency Machine Translation Algorithm with a Target Dependency Language Model. *In Proceedings of ACL*, pages 577 – 585, Columbus, Ohio, USA, June.

[9] Zhongjun He, Yao Meng, Yajuan Lv, Hao Yu and Qun Liu. 2009. Reducing SMT Rule Table with Monolingual Key Phrase. *In Proceedings of ACL-IJCNLP*, pages 121 – 124, Suntec, Singapore, August.

[10] Zhongjun He, Yao Meng and Hao Yu. 2009. Discarding Monotone Composed Rule for Hierarchical Phrase-based Statistical Machine Translation. *In Proceedings of IUCS*, pages 25 – 29, Tokyo, Japan, December.

[11] Zhiyang Wang, Yajuan Lv, Qun Liu and Young-Sook Hwang. 2010. Better Filtration and Augmentation for Hierarchical Phrase-Based Translation Rules. *In Proceedings of ACL*, pages 142 – 146, Uppsala, Sweden, July.

[12] Gonzalo Iglesias, Adria de Gispert, Eduardo R. Banga and William Byrne. 2009. Rule Filtering by Pattern for Efficient Hierarchical Translation. *In Proceedings of EACL*, pages 380-388, Athens, Greece, April.

[13] Nadi Tomeh, Nicola Cancedda and Marc Dymetman. 2009. Complexity-Based Phrase-Table Filtering for Statistical Machine Translation. *In Proceedings of MT Summit XII*, Ottawa, Ontario, Canada, August.

[14] Mei Yang and Jing Zheng. 2009. Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT. *In Proceedings of ACL-IJCNLP*, pages 237 – 240, Suntec, Singapore, August.

[15] Matthias Huck and Hermann Ney. 2012. Insertion and Deletion Models for Statistical Machine Translation. *In Proceedings of NAACL*, pages 347 – 351, Montreal, Canada, June.

[16] Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. *In Proceedings of ICSLP*, volume 30, pages 901-904.

[17] Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *In Proceedings of HLT/NAACL*, Edmonton, Canada, July.

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

[19] Franz Josef Och. 2003. Minimum error rate training in statisticalmachine translation. *In Proceedings of ACL*, pages 160-167, Sapporo, Japan, July.

[20] David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*,24(11):503-512.