[PatentMT] Summary Report of Team III_CYUT_NTHU

Joseph Chang, Shih-ting Huang, Ho-ching Yen,
CS, NTHU, Taiwan, R.O.C.
{bizkit.tw, koromiko1104, fi26.tw}@gmail.com

Ming-jhuan Jiang, Chung-chi Huang,
ISA, NTHU, Taiwan, R.O.C.
{raconquer,u901571}@gmail.com

Jason S. Chang, *Ping-che Yang
CS, NTHU, Taiwan, R.O.C.
*III, Taipei, Taiwan, R.O.C.
{jason.jschang, maciaclark}@gmail.com

ABSTRACT
In this report paper, we investigate two issues facing phrase-based machine translation (MT) systems such as Moses (Koehn et al., 2007): out-of-vocabulary (OOV) words and singletons. MT systems typically ignore and directly output unknown or OOV source words into the target translation. On the other hand, for words which do not couple with their preceding or following words as phrases, as referred to as singletons, MT systems typically leave their translation disambiguation to language model within which knowledge is somewhat limited and determined by the preset length of words. In this paper, we first analyze the proportion of OOV words and singletons in translation task, summarize types of OOV words, and manually evaluate the impact of singletons on phrase-based MT systems. We also introduce methods for dealing with these two issues without changing the underlying phrase-based decoder.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *machine translation*, *text analysis*.

General Terms
Algorithms, Performance, Design, Experimentation, Languages.

Keywords
Machine translation, out-of-vocabulary words, singletons, phrase-based machine translation system, Moses, language model, translation model, decoder.

TeamName: [III_CYUT_NTHU]

Subtasks/Languages: [Chinese-to-English PatentMT]

External Resources Used: [Giza++, Moses, ip.com, NICT]

1. INTRODUCTION
Training data can not cover all the words or phrases. Some of the words in the source-language texts are unknown to the MT system, to be specific, unknown to its phrase tables or syntax-based translation rules. Whenever encountering an OOV word, MT systems typically ignore and leave the word untranslated to the target translation, which may in turn degrade the readability (let alone the readers who do not understand the source language) and the overall translation quality. The OOV problem could be better handled if a system recognized the constituents of an unknown word and leveraged the translations of the constituents to form possible translations for the OOV.

Many factors cause the source words to be out-of-vocabulary to a system. Some OOVs result from segmentation error in the source language (if the language does not have clear word boundary), some from name entities, such as person name, place, and organization, and others from low-frequency abbreviations (e.g., 體協 athletic association) and combination forms (e.g., eyebank, widebody) of common words (e.g., bank, body). According to our OOV analysis, the last type accounts for one fourth of OOV cases and takes up the same proportion as the OOV which can be paraphrased highly valued in the previous work (Marton et al. (2009); Mirkin et al. (2009)).

On the other hand, phrase-based MT systems have little means to disambiguate singletons in context except for their language models that are usually blamed for their lack of syntax knowledge due to limited word length. Phrases, or sequences of words to be exact, benefit phrase-based systems particularly in sense/translation disambiguation (one sense per discourse or collocation (Yarowsky, 1995)). According to our preliminary analyses, phrase-based systems such as Moses do not translate singletons very well especially nominal and verbal singletons.

In this report paper, we manage to propose methods to tackle with the OOV and singleton issues facing phrase-based decoders (syntax-based ones would definitely have OOV problem too). And we also report the analyses concerning the types of OOVs and the translation accuracy of singletons from an existing phrase-based system in a specific translation task.

2. RELATED WORK
Machine Translation (MT) has been an area of active research. In this paper, we focus on the OOV and singleton problems in phrase-based MT systems.

Recent work has been done on translating different OOV cases: name entities (NE), and compounds. Knight and Graehl (1997) introduced a statistical machine transliteration model to tackle proper names, while Hassan and Sorensen (2005) presented a NE translating approach that combines NE translation and transliteration. On the other hand, Cao and Li (2002) focused on translating compound words, especially noun phrases, using statistical approach and translation templates to translate noun phrases. Little research takes note of OOVs resulting from abbreviations of source phrases (e.g. 體協) or combination form of common words (e.g., 血庫). These two types cover some

portion of name entities and noun-noun and adjective-noun compounds (e.g., 邊貿 border trade (NN), and 新規 new regulations (AN)).

In the studies more closely related to our work, Marton et al. (2009) and Mirkin et al. (2009) proposed statistical paraphrase models that replace the OOV words with their corresponding in-vocabulary equivalents. For example, the OOV word "訪談" (interview) can be translated by replacing with the in-vocabulary word "訪問" (visit). Paraphrases in Marton et al. (2009) are learnt based on the word alignment information computed over a large additional set of bitexts. On the other hand, Mirkin et al. (2009) paraphrased OOV words using contextual entailment rules that are derived from monolingual corpora as well as manually compiled synonym thesaurus. These studies are similar in spirit to our work. However, we do not directly address the problem via paraphrasing. We translate the OOV by combining the translations of its constituents via wildcard search in the existing bilingual resources.

3. ANALYSES

In this section, we report the statistics on OOVs, singletons, different types of OOVs, and the retrieval rates of OOVs based on different query types and bilingual resources.

3.1 Analyses on OOV Words

To study the problem of translating OOV words, we used NIST MT-08 Chinese-to-English test set consisting of 1,273 unknown words in 637 sentences out of a total of 1,357 sentences. Among these 1,273 distinct OOV words, we inspected the number of OOVs with respect to their lengths, i.e., the number of characters (See Table 1). As can be seen in Table 1, OOV words of two characters accounted for more than half of the OOV cases. As a result, we focused on translating two-character unknown words. To analyze the OOV types and proportions, formulate partial-matched wildcard queries, and determine good bilingual resources for sublexical/constituent translation retrieval, we randomly selected 100 additional sentences containing only OOV words of two characters.

Table 1. The number of OOVs with respect to OOVs' lengths.

Length	Number of OOVs	Percentage(%)
1	56	4.4
2	683	53.7
3	352	27.7
4	115	9
5+	67	5.3

OOV words in these 100 sentences are manually classified into 10 types shown in Table 2 taking into consideration their reference translations (manually) extracted from reference sentences. Since our model was designed to retrieve and combine sublexical translations of an OOV word's constituents, it targets specifically on translating OOV words of the *combination forms* which accounts for one fourth of the OOV words. See Table 2 for more details.

Table 2. OOV types and their descriptions and examples.

OOV Type	Description of OOV Type	Example	# OOVs
Order Variants	Sequence of characters reversed without changing the original meaning	療治(治療) (treat)	1
Writing Variants	Replacement between simplified and traditional Chinese characters	念書 (唸書) (study)	1
Domain Specific	Domain specific terminologies	勤務 (service support) 二傳 (setter)	2
Word + Suffix	Words composed by a content character (underscored character) and a not translated function character	忙著 (busy) 爐子 (stove)	4
Informal	Used in conversation or informal writing	看頭 (worth watching) 幹麼 (what)	6
Old Use	Words rarely in use now	古稀 (60 years old) 橫流 (all over)	8
Name Entity	Name entities could be transliterated such as person, place, and organization	布希 (bush) 膠州 (jiaozhou)	12
Segmentation Error	Words erroneously split by the segmentation system	領式 (開領式) 會兒 (這會兒)	16
Rare Paraphrase	Words could be translated by replacing with its paraphrases	踐行 (practice) 訪談 (interview)	25
Combination Form	Words could be translated by combining sublexical translations	上肢 (upper limbs) 肌力 (muscle strength)	25

Table 3. The number of OOVs using the query type of 2 characters c_1, c_2.

Query type	Number of translatable OOVs	Example	
		OOV	Matched words
c_1 * and c_2 *	17	上肢 (upper limbs)	(上方 肢體)
c_1 * and * c_2	9	上肢 (upper limbs)	(上方 四肢)
* c_1 and c_2 *	2	震魔 (quake demon)	(地震 魔鬼)

* c_1 and * c_2	1	鐘 體 (bell body)	(時鐘 身體)

Intuitively, there were four ways to formulate the query for sublexical translations for a two-character OOV word c_1, c_2. Table 3 shows the first and the second query formulations of adding wildcard * can retrieve most relevant translations.

We further determine the effectiveness of various bilingual resources for finding sublexical translations. In other words, we compared translation hit rates using different resources based on the 1st and 2nd query types. Among resources, Lin Yutang's Chinese-English Dictionary, LDC translation lexicon (LDC2002L27), character-based phrase table and word-based phrase table, hit rates of 25 OOVs were 0.64, 0.68, 0.60, and 0.88, respectively. Word-based phrase table results in the highest hit rate, thus chosen as our bilingual resource for sublexical translation retrieval. It has advantages like different inflectional word forms and match of translation domain.

3.2 Analyses on Singletons

To investigate the percentage of singletons and their effect on MT translation quality, we randomly selected fifty sentences from NIST MT-08 English-to-Chinese test set. Singletons accommodate 6% of the words and are primary nouns (72%) and verbs (21%). Analyses on the translation quality of different parts-of-speech reveal that translation quality varies and while nouns achieve 50% accuracy, verbs achieve as low as 20%. Table 4 shows an example of the manually acquired reference translation for singletons.

Table 4. An example of singletons' reference translation.

Source sentence: 在 不足 半 個 月 時間 內 ， 王燕 的 **上肢** 肌力 恢復 了 兩級。

Reference 1: in less than half a month 's time , wang yan 's arm strength has recovered by two points .

Reference 2: muscle strength in wang yan 's arm has recover by two grades in less than half a month .

Reference 3: the muscle strengths of wang yan 's upper limbs improved by two levels in less than half a month .

Reference 4: in less than half a month , the muscle strength of wang yan 's upper limbs has regained by two levels .

Singletons pose a problem too in patent translation: singletons take up 5% based on the development data (2000 Chinese sentences) from NTCIR-11 patent translation, much higher than the OOV percentage (approximately 1.1%).

4. PROPOSED METHODS

4.1 Method for OOV

After obtaining the sublexical translations by qerying the abovementioned two types of sublexical/constituent (in Section 3.1) queries, we generate and rank the translation candidates for OOVs using the procedure in Figure 1. Note that such OOV model serves as a preprocessing component, translating unknown words prior to the MT systems.

procedure EvaluateCandidates (O, N, θ)

for each constituent c in OOV O
(1a) *SubTrans* = RetrieveSublexicalTranslations (c, O)
(1b) *CandList*[position (c, O)] = BilingualInfo(*SubTrans*, c)

(2a) *Straight* = *CandList*[1]
(2b) *Inverted* = *CandList*[len(O)]

for each constituent position cp >1 in ascending constituent positions of O
(3a) *Straight*× = *CandList*[cp]
for each constituent position cp <len(O) in descending constituent positions of O
(3b) *Inverted*× = *CandList*[cp]

(4a) *Straight* = MonolingualInfo (*Straight*)
(4b) *Inverted* = MonolingualInfo (*Inverted*)
Candidates = *Straight* + *Inverted*
(5) *RankedCandidates*=Sort *Candidates* in decreasing order of *Score*
(6) Return top N *RankedCandidates* with *Scores* exceeding θ

Figure 1. Generation and Ranking Procedure for OOV.

In Step (1b), elements in *CandList* are of the form (c, <*source word*, *target N-gram*>, P (*target N-gram*|c) · P (c|*target N-gram*)). Bidirectional conditional probabilities P (*target N-gram*|c) and P (c|*target N-gram*) are trained on large parallel corpus aligned on the level of constituents instead of words.

Although the translation scope of an OOV word is much smaller than that of a whole sentence, re-ordering can still occur between constituents' translations. Hence, both straight and inverted cases of translation combination are considered (Step (2)). During candidate generation, *Straight* and *Inverted* iteratively cover the span of the OOV word (Step (3)) collecting constituent translations and multiplying translations' scores at the same time. For each assembled translation candidate *TransCand*, its lexical translation score is estimated by the product of bidirectional conditional probabilities of the sublexical translation pairs as:

$$P_{trans} = \sqrt[|o|]{\prod_{c_i \in o} p(c_i \mid target\ N-gram_{ij}) \cdot P(target\ N-gram_{ij} \mid c_i)}$$

where c_i stands for constituent of the OOV word and *target N-gram*$_{ij}$ for one of the sublexical translations for c_i composing *TransCand*.

Note that apart from bilingual information, we further leverage monolingual information on target-language side to estimate the translation candidates (Step 4), using

$$Score(TransCand) = P_{trans}(TransCand)^{\lambda_1} \cdot P_{TLM}(TransCand)^{\lambda_2}$$

where λ_i is the combination weight and TLM stands for target-language language model.

4.2 Method for Singletons

For a singleton-annotated sentence, we plan to translate the contexts with singletons prior to the phrase-based MT systems. More specifically, we first automatically annotate the source sentences with singleton information, exploit skipped phrase table to translate the annotated patterns like "w_1 [s] w_2" and "w_1 [s] [s'] w_2" where w_i stands for a word and "[s]" and "[s']" for the singletons, finalized with submitting the partially translated sentences to the underlying MT decoders such as Moses.

Methods in Section 4.1 and Section 4.2 are proposed as pre-translation modules for phrase-based MT systems without changing the internal components in the systems. The state-of-the-art open-source MT system, Moses, provides XML markup language for us to easily incorporate our models' translation results in OOVs and singletons. Secondly, we generated binary on-demand version of reordering and phrase tables to same time at run-time translation. We also omitted the step of tuning decoding feature weights for quick system set-up. And our baseline achieved around 20.82 BLEU scores on the NTCIR 2011 development set.

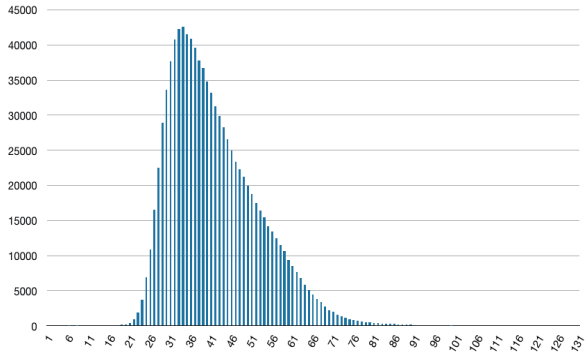

Figure 2. The number of training sentences (y-axis) with respect to their lengths (x-axis).

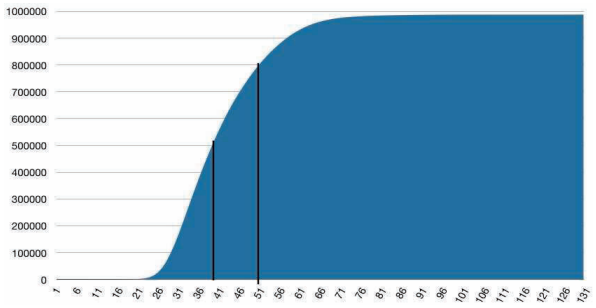

Figure 3. Accumulated number of sentences at different lengths.

5. SYSTEM SETUP AND ENVIRONMENT

We built up Moses (version 2009/08/31 svn checkout, packaged by "The Ubuntu NLP Repository v6.06" at http://cl.naist.jp/~eric-n/ubuntu-nlp/dists/dapper/all/) on top of the stable version of Debian GNU/Linux. Our MT decoder ran under a two quad-core XEON 2.49GHz and eight gigabyte ECC RAM. We set up our baseline system following most of the data preparation and training steps of Baseline System 2 as published on the NTCIR 2011 workshop website except for the following. In the stage of data preparation, we filtered out sentences longer than 50 words instead of 40 words to include more training data (from 477,596 to 754,315 training sentences). See Figure 2 and 3 for the statistics on the lengths of training sentences.

In the task of patent translation (Goto et al., 2011), we found that many of the OOV cases are name entities and domain-specific terms or terminologies. We further designed an MT decoder for these two types of OOVs. Such decoder was trained on Chinese-English Terminology Dictionary from National Institute for Compilation and Translation of Taiwan (NICT) *or* English-Chinese title pairs crawled from ip.com, excluding patents published in 2006-2007 as required.

We manually evaluated 48 OOVs of the NIST 2011 development data with the measurement below. Table 5 shows twelve these OOV examples translated by the additional decoder trained on two different knowledge sources: NICT and ip.com. Generally speaking, the OOV decoder trained on NICT underperformed the one on ip.com data in which of the 48 OOVs 31.2% was evaluated as score 1, 33.3% as 0.5, and 35.4% as 0.

Score	Criteria for the score
0	Incorrect
0.5	Partially correct
1	Correct/near-correct

Table 5. Twelve example OOVs translated by additional decoder.

OOV	NICT	ip.com	reference	Max of score
阔叶	broad leaved	broad leaf	broadleaf	1
流通型	flow through type	flow type	flow-through	0.5
如此之多	thusness of multiple	so between multiple	such a size	0
井涌	well bore	well surge	the kick	0
肽立即	肽 immediate	peptide immediate	peptide is immediately	0.5
机外	external	machine	outside the generator	0
冷机	cooling machine	cooling machine	engine cold state	0
脑靶	cerebral target	brain target	brain target	1
相应帧	phase response frame	correspondin g frame	correspondin g frame	1
黄化	etiolation	yellowing	yellow	0.5
铂络	platium network	platinum complex	platium	0.5

联结环	coupling link	coupling ring	coupling ring	1

6. SUMMARY

In summary, we have examined two of the important issues facing the phrase-based machine translation systems: OOVs and singletons. We describe two methods for dealing with OOVs, one as a preprocessing component, i.e. partial-matched OOV translation model for combination form, and the other as a post-processing one, i.e. fully-fledged OOV decoder, (with an underlying phrase-based MT decoder such as Moses), and one methods for handling singletons. We have analyzed these issues and our proposed methods in the translation tasks of NIST 2006, NIST 2008, and NTCIR 2011. We look forward to bettering our system using hierarchical-like patterns for singletons and sentence domain classifier for domain-specific MT translation and language models.

7. ACKNOWLEDGMENTS

This work was conducted under the "Project Digital Convergence Service Open Platform" of the Institute for Information Industry, Taipei, Taiwan, which is subsidized by the Ministry of Economic Affairs of the Republic of China.

8. REFERENCES

[1] Yunbo Cao and Hang Li. 2002. Base Noun Phrase translation using web data and the EM algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1-7.

[2] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of the 9th NTCIR Workshop*.

[3] Hany Hassan and Jeffrey Sorensen. 2005. An IntegratedApproach for Arabic-English Named Entity Translation. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 87-93.

[4] Kevin Knight and Jonathan Graehl. 1997. Machine Transliteration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 128-135.

[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

[6] Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Nature Language Processing*, pages 381-390.

[7] Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language Entailment Modeling for Translating Unknown Terms. In *Proceedings of the 47th Annual Meeting of ACL and the 4th IJCNLP of the AFNLP*, pages 791–799.

[8] David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the ACL*, pages 189-196.