Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation

Nizar Habash
Center for Computational Learning Systems
Columbia University
habash@ccls.columbia.edu

Abstract

We present four techniques for online handling of Out-of-Vocabulary words in Phrase-based Statistical Machine Translation. The techniques use spelling expansion, morphological expansion, dictionary term expansion and proper name transliteration to reuse or extend a phrase table. We compare the performance of these techniques and combine them. Our results show a consistent improvement over a state-of-the-art baseline in terms of BLEU and a manual error analysis.

1 Introduction

We present four techniques for *online* handling of Out-of-Vocabulary (OOV) words in phrase-based Statistical Machine Translation (SMT).[1] The techniques use morphological expansion (MORPHEX), spelling expansion (SPELLEX), dictionary word expansion (DICTEX) and proper name transliteration (TRANSEX) to reuse or extend phrase tables online. We compare the performance of these techniques and combine them. We work with a standard Arabic-English SMT system that has been already optimized for minimizing data sparsity through the use of morphological preprocessing and orthographic normalization. Thus our baseline token OOV rate is rather low (average 2.89%). None of our techniques are specific to Arabic and all can be retargeted to other languages given availability of technique-specific resources. Our results show that we improve over a state-of-the-art baseline by over 2.7% (relative BLEU score) and handle *all* OOV instances. An error analysis shows that, in 60% of the time, our OOV handling successfully produces acceptable output. Additionally, we still improve in BLEU score even as we increase our system's training data by 10-fold.

[1]This work was funded under the DARPA GALE program, contract HR0011-06-C-0023.

2 Related Work

Much work in MT has shown that orthographic and morpho-syntactic preprocessing of the training and test data reduces data sparsity and OOV rates. This is especially true for languages with rich morphology such as Spanish, Catalan, and Serbian (Popović and Ney, 2004) and Arabic (Sadat and Habash, 2006). We are interested in the specific task of *online OOV handling*. We will not consider *solutions* that game precision-based evaluation metrics by deleting OOVs. Some previous approaches anticipate OOV words that are potentially morphologically related to in-vocabulary (INV) words (Yang and Kirchhoff, 2006). Vilar et al. (2007) address spelling-variant OOVs in MT through online re-tokenization into letters and combination with a word-based system. There is much work on name transliteration and its integration in larger MT systems (Hassan and Sorensen, 2005). Okuma et al. (2007) describe a dictionary-based technique for translating OOV words in SMT. We differ from previous work on OOV handling in that we address spelling and name-transliteration OOVs in addition to morphological OOVs. We compare these different techniques and study their combination. Our morphology expansion technique is novel in that we automatically learn which source language morphological features are irrelevant to the target language.

3 Out-of-Vocabulary Words in Arabic-English Machine Translation

Arabic Linguistic Issues Orthographically, we distinguish three major challenges for Arabic processing. First, Arabic script uses *optional* vocalic diacritics. Second, certain letters in Arabic script are often spelled inconsistently, e.g., variants of Hamzated Alif, أ Â[2] or إ Ă, are often written without

[2]Arabic transliteration is provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007).

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 57–60,
Columbus, Ohio, USA, June 2008.

Hamza: أ *A*. Finally, Arabic's alphabet uses *obligatory* dots to distinguish different letters (e.g., ب *b*, ت *t* and ث *θ*). Each letter base is ambiguous two ways on average. Added or missing dots are often seen in spelling errors. Morphologically, Arabic is a rich language with a large set of morphological features such as gender, number, person and voice. Additionally, Arabic has a set of very common clitics that are written attached to the word, e.g., the conjunction و+ *w+* 'and'. We address some of these challenges in our baseline system by removing all diacritics, normalizing Alif and Ya forms, and tokenizing Arabic text in the highly competitive Arabic Treebank scheme (Sadat and Habash, 2006). This reduces our OOV rates by 59% relative to raw text. So our baseline is a real system with 2.89% token OOV rate. The rest of the challenges such as spelling errors and morphological variations are addressed by our OOV handling techniques.

Profile of OOV words in Arabic-English MT In a preliminary study, we manually analyzed a random sample of 400 sentences containing at least one OOV token extracted from the NIST MTEval data sets. There were 686 OOV tokens altogether. 40% of OOV cases involved proper nouns. 60% involved other parts-of-speech such as nouns (26.4%), verbs (19.3%) and adjectives (14.3%). The proper nouns seen come from different origins including Arabic, Hebrew, English, French, and Chinese. In many cases, the OOV words were less common morphological variants of INV words, such as the nominal dual form. The different techniques we discuss in the next section address these different issues in different ways. Proper name transliteration is primarily handled by TRANSEX. However, an OOV with a different spelling of an INV name can be handled by SPELLEX. Morphological variants are handled primarily by MORPHEX and DICTEX, but since some morphological variations involve small changes in lettering, SPELLEX may contribute too.

4 OOV-Handling Techniques

Our approach to handling OOVs is to extend the phrase table with possible translations of these OOVs. In MORPHEX and SPELLEX techniques, we match the OOV word with an INV word that is a possible variant of the OOV word. Phrases associated with the INV token in the phrase table are "recycled" to create new phrases in which the INV word is replaced with the OOV word. The translation weights of the INV phrase are used as is in the new phrase. We limit the added phrases to source-language unigrams and bigrams (determined empirically). In DICTEX and TRANSEX techniques, we add completely new entries to the phrase table. All the techniques could be used with other approaches, such as input-text lattice extension with INV variants of OOVs or their target translations. We briefly describe the techniques next. More details are available in a technical report (Habash, 2008).

MORPHEX We match the OOV word with an INV word that is a possible *morphological* variant of the OOV word. For this to work, we need to be able to morphologically analyze the OOV word (into lexeme and features). OOV words that fail morphological analysis cannot be helped by this technique. The morphological matching assumes the words to be matched agree in their lexeme but have different inflectional features. We collect information on possible inflectional variations from the original phrase table itself: in an off-line process, we cluster all the analyses of single-word Arabic entries in our phrase table that (a) translate into the same English phrase and (b) have the same lexeme analysis. From these clusters we learn which morphological inflectional features in Arabic are irrelevant to English. We create a rule set of morphological inflection maps that we then use to relate analyses of OOV words to analyses of INV words (which we create off-line for speedy use). The most common inflectional variation is the addition or deletion of the Arabic definite article ال+ *Al+*, which is part of the word in our tokenization.

SPELLEX We match the OOV token with an INV token that is a possible correct spelling of the OOV token. In our current implementation, we consider four types of spelling correction involving one letter only: letter deletion, letter insertion, letter inversion (of any two adjacent letters) and letter substitution. The following four misspellings of the word فلسطيني *flsTyny* 'Palestinian' correspond to these four types, respectively: فلسطني *flsTny*, فلسطيننني *flsTynny*, فلطسيني *flTsyny* and قلسطيني *qlsTyny*. We only allow letter substitution from a limited list of around 90 possible substitutions (as opposed to all 1260 possible substitutions). The substitutions we considered include cases we deemed harder than

usual to notice as spelling errors: common letter shape alternations (e.g., ر *r* and ز *z*), phonological alternations (e.g., ص *S* and س *s*) and dialectal variations (e.g., ق *q* and ىء ŷ). We do not handle misspellings involving two words attached to each other or multiple types of single letter errors in the same word.

DICTEX We extend the phrase table with entries from a manually created dictionary – the English glosses of the Buckwalter Arabic morphological analyzer (Buckwalter, 2004). For each analysis of an OOV word, we expand the English lemma gloss to all its possible surface forms. The newly generated pairs are equally assigned very low translation probabilities that do not interfere with the rest of the phrase table.

TRANSEX We produce English transliteration hypotheses that assume the OOV is a proper name. Our transliteration system is rather simple: it uses the transliteration similarity measure described by Freeman et al. (2006) to select a best match from a large list of possible names in English.[3] The list was collected from a large collection of English corpora primarily using capitalization statistics. For each OOV word, we produce a list of possible transliterations that are used to add translation pair entries in the phrase table. The newly generated pairs are assigned very low translation probabilities that do not interfere with the rest of the phrase table. Weights of entries were modulated by the degree of similarity indicated by the metric we used. Given the large number of possible matches, we only pass the top 20 matches to the phrase table. The following are some possible transliterations produced for the name باستور *bAstwr* together with their similarity scores: pasteur and pastor (1.00), pastory and pasturk (0.86) bistrot and bostrom (0.71).

5 Evaluation

Experimental Setup All of our training data is available from the Linguistic Data Consortium (LDC).[4] For our basic system, we use an Arabic-English parallel corpus[5] consisting of 131K sentence pairs, with approximately 4.1M Arabic tokens and 4.4M English tokens. Word alignment is done with GIZA++ (Och and Ney, 2003). All evaluated systems use the same surface trigram language model, trained on approximately 340 million words from the English Gigaword corpus (LDC2003T05) using the SRILM toolkit (Stolcke, 2002). We use the standard NIST MTEval data sets for the years 2003, 2004 and 2005 (henceforth MT03, MT04 and MT05, respectively).[6]

We report results in terms of case-insensitive 4-gram BLEU (Papineni et al., 2002) scores. The first 200 sentences in the 2002 MTEval test set were used for Minimum Error Training (MERT) (Och, 2003). We decode using Pharaoh (Koehn, 2004). We tokenize using the MADA morphological disambiguation system (Habash and Rambow, 2005), and TOKAN, a general Arabic tokenizer (Sadat and Habash, 2006). English preprocessing simply included down-casing, separating punctuation from words and splitting off "'s".

OOV Handling Techniques and their Combination We compare our baseline system (BASELINE) to each of our basic techniques and their full combination (ALL). Combination was done by using the union of all additions. In each setting, the extension phrases are added to the baseline phrase table. Our baseline phrase table has 3.5M entries. In our experiments, on average, MORPHEX handled 60% of OOVs and added 230 phrases per OOV; SPELLEX handled 100% of OOVs and added 343 phrases per OOV; DICTEX handled 73% of OOVs and added 11 phrases per OOV; and TRANSEX handled 93% of OOVs and added 16 phrases per OOV.

Table 1 shows the results of all these settings. The first three rows show the OOV rates for each test set. $\text{OOV}_{sentence}$ indicates the ratio of sentences with at least one OOV. The last two rows show the best absolute and best relative increase in BLEU scores above BASELINE. All conditions improve over BASELINE. Furthermore, the combination improved over BASELINE and its components. There is no clear pattern of technique rank across all test sets. The average increase in the best performing conditions is around 1.2% BLEU (absolute) or 2.7% (relative). These consistent improvements are not statistically significant. However, this is still a nice

[3] Freeman et al. (2006) report 80% F-score at 0.85 threshold.

[4] http://www.ldc.upenn.edu

[5] The parallel text includes Arabic News (LDC2004T17), eTIRR (LDC2004E72), Arabic Treebank with English translation (LDC2005E46), and Ummah (LDC2004T18).

[6] The following are the statistics of these data sets in terms of (sentences/tokens/types): MT03 (663/18,755/4,358), MT04 (1,353/42,774/8,418) and MT05(1,056/32,862/6,313). The data sets are available at http://www.nist.gov/speech/tests/mt/.

Table 1: OOV Rates (%) and BLEU Results of Using Different OOV Handling Techniques

	MT03	MT04	MT05
$\text{OOV}_{sentence}$	40.12	54.47	48.30
OOV_{type}	8.36	13.32	11.38
OOV_{token}	2.46	3.21	2.99
BASELINE	44.20	40.60	42.86
MORPHEX	44.79	41.18	43.37
SPELLEX	45.09	41.11	43.47
DICTEX	44.88	41.24	43.46
TRANSEX	44.83	40.90	43.25
ALL	**45.60**	**41.56**	**43.95**
Best Absolute	1.40	0.96	1.09
Best Relative	3.17	2.36	2.54

result given that we only focused on OOV words.

Scalability Evaluation To see how well our approach scales up, we added over 40M words (1.6M sentences) to our training data using primarily the UN corpus (LDC2004E13). As expected, the token OOV rates dropped from an average of 2.89% in our baseline to 0.98% in the scaled-up system. Our average baseline BLEU score went up from 42.60 to 45.00. However, using the ALL combination, we still increase the scaled-up system's score to an average BLEU of 45.28 (0.61% relative). The increase was seen on all data sets.

Error Analysis We conducted an informal error analysis of 201 random sentences in MT03 from BASELINE and ALL. There were 95 different sentences containing 141 OOV words. We judged words as *acceptable* or *wrong*. We only considered as *acceptable* cases that produce a correct translation or transliteration *in context*. Our OOV handling successfully produces acceptable translations in 60% of the cases. Non-proper-noun OOVs are well handled in 76% of the time as opposed to proper nouns which are only correctly handled in 40% of the time.

6 Conclusion and Future Plans

We have presented four techniques for handling OOV words in SMT. Our results show that we consistently improve over a state-of-the-art baseline in terms of BLEU, yet there is still potential room for improvement. The described system is publicly available. In the future, we plan to improve each of the described techniques; explore better ways of weighing added phrases; and study how these techniques function under different tokenization conditions in Arabic and with other languages.

References

T. Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium (LDC2004L02).

A. Freeman, S. Condon, and C. Ackerman. 2006. Cross Linguistic Name Matching in English and Arabic. In *Proc. of HLT-NAACL*.

N. Habash. 2008. Online Handling of Out-of-Vocabulary Words for Statistical Machine Translation. CCLS Technical Report.

N. Habash, A. Soudi and T. Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Springer.

N. Habash and O. Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proc. of ACL'05*.

H. Hassan and J. Sorensen. 2005. An integrated approach for Arabic-English named entity translation. In *Proc. of the ACL Workshop on Computational Approaches to Semitic Languages*.

P. Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proc. of AMTA*.

F. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52.

F. Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of ACL*.

H. Okuma, H. Yamamoto, and E. Sumita. 2007. Introducing translation dictionary into phrase-based SMT.. In *Proc. of MT Summit.*

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*.

M. Popović and H. Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proc. of LREC.*

F. Sadat and N. Habash. 2006. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proc. of ACL*.

A. Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proc. of ICSLP*.

D. Vilar, J. Peter, and H. Ney 2007. Can we translate letters?. In *Proc. of ACL workshop on SMT*.

M. Yang and K. Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proc. of EACL*.