# Automatic Learning of Morphological Variations for Handling Out-of-Vocabulary Terms in Urdu-English Machine Translation

**Nizar Habash**
Center for Computational Learning Systems
Columbia University
475 Riverside Dr, New York, NY 10115
habash@ccls.columbia.edu

**Hayden Metsky**
Millburn High School
462 Millburn Ave, Millburn, NJ 07041
hmetsky@gmail.com

## Abstract

We present an approach for online handling of Out-of-Vocabulary (OOV) terms in Urdu-English MT. Since Urdu is morphologically richer than English, we expect a large portion of the OOV terms to be Urdu morphological variations that are irrelevant to English. We describe an approach to automatically learn English-irrelevant (target-irrelevant) Urdu (source) morphological variation rules from standard phrase tables. These rules are learned in an unsupervised (or lightly supervised) manner by exploiting redundancy in Urdu and collocation with English translations. We use these rules to hypothesize invocabulary alternatives to the OOV terms. Our results show that we reduce the OOV rate from a standard baseline average of 2.6% to an average of 0.3% (or 89% relative decrease). We also increase the BLEU score by 0.45 (absolute) and 2.8% (relative) on a standard test set. A manual error analysis shows that 28% of handled OOV cases produce acceptable translations in context.

## 1 Introduction

The problem of Out-of-Vocabulary (OOV) terms is a common theme in many NLP applications, especially automatic speech recognition (ASR) and machine translation (MT). Habash (2008) points out that low token OOV rates can be quite deceptive since they affect a significant proportion of the sentences in MT. For example, a 3% token OOV rate can negatively affect the fluency and accuracy of 40% of all sentences.

We are interested in the specific task of *online OOV Handling* as a way to address terms that are not modeled in the offline built MT system. We work with a standard phrase-based MT system on Urdu-English MT. Since Urdu is morphologically richer than English, we expect a large portion of the OOV terms to be Urdu morphological variants that are irrelevant to English. In this paper we describe an approach to automatically learn English-irrelevant (target-irrelevant) Urdu (source) morphological variation rules from standard phrase tables. These rules are learned in an unsupervised (or lightly supervised) manner by exploiting redundancy in Urdu and collocation with English translations. We use these rules to hypothesize invocabulary (INV) alternatives to the OOV terms. We then use phrases associated with the INV terms to add to the phrase table additional phrases in which we replace the INV term with its corresponding OOV term. Our results show that we reduce the OOV rate from a standard baseline average of 2.6% to an average of 0.3% (or 89% relative decrease). We also increase the BLEU score by 0.45 (absolute) and 2.8% (relative) on a standard test set. A manual error analysis shows that 28% of handled OOV cases produce acceptable translations in context. We also present additional results comparing and combining this technique with two other techniques that target proper names and spelling errors.

This paper is structured as follows. Section 2 presents previous related research. Section 3 presents some relevant background on Urdu linguistics and profiles specific problems for Urdu-English MT. Section 4 describes our baseline MT system. Section 5 details the morphology variation rule learning approach and discusses the different types of learned rules. Section 6 presents our system evaluation and results.

## 2 Related Work

The work presented in this paper is in the intersection of multiple active areas of research. In particular we briefly describe three areas: unsupervised multilingual learning of morphology, OOV handling in Machine Translation and Urdu NLP.

**Unsupervised Multilingual Morphology Learning** Snyder and Barzilay (2008) describe an approach for unsupervised learning of cross-lingual morphological segmentation using parallel corpora for three Semitic languages (Arabic, Hebrew and Aramaic) and English. Their models jointly induce morpheme boundaries for the studied languages and identified cross-lingual morpheme patterns. Their work overlaps research in unsupervised morphology learning and research in multilingual learning. Much research has been done in multilingual learning to build tools exploiting parallel data from morphology to word sense tagging (Yarowsky et al., 2001; Diab and Resnik, 2002; Rogati et al., 2003). Research in unsupervised morphological learning explores ways of deriving morphology information from redundancy in the data (Goldsmith, 2001; Creutz and Lagus, 2007).

**OOV Handling in Machine Translation** 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 (Lee, 2004; Habash and Sadat, 2006). We are interested here in the specific task of *on-line OOV handling*. The most common solution for such OOV words is to delete them from the output – thus gaming precision-based evaluation metrics such as BLEU (Papineni et al., 2002). We will not consider this "solution." Some previous approaches anticipate OOV words that are potentially morphologically related to in-vocabulary (INV) words. For example, Yang and Kirchhoff (2006) extend phrase tables with back-off phrase variants that are segmented into smaller morphological units. Test data OOV terms are segmented in a similar manner. Talbot and Osborne (2006) propose a language-independent approach for modeling lexical redundancy for MT. They use this approach to smooth phrase-based translation models. Their approach does not target OOVs in particular, but clearly helps address many OOV cases. Vilar et al. (2007) address spelling-variant OOVs in MT through on-line re-tokenization into letters and combination with a word-based system. Habash (2008) compares and combines four techniques for online handling of Out-of-Vocabulary words in Arabic-English phrase-based MT. The techniques used are spelling expansion, morphological expansion, dictionary term expansion and proper name transliteration. The techniques are used to extend the phrase table with recycled or novel phrases. His results show a consistent improvement over a state-of-the-art baseline in terms of BLEU and a manual error analysis.

**Urdu NLP** Relative to other languages with similar populations, Urdu has not received a lot of attention (Hussain, 2004b). A close sister language of Urdu, Hindi, has received relatively more attention. One particular publication on Hindi is relevant here as it explores similar issues: Mahesh and Sinha (2007) exploit rich morphology in Hindi to handle translation divergences between Hindi and English in a rule-based MT approach. One of the earlier papers we could find on Urdu and MT is by Jones and Havrilla (1998), in which they described a formalism for learning transfer rules for Urdu-English MT. Humayoun (2006) describes a suite of resources for Urdu processing and Hussain (2004a) discusses in great details the workings of a morphological analyzer for Urdu. In 2008, Urdu was chosen as one of languages from which to translate into English in the National Institute of Standards and Technology (NIST) MT Evaluation competition.[1] We use the data provided by NIST in this paper and report results on its development and test sets.

In this paper, we describe and evaluate an approach to on-line OOV handling in the context of Urdu-English MT using automatically learned morphological variation rules learned in an unsupervised manner from multilingual data (specifically phrase tables extracted from automatically aligned parallel data, which are arguably "lightly supervised"). The morphological rules learned cluster morphological phenomena in the source language (Urdu) that are

[1] http://www.nist.gov/speech/tests/mt/2008/doc/

not relevant to the target language (English). By relating an OOV term to an INV term using one of these rules, we can expand existing phrase tables with "recycled phrases" of the INV terms. This approach is similar to Habash (2008)'s work on Arabic online OOV handling except that unlike his work on morphological expansion which required a morphological analyzer, we do not need one; instead we learn the morphology mapping automatically. As such, we restrict ourselves to not using any of the existing morphological analyzers for Urdu (Hussain, 2004a; Humayoun, 2006). The work of Snyder and Barzilay (2008) is close to our work; however, unlike them, we are asymmetrically interested in modeling aspects of one language (Urdu) that are irrelevant to the other language (English). We expect our approach to be more useful for morphologically rich source languages being translated to morphologically poor languages. Our work is closer to Talbot and Osborne (2006), except in that they do not target OOVs in particular. The features they learn to determine lexical-redundancy cluster membership are similar to the rules we learn in this paper. Finally, we differ in general from previous work in multilingual learning and morphology learning in that we work on Urdu and in that we use and evaluate our rules for the task of OOV handling.

## 3 Urdu Linguistic Challenges

Urdu is the official language of Pakistan and one of India's 23 languages. Despite being spoken by over 60M native speakers and over 100M second language speakers, Urdu has only recently started to receive computational attention. Urdu is an Indo-European language from the Indo-Iranian branch. Urdu is known to closely resemble Hindi (forming together what is sometimes called "Hindustani"). However, Urdu differs from Hindi in that it is written in an extended form of the Arabic script and in that it shows a lot of influences from Persian (another Indo-European language) and Arabic (a Semitic language) compared to Hindi.

In this section we discuss the orthographic and morphological challenges for computational processing of Urdu. For a much more detailed discussion of Urdu orthography and morphology from a computational point of view, see (Hussain, 2004a; Humayoun, 2006). We will not discuss syntactic issues in this work. We also present a preliminary analysis of the types of OOVs seen in Urdu to further motivate our work.

### 3.1 Urdu Orthography

Urdu is written using an extended version of the right-to-left context-sensitive Perso-Arabic alphabet consisting of 44 basic letter forms and 15 optional diacritical marks (Humayoun, 2006). The following are some of the prominent challenges for Urdu orthography.

- **Diacritics** As in Arabic, diacritics are often not written in Urdu. Diacritics' general absence adds to the ambiguity challenge of translating from Urdu to English. For example, the word بن is ambiguous depending on its vowelization as the noun بِن *bin* 'son' or the verb بَن *ban* 'make'.
- **Letter Marks** Arabic's alphabet uses *obligatory* marks (typically dots) to distinguish different letters (e.g., ب *b*, ت *t* , ث *θ*, پ *p* and ٹ *ṯ*).[2] This is different from using diacritics. The number of basic letter forms is 18, less than half the number of letters. As such there is a high likelihood that spelling errors involving these marks take place.
- **Disconnective Letters** Although the Arabic script is a mostly connective cursive script, there are a few letters that do not connect to the letters that follow them: آ *Ā* ا *A* ر *r* ز *z* ژ *ž* د *d* ذ *ð* و *w*. This leads to the presence of a tiny word-internal space that sometimes is confused for a word separator. As a result, some

[2] All Arabic script transliterations are provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007). This scheme extends Buckwalter's transliteration scheme (Buckwalter, 2002) to increase its readability while maintaining the 1-to-1 correspondence with the orthography as represented in Unicode. The following are the only differences from Buckwalter's scheme (which is indicated in parentheses): آ Ā (|), أ Â (>), ؤ ŵ (&), إ Ă (<), ئ ŷ (}), ة ħ (p), ث θ (v), ذ ð (*), ش š ($), ظ Ď (Z), ع ς (E), غ γ (g), ى ý (Y), ـٔ ã (F), ـٌ ũ (N), ـٍ ĩ (K). For Urdu-specific extensions of the Arabic script, we extend the Habash-Soudi-Buckwalter transliteration scheme as follows: ہ *h*, ھ *h*, ٹ *t*, ڈ *d*, ڑ *r*, ں *n*, ے *y*, پ *p*, ژ *ž*, چ *č*, گ *g*.

words may be broken into two parts or more; and incorrect words are made up of two or more words. See the first spelling error example in Table 1.

- **Homophone Letters** Given that Urdu has a lot of borrowings from Arabic that retain their Arabic spelling even though they change their pronunciation, there are cases of spelling errors of Arabic words spelled as pronounced in Urdu. See the second spelling error example in Table 1, which is the result of the two letters ذ ð and ز *z* having the same pronunciation in Urdu /z/.

We do not address the diacritization issue in this work. And although we also do not address spelling directly, we in fact address some of the spelling cases indirectly as they are confusable with source morphological variations in which a letter is substituted for another without an effect on the target-language.

### 3.2 Urdu Morphology

Urdu is a weakly inflected language with multiple productive inflection/derivation morphological mechanisms that reflect the different language-origins of its words. For instance, although Urdu is primarily a suffixation language, Arabic templatic morphology also appears in Urdu, e.g., broken plurals: the plural of جزیرہ *jzyrh* 'island' is جزائر *jzAŷr* 'islands'. Another example highlighting this complexity is the presence of multiple productive feminine morphemes: for words of Arabic origin, it is ہ+ +*h*, e.g., والد *wAld* 'father' becomes والدہ *wAldh* 'mother'; however, the productive feminine morpheme for words of Hindi origin is ی+ +*y*, e.g., لڑکا *lṛkA* is 'boy' but لڑکی *lṛky* is 'girl'.

Urdu nouns inflect in gender (masculine and feminine), number (singular and plural) and case (nominative [NOM], oblique [OBL] and vocative [VOC]). For example, the plural of the Urdu word کتاب *ktAb* 'book' has three case-variant forms: کتابیں *ktAbyn* (NOM) کتابوں *ktAbwn* (OBL) and کتابو *ktAbw* (VOC). The oblique case is further modified syntactically using a variety of post-positions leading to a total of nine (extended) cases: nominative, oblique, vocative, ergative (OBL+نے *ny*), accusative (OBL+کو *kw*), dative (OBL+کو/کے *kw/ky*), instrumental (OBL+سے *sy*), genitive (OBL+کا/کی/کے *kA/ky/ky*) and locative (OBL+میں/پر/تلے/... *myn/pr/tly/...*). Post-positions are typically written separate from the word whose function they specify, but often, due to orthography features discussed above, the post-positions are attached to the word, effectively extending its ortho-morphology.

Urdu verbs morphologically inflect in mood-aspect-tense (infinitive, subjunctive, perfective and imperfective), person (first, second, third), gender (masculine and feminine) and number (singular and plural). With few exceptions, conjugation is very regular. Urdu also has productive suffixes that generate so-called causatives and double causatives from basic verbs. For example, the root بن *bn* in the basic verb بننا *bn**n**A* 'to make (unaccusatively *be made*)',[3] can be extended with the suffix ا + +*A* to create بنانا *bn**A**nA* 'to make by self (direct causative)' and also with the suffix وا+ +*wA* to create بنوانا *bn**w**A**n**A* 'to make through another person (indirect causative)'.

Much of these inflectional variations are just "noise" from the point of view of English but some are not. In the work presented here we attempt to automatically learn the patterns of what English is truly blind to and what it is not.

### 3.3 Preliminary Analysis of Urdu OOVs

To understand the kind of phenomena we need to handle when solving the OOV problem in Urdu-English MT, we took a sample of 100 sentences (1,778 words) from our baseline system and classified the OOV tokens in it. 48 OOV cases (2.7% of words) appeared in 37 sentences (37% of sentences). The OOV cases are (a.) spelling errors (19 case or 39.6%), (b.) morphology variants unseen in data (18 cases or 37.5%) and (c.) proper nouns requiring transliteration (11 cases or 22.9%). We exemplify these three classes in Table 1. In this work, we primarily address morphology issues (almost two-fifths of all OOVs) and we touch on spelling issues and transliteration of proper nouns in as much as these could be interpreted as morphological variants from the point of view of English. In Section 6.3, we present two techniques for handling spelling er-

[3] نا+ +*nA* is the infinitive marker.

| Class | Urdu | English |
|---|---|---|
| Proper Noun | ہوگرڈ *hwGrd* | Hoggard |
| | جھارکھنڈ *jhArkhnd* | Jharkhand |
| Morphology | بیویاں *bywyAn* related to | wives (pl) |
| | بیوی *bywy* | wife (sg) |
| | برقعے *brqςy* related to | veil (obl) |
| | برقع *brqς* | veil (nom) |
| Spelling | ساجدکو *sAjdkw* incorrect form of: ساجد کو *sAjd kw* | SAjid (dat) |
| | مذاحمت *mðAHmt* incorrect form of: مزاحمت *mzAHmt* | friction/ resistance |

Table 1: Three classes of OOVs in Urdu-English MT

rors and proper nouns and we compare them to and combine them with the work on morphological variations.

## 4 Urdu-English Baseline MT

In this section, we describe our baseline Urdu-English MT system.

### 4.1 Data

The data we use here is restricted to the resource package made available by NIST for their 2008 MT Evaluation (Urdu-English Track). We even follow the restriction to not use any additional monolingual data outside that package. Among other things, the package includes a parallel corpus of Urdu and English, a lexicon of Urdu with English glosses, a morphological analyzer and a transliterator. We only make use of the parallel corpus and the lexicon in this paper.

Although some amount of simple preprocessing was done in the provided data, we still needed to do additional preparation before we could use it. In particular, we use an implementation of Gale and Church (1993)'s sentence alignment algorithm to align the Urdu and English sentences in the parallel corpus. We extend the corpus with paired Urdu-English entries from the lexicon. The presence of the lexicon may positively bias the automatic learning process of the morphological variations, but it is by no means necessary. A random set of 1952 sentences (15K words) is extracted and used for tuning. Training data consists of 253,260 sentences with 1.8M words of English and 1.9M words of Urdu.

### 4.2 Orthographic Preprocessing

Although the encoding of Arabic script and its extensions is standardized in a context-insensitive manner in Unicode,[4] almost all context-sensitive glyphs are possible to use directly.[5] This is not recommended; however, it is sometimes done. As a result, the same appearance of a word on screen/page may be implemented using different sequences of letters. This has the effect of increasing sparsity for any natural language processing system. We address this issue through a special cleaning step that collapses the various glyphs into their correct letter form encoding.

In addition, we remove all kashidas (elongation markers in Arabic script) and all diacritics. We also collapse the two forms of Heh in Urdu (ہ *h* and ھ *h*), the two different forms of Yeh (ی *y* and ے *y*) and the two different forms of Nuun (ن *n* and ں *n*). The decisions to collapse these forms were empirically determined using a development set on which we received around a 5% relative improvement in BLEU score. The collapse of some of these characters helps reduce variations resulting from spelling errors, but also from morphological alternatives. In particular the two different forms of Yeh can be morphologically distinctive in Urdu.

English preprocessing simply includes downcasing, separating punctuation from words and splitting off "'s".

### 4.3 Building the Phrase-based MT Baseline

We built our baseline system using *standard* resources for phrase-based MT. Word alignment is done with GIZA++ (Och and Ney, 2003). Phrase table extraction and decoding are done using resources from the Pharaoh system suite (Koehn, 2004). Tuning was done use Och's Minimum Error Training (MERT) method (Och, 2003). A trigram English language model was implemented using the SRILM

[4] http://unicode.org/charts/PDF/U0600.pdf
[5] http://unicode.org/charts/PDF/UFB50.pdf

toolkit (Stolcke, 2002) applied to the English side of the training data.

Section 6 contains the evaluation results for the baseline system.

## 5 Automatic Learning of Morphology Variation Rules

Our basic approach for handling OOVs using morphology information is as follows: we match the OOV token with an INV token that is a possible morphological variant of the OOV token. Then, phrases associated with the INV token in the phrase table are used to create new phrases in which the INV token is replaced with the OOV token. For this approach to work, we only allow mappings that are "noise" from the point of view of English. For example, case-variant forms of an Urdu noun are all interchangeable. We describe next how we learn these morphology variation rules. Then we present an analysis of the different types of learned rules.

### 5.1 Learning Urdu Morphology Variation Rules

We collect information on possible inflectional variations from the original phrase table itself. In an offline process, we cluster all the Urdu phrases with single word entries in our phrase table that translate into the same English phrase. For every two Urdu words, $i$ and $j$ in the same cluster, we try to produce a three-way segmentation into *prefix stem suffix* such that $stem_i$ equals $stem_j$, $stem_i$ is at least one character long, and $stem_i$ and $stem_j$ are in fact the longest shared sub-strings in words $i$ and $j$. Once a segmentation is found, a bidirectional rule of the following form is created (if seen for the first time), and its weight (measured by number of supporting examples) is incremented: $prefix_i$ _ $suffix_i \Leftrightarrow prefix_j$ _ $suffix_j$.

During translation time, an OOV word is matched against all rules (by matching prefix and suffix conditions). Once a match is found, a morphological expansion is created. We check if the expanded form is an INV. If it is not, we ignore it. However, if it is, we copy all the phrases showing the INV word as a singleton entry and replace the INV word with the OOV word. The translation weights of the INV phrase are used as is in the new phrase. In the future we plan to investigate how to modify the weights using the probabilities of the learned rules.

### 5.2 Analysis of Learned Rules

Our system learned 2,274,392 rules (1,137,196 bidirectional rules), which we rank based on redundancy in supporting examples in the data (or weight). There are 123 different unique ranks with a Zipfian distribution showing an expected very long tail: 96.4% of all rules are with 1, 2 or 3 supporting examples only. Only 88 rules (in 43 ranks) have 100 supporting examples or more. Table 2 describes the distribution of rule counts by rank level. We distinguish three classes of rules: Prefixing rules (PRE) involve adding/deleting a prefix or replacing a prefix with another prefix. Suffixing rules (SUF) similarly involve adding/deleting a suffix or replacing a suffix with another suffix. Circumfixing rules (CIRC) include all other possible rules: adding/deleting/replacing circumfixes and all rules mixing prefixes/suffixes and circumfixes. That is, we count a rule replacing a prefix with a suffix as a circumfixing rule. For example, the fifth row in Table 2 says that in the top 50 rank levels, there are 104 rules, 88% of which are suffixing rules, 8% of which are prefixing, and 4% of which are circumfixing.

The shifting distribution of the rules shows a nice consistency with what we know about Urdu morphology: Urdu is a primarily suffixational language. The highest rank SUF rules are all nominal/verbal inflections that are not present in English such as deleting the infinitive marker نا+ *+nA* or replacing the plural nominative suffix یں+ *+yn̲* with the plural oblique suffix وں+ *+wn̲*. The most common SUF rule allows deleting/adding the suffix ی+ *y*, which actually collapses the distinction between the highly ambiguous suffixes ی+ *y* and ے+ *y̲* in our system. The ے *y̲* suffix refers to past masculine plural, non-past third/second person singular, and nominative masculine plural; while the suffix ی+ *y* refers to past feminine singular, feminine singular or adjectival derivations, among others. The top 20 rules (10 bidirectional rules) are presented in Table 3.

Among medium frequency rules, we find examples of spelling correction rules that reflect phonological similarities, e.g., a rule that replaces ذ ð with

ز *z* corrects the word ذیادتی *ðyAdty* to زیادتی *zyAdty* 'excess'. Similarly, some of these spelling correction rules correct shape-based errors such as spelling شاهی *šAhy* 'royal' as ساهی *sAhy*. We also find SUF rules that correct space-spelling errors where a postcilitic is written attached to a word ending with a disconnective letter, e.g., اکتوبرکو *Aktwbrkw* 'in October' becomes اکتوبر *Aktwbr*.

The first CIRC rule is ranked 45th. It refers to a phenomenon called "strengthening", where a word-stem vowel is lengthened as part of a derivational process to create its causative form. The specific rule that leads the CIRC list also adds the infinitive suffix نا+ +*nA*. For example, کٹ *kṯ* 'be cut/lose' کاٹنا becomes *kAṯnA* 'to cut/bite'. This rule, expressed as [k_ ⇔ kA_nA], reflects a current limitation, namely that infixation is not modeled. As a result, the rule is too specific in that it incorrectly encodes part of the basic word stem (*k*) as a prefix.

Among very low frequency rules, we find examples that link words for morphologically meaningless reasons. For instance, the following words are linked to each other through multiple rules since they all map to different senses of the English word 'space' and contain the letter ا 'A' (a purely hypothetical stem): کمرا *kmr**A*** (space as in room, bedroom), اکاش ***A**kAš* (space as in ether, firmament), فاصلہ *f**A**Slh* (space as in distance, break, discontinuation), مقام *mq**A**m* (space as in locality, abode, dwelling), خلاء *xl**A**'* (space as in aerospace, vacuum), and فضا *fĎ**A*** (space as in atmosphere). Some of these rules are PRE or SUF, but most are CIRC.

A large portion of the rules is very noisy as a result of bad alignment, non-inflectional clustering, English-semantic alignments that are not meaningful in Urdu, or the loose definition of "STEM", i.e. *one shared letter or more*, which allows for a lot of implausible and infrequent rules to be generated. In addition, the simple model we use does not allow learning independent rules that can be applied in a hierarchical manner or in a collective manner.

## 6 Evaluation

We report results on the DEVSET set provided in the NIST package. DEVSET includes 4,975 sentences and has one translation reference per sentence. We also report on the NIST MT Eval 2008 official test set (MT08), which has 1,862 sentences with four translation references. We report results in terms of case insensitive 4-gram (standard) BLEU (Papineni et al., 2002) metric scores. Other metrics such as NIST were considered but gave no additional information.

| Rank | Total | PRE | SUF | CIRC |
|---|---|---|---|---|
| 10 | 20 | 0% | 100% | 0% |
| 20 | 40 | 0% | 100% | 0% |
| 30 | 60 | 3% | 97% | 0% |
| 40 | 82 | 7% | 90% | 2% |
| 50 | 104 | 8% | 88% | 4% |
| 60 | 126 | 11% | 86% | 3% |
| 70 | 160 | 18% | 80% | 3% |
| 80 | 200 | 21% | 74% | 5% |
| 90 | 314 | 24% | 66% | 10% |
| 100 | 500 | 27% | 52% | 21% |
| 110 | 1,882 | 28% | 31% | 41% |
| 120 | 38,784 | 19% | 15% | 66% |
| 123 | 2,274,392 | 8% | 8% | 84% |

Table 2: Distribution of different learned-rule types over rank. **Rank** refers to the top *n* rules by amount of supporting evidence. **Total** refers to the actual number of rules in rank level. **PRE**, **SUF** and **CIRC** refer to the percentage of prefixing, suffixing and circumfixing rules, respectively.

In the following section, we present the results of applying different subsets of learned rules. We then present a manual error analysis of the MT output. Finally, we present some additional results comparing the approach we use to other techniques for OOV handling.

### 6.1 Evaluation of Morphology Variation Rules

The results are shown in Table 4. The number of words and OOV words in both DEVSET and MT08 are shown. In addition, the BLEU scores (multiplied by 100) are presented for the BASELINE system when not using any morphology variation rules and when it is supplemented with the top $n$ ranks of rules. With few exceptions, adding more rules corresponds to better performance as measured by BLEU. The column marked as LOOV displays the ratio of leftover OOV words in the output. Using more rules allows more OOV words to be handled. All leftover

| Rank | Rule |
|---|---|
| 1 | _ ⇔ _ y |
| 2 | _ ⇔ _ nA |
| 3 | _ ⇔ _ A |
| 4 | _ ⇔ _ AnA |
| 5 | _ ⇔ _ h |
| 6 | _ A ⇔ _ y |
| 7 | _ ⇔ _ wn |
| 8 | _ ⇔ _ t |
| 9 | _ h ⇔ _ y |
| 10 | _ ⇔ _ ny |

Table 3: Top 10 bidirectional rules learned by our system.

OOVs are kept in the output (not deleted). Overall, we reduce the OOV rate on BASELINE from an average of 2.6% to an average of 0.3% (or 89% relative decrease) and increase the BLEU score on BASELINE by 0.45 (absolute) for MT08 and 0.22 (absolute) for DEVSET – an average relative increase of 2.6%.

### 6.2 Error Analysis

We conducted an error analysis of 100 sentences selected randomly from DEVSET. The sample contains 45 OOV words. We handle all of them except for two. We judge the handled 43 OOV words as *acceptable* or *wrong*. We only consider as *acceptable* cases that produce a correct translation or transliteration *in context*. There are 12 acceptable cases (28%). Given that our approach is unsupervised and does not use any morphological analysis resources (as did (Habash, 2008)), this is a good result for handling words that otherwise are not translated. Two of the *acceptable* cases (17%) are proper nouns and the rest are nouns, adjectives and verbs. Five of the 12 *acceptable* cases (42%) do not match any reference words, i.e., they cannot be captured by BLEU.

Of the 31 *wrong* cases, six (19%) result from deleting the OOV word through mapping it to an INV term whose English translation does not translate the words completely. Such bad phrase table entries are created when phrase extraction faces bad/sparse alignments. In the rest of the *wrong* cases, the decoder made a bad selection. Ten of the 31 *wrong* cases are proper nouns (32%) and the rest are nouns, adjectives and verbs. Overall, there are

| | DEVSET | | MT08 | |
|---|---|---|---|---|
| **Words** | 90454 | | 42196 | |
| **OOV** | 2087 | | 1180 | |
| | BLEU | LOOV | BLEU | LOOV |
| BASELINE | 9.47 | 2.31% | 15.95 | 2.80% |
| $\text{top}_{10}$ | 9.49 | 1.82% | 15.98 | 2.12% |
| $\text{top}_{20}$ | 9.52 | 1.61% | 16.00 | 1.89% |
| $\text{top}_{30}$ | 9.53 | 1.52% | 16.00 | 1.77% |
| $\text{top}_{40}$ | 9.55 | 1.45% | 16.00 | 1.67% |
| $\text{top}_{50}$ | 9.55 | 1.41% | 16.01 | 1.63% |
| $\text{top}_{60}$ | 9.55 | 1.40% | 16.00 | 1.62% |
| $\text{top}_{70}$ | 9.57 | 1.32% | 16.03 | 1.46% |
| $\text{top}_{80}$ | 9.56 | 1.30% | 16.07 | 1.43% |
| $\text{top}_{90}$ | 9.56 | 1.20% | 16.08 | 1.32% |
| $\text{top}_{100}$ | 9.55 | 1.13% | 16.12 | 1.21% |
| $\text{top}_{110}$ | 9.57 | 0.92% | 16.21 | 0.99% |
| $\text{top}_{120}$ | 9.62 | 0.60% | 16.25 | 0.68% |
| $\text{top}_{123}$ | 9.69 | 0.26% | 16.40 | 0.32% |

Table 4: The BLEU scores comparing the baseline system to its performance when supplemented with $\text{top}_n$ ranks of rules. LOOV refers to the leftover OOV that are not handled.

12 proper nouns (28%) among the handled OOVs. However the ratio of *acceptable* proper nouns to all proper nouns (17%) is around half the ratio of *acceptable* non-proper nouns to all non-proper nouns (32%). This result is not unexpected since we did not focus on proper nouns in this paper.

### 6.3 Comparing with other Techniques for OOV Handling

Following Habash (2008), we compare our morphology variation approach with two techniques for OOV handling. In the first technique, SPELLVAR, we produce spelling variation hypotheses that assume the word is misspelled by letter deletion, addition, substitution or inversion (alternating the position of two adjacent letters). We allow one spelling modification at a time. This is a very simple technique to implement and does not require any additional resources. The spelling hypotheses are used to link an OOV word to an INV word. Then the phrases associated with the INV word are recycled in a similar manner to what we did in morphology variation.

| | DEVSET | MT08 |
|---|---|---|
| BASELINE | 9.47 | 15.95 |
| MORPHVAR | 9.69 | 16.40 |
| SPELLVAR | 9.70 | 16.34 |
| TRANSVAR | 9.76 | 16.55 |
| ALL | 9.79 | 16.57 |

Table 5: Results of OOV handling using different techniques: MORPHVAR is our morphology variation approach, SPELLVAR is a spelling variation approach and TRANSVAR is an approach that produces transliteration hypotheses; ALL is a union combination of all phrases created by the three techniques.

The second technique, TRANSVAR, is more complex as it involves introducing completely novel phrases through using a transliteration component. We retarget a publicly available transliteration system for Arabic-English (Habash, 2008) by converting Urdu words to an "Arabic form." The conversion includes simple substitution of letters only used in Urdu to their closest Arabic variant: e.g., ڑ *ṛ* becomes ر *r*, ٹ *ṭ* becomes ت *t* and ژ *ž* becomes ج *j*. Then, the Arabic-to-English transliterator is used. The newly generated pairs are assigned very low translation probabilities that do not interfere with the rest of the phrase table. Weights of entries are modulated by the degree of similarity indicated by the confidence measure returned by the transliterator. Given the large number of possible matches, we only pass the top 20 matches to the phrase table.

Table 5 shows the results comparing our baseline with the morphology variation (MORPHVAR) approach as well as the SPELLVAR and TRANSVAR approaches. We also combine all of these approaches (ALL) by simply taking the union of all the new phrases. Each of the techniques clearly improves over the baseline and the combination improves the most. Although these scores are not strictly statistically significant, they do show a consistent trend across two different test sets.

The TRANSVAR approach shows the biggest single improvement in BLEU score, which is expected given that it uses additional outside resources. However, SPELLVAR does not consistently beat MORPHVAR, even though it comes close. We believe that although many of the morphological variations can be thought of as simply "spelling errors" from the point of view of English, some are too complex to handle as such since this may lead to a huge overgeneration of spelling hypotheses.

It is important to remember that these approaches are only working on OOV words, which are around 2.6% of all words, yet they (the approaches) show an average increase of 0.47 BLEU from the BASELINE (3.6% relative). By comparison, gaming the BLEU metric by deleting all OOV terms increases the score of DEVSET to 9.92 BLEU. This means that the combined techniques give us (without gaming) 71% of the score increase that we could have received through gaming. The morphology variation technique currently gives 49% of the gaming score increase.

## 7 Conclusion and Future Plans

We presented an approach for automatic unsupervised learning of morphological variation rules for the purpose of *online OOV Handling* in Urdu-English MT. We reduce the OOV rate from a standard baseline average of 2.6% to an average of 0.3% (or 89% relative decrease). We also increase the BLEU score by 0.45 (absolute) and 2.8% (relative) on a standard test set. A manual error analysis shows that 28% of handled OOV cases produce acceptable translations in context.

In the future we plan to improve our morphology learning model to allow learning independent rules that can be applied in a sequential manner. We will also consider using morphological analyzers for Urdu to help with the creation of rules. Finally, we plan to investigate the use of different weighing schemes to manipulate the probabilities in the recycled phrases. We have done some preliminary experiments that show some promise.

## Acknowledgement

The first author was funded under the DARPA GALE program, contract HR0011-06-C-0023.

## References

Tim Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0.

Mathias Creutz and Krista Lagus. 2007. Unsupervised Models for Morpheme Segmentation and Morphology

Learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1).

Mona Diab and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging Using Parallel Corpora. In *Proceedings of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA.

W.A. Gale and K.W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19:75–102.

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27.2:153–198.

Nizar Habash and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL-06)*, New York, NY.

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

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of ACL-08*, Columbus, OH.

Muhammad Humayoun. 2006. Urdu Morphology, Orthography and Lexicon Extraction. Master's thesis, Chalmers University of Technology and Göteborg University.

Sara Hussain. 2004a. Finite-State Morphological Analyzer for Urdu. Master's thesis, National University of Computer and Emerging Sciences, Lahore, Pakistan.

Sarmad Hussain. 2004b. Urdu Localization Project. In *Proceedings of the International Conference on Computational Linguistics (COLING-2004) Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland.

Douglas Jones and Rick Havrilla. 1998. Twisted Pair Grammar: Support for Rapid Development of Machine Translation for Low Density Languages. In *Proceedings of AMTA-98*, Langhorne, PA.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proceedings of AMTA-04*, Washington, DC.

Young-Suk Lee. 2004. Morphological Analysis for Statistical Machine Translation. In *Proceedings of HLT-NAACL-04*, Boston, MA.

R. Mahesh and K. Sinha. 2007. Using Rich Morphology in Resolving Certain Hindi-English Machine Translation Divergence. In *Proceedings of the Machine Translation Summit (MT SUMMIT XI)*, Copenhagen, Denmark.

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

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of ACL-03*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL-02*, Philadelphia, PA.

Maja Popović and Hermann Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

Monica Rogati, J. Scott McCarley, and Yiming Yang. 2003. Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. In *Proceedings of ACL-03* Sapporo, Japan.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of ACL-08*, Columbus, OH.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-02)*.

David Talbot and Miles Osborne. 2006. Modelling Lexical Redundancy for Machine Translation. In *Proceedings of ACL-06*, Sydney, Australia.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can We Translate Letters? In *Proceedings of ACL-07 Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based Backoff Models for Machine Translation of Highly Inflected Languages. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy.

D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing Multilingual Text Analysis Tools Via Robust Projection Across Aligned Corpora. In *Proceedings of HLT-01*.