A Taxonomy of Weeds: A Field Guide for Corpus Curators to Winnowing the Parallel Text Harvest

Katherine M. Young[†] katherine.young.ctr.1@us.af.mil
N-Space Analysis, LLC

Jeremy Gwinnup jeremy.gwinnup.1@us.af.mil
Air Force Research Laboratory

Lane O.B. Schwartz lanes@illinois.edu
University of Illinois

Abstract

Modern machine translation techniques rely heavily on parallel corpora, which are commonly harvested from the web. Such harvested corpora commonly exhibit problems in encoding, language identification, sentence alignment, and transliteration. Just as agricultural harvests must be threshed and winnowed to separate grain from chaff, electronic harvests should be carefully processed to ensure the quality and usability of the resulting corpora. In this work, we catalog a taxonomy of problems commonly found in harvested parallel corpora, and outline approaches for detecting and correcting these problems.

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some problems; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. While this work is by no means comprehensive of all problems extant in corpus creation and curation, we nevertheless believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

1 Introduction

Statistical machine translation typically requires large amounts of translated parallel text to serve as training data for statistical translation models. End-users of machine translation may use in-house data developed from years of prior human translation efforts (Plitt and Masselot, 2010; Hellstern and Marciano, 2014). A perhaps more common practice, developed over the past fifteen years (Resnik, 1998), involves the automatic harvest of parallel corpora from online

[†]This work is sponsored by the Air Force Research Laboratory under Air Force contract FA-8650-09-D-6939-029.

resources, such as bilingual web sites (Smith et al., 2013) or the crowd-sourced translations of the TED Talk transcripts (Cettolo et al., 2012).

Just as agricultural harvests must be threshed and winnowed to separate grain from chaff, electronic harvests may be carefully processed to ensure the quality and usability of the resulting corpora. Simard (2014) suggested the metaphor of weeds choking out cultivated plants to be more apropos than that of cleaning "dirt" from corpora. We adopt this terminology, identifying a broad variety of such *weeds* found growing wild in online data, potentially degrading the quality of harvested corpora. In keeping with this botanic metaphor, we use *zizania*, a Greek term for a type of weed that grows intermixed with wheat,[1] as a basis for our taxonomic nomenclature.

In this work, we present a taxonomy of weeds commonly found in harvested parallel corpora, and outline approaches for detecting and correcting these problems. At the highest rank, the taxa we present are categorized based on provenance: Do the errors originate from problems during automatic processing of the text (*zizania ex machina*) or from human failure (*zizania ex homine*)? We categorize six major types of the former (§2.1–2.6), as well as six major types of the latter (§3.1–3.6). Throughout this work, we consider weeds that have been previously identified in the established literature, as well as weeds that we have encountered that have not heretofore been described in the literature.

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some weeds; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. We believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

2 *Zizania ex machina*: Weeds of mechanical origin

We now survey various *zizania ex machina*: weeds that originate during automated corpus processing.

2.1 Wrong Language Text

Wrong-language text errors can occur during automatic collection of parallel text from websites. The scraping program may mis-identify similar languages, or the program may fail to notice a section of foreign text within a page produced in the correct language. For example, if the program is scraping an English-language site with hotel reviews, it may pick up some reviews written in French. Alternatively, the program may fail to exclude a section of text that has remained untranslated across pages of a multilingual site. These failures create two types of errors that can be automatically detected, *Source-Source* errors, and *Source-Other* errors.

[1] See, for example, the usage of *zizania* in the Greek New Testament (Matthew 13:25).

2.1.1 Source-Source instead of Source-Target

An example of Source-Source error occurred in the initial release of the IWSLT 2014 data (Cettolo et al., 2012), in which some of the parallel English-French text was provided untranslated, creating English-English data. This was subsequently corrected. Source-Source errors can be detected automatically by searching for sentences that are duplicated across parallel text; these are usually untranslated sections. Short duplicate sentences should be examined separately, since there can be some legitimate duplication if the text contains URLs, named entities, borrowed words, or quotations. Legitimate duplication at the token level can also be caused by cognates (for example, the English word *importance* matches French *importance*).

2.1.2 Source-Other instead of Source-Target

Examples of Source-Other errors can be found in the French side of the 10^9 English-French corpus (Callison-Burch et al., 2009), in which we find paragraphs in Greek, Russian, German, and other languages. Such Source-Other errors can be detected easily if the incorrect language has a different character set than the correct language. For example, a section of Greek within a supposedly French document can be easily filtered out by specifying a desired range of permitted Unicode code points.

For languages with similar alphabets, we apply a simple dictionary-based program to remove sentences with a majority of unknown words. Recent work (Zampieri et al., 2014; Lui et al., 2014) leverages character n-grams, POS sequences, and other features to train language discrimination systems for similar languages.

Depending on the application, thresholding may be desired to allow a specified amount of wrong-language text (for foreign names, borrowed words, quotations, etc.). On the other hand, web-scraped text from multi-lingual sites often contains isolated wrong-language phrases that we may want to remove, such as hyperlinks in multiple languages. Multi-lingual sites can also contain stock phrases like "Click here to login" that may remain untranslated across the site; these might also need to be removed.

2.1.3 An illustration of a specific language identification clean-up process

For languages with similar but not identical alphabets, detection programs can be written that are specific to that language pair. For example, the English-Russian Common Crawl data includes sections which are actually English-Ukrainian. Ukrainian has four characters not found in Russian which can be used to identify unwanted Ukrainian segments: UKRAINIAN I (i I), YI(ї Ї), GHE WITH UPTURN (ґ Ґ) or IE (є Є). We make an exception to allow UKRAINIAN I in Russian segments when it occurs in a potential context for a Roman numeral (adjacent to Latin X, I, V, x, i, v, or their Cyrillic counterparts).

Second, on the English side of the Russian-English Common Crawl, we find sections of text in other languages such as French. Both English and French use the Latin character set, but French uses special characters not typically found in English such as à é ê î ô œ ç; these could be used to identify the presence of French, with some proportion of exceptions allowed for borrowed words like *café*. However, for the Common Crawl we we also want to detect other non-English languages like Spanish. Instead of relying on specific accented characters to detect

Experiment	Corpus Size	Filtered Corpus Size	Avg. Cased BLEU	Avg. Uncased BLEU
Baseline	878386	732129	25.39	26.59
Cleaned	772530	642746	25.73	26.95

Table 1: Before and After Common Crawl experiment results reported in BLEU

non-English text, we apply a spell checker to identify English text. We use the `aspell`[2] spell-checker to determine the proportion of words that are not recognized as English, and compare this to a set threshold to identify the wrong-language sections. We exclude from consideration words of 3 characters or less, because many short words have false friends in other languages (e.g., *die* in English and German, *on* in English and French).

We demonstrate the effectiveness of these techniques by taking a baseline WMT15 MT system and replacing the phrase and lexicalized reordering tables with ones generated from the Common Crawl corpus in both original and cleaned configurations. Table 1 shows the cleaned corpus yields a +0.34 BLEU improvement over the non-processed baseline even with a 12% reduction in corpus size.

2.2 Historical Encoding Errors

Portions of a corpus are sometimes encoded using a different character encoding scheme than the rest of the document. If not detected and corrected, this leads to an encoding cipher, where sentences appear shifted to an incorrect character range. Encoding errors of this type can also occur when extracting text from a PDF document.

In the Russian-English Common Crawl parallel corpus, a number of Russian source sentences are encoded using the 8-bit Windows-1251 character encoding scheme. Most sentences in this corpus are encoded using UTF-8; when Windows-1251 encoded sentences are interpreted as UTF-8, the Cyrillic characters incorrectly appear as characters from the Latin-1 supplement block. This can be corrected by shifting these characters ahead by 350_{hex} code points into the correct Unicode Cyrillic character range. An example of this code point shift is shown in Figure 1 below:

(**a**) Справка по городам России и мира.
(**b**) Ñïðàâêà ïî ãîðîäàì Ðîññèè è ìèðà.

Figure 1: Russian sentence (**a**) originally encoded as Windows-1251, interpreted as UTF-8 (**b**)

Encoding errors may also show up in isolated characters. We see this in some of the Common Crawl data, in which French accented characters have been converted to Cyrillic characters. For example, we find the words *équipe* and *château* written as *ўquipe* and *chвteau*. This is the reverse of the Russian code point shift described above, and these errors can also be corrected automatically if we know that the Cyrillic characters are out of range for our text. The Common Crawl exhibits a variety of code point encoding problems in addition to those shown here. Out of range characters should be examined for code point shifts and encoding problems that could possibly be corrected.

[2] http://www.aspell.net

Lang.	Set	Sentences w. repeat errors	Total sentences
French	dev2010	11	887
Chinese	dev2010	87	887
	tst2010	81	1570
	tst2014	13	1068
Farsi	tst2010	1	885
	tst2011	22	1132
	tst2012	343	1375
	tst2013	187	923
	tst2014	53	1131

Table 2: Number of sentences containing segment-internal repetition errors in IWSLT dev and test sets

Lang.	Year	Sentences w. repeat errors	Total sentences
Arabic	2013	3	155,047
	2014	5	186,467
Chinese	2014	550	177,901
Farsi	2013	5,749	81,872
	2014	8,987	112,704
French	2013	173	162,681
	2014	373	186,510
Russian	2013	109	135,669
	2014	145	185,205

Table 3: Number of sentences containing segment-internal repetition errors in IWSLT training sets

There can also be encoding problems with individual characters. A confusion between UTF-8 encoding and Windows-1252 encoding can lead to a single character such as 0xE28099 (’) being interpreted as multiple, single-byte characters: 0xE2 (â), 0x80 (€) and 0x99 (™) Notenbloom (2009). These single-byte/multiple-byte encoding errors can be corrected programmatically with existing tools.

Finally, we note that the character U+FEFF may appear in some files as the residue of a byte order marker at the start of a file; this should be deleted to avoid confusion with the Arabic script character U+FEFF, which is a zero-width non-breaking space.

2.3 Bidirectional Reversal

Adobe’s Portable Display Format (PDF) is meant as a display format and does not focus on the orderly layout of data in the document’s container. This presents issues when extracting text in an orderly fashion from PDF documents. Extraction issues are compounded when dealing with custom fonts and historical encoding schemes. Additional issues involve the display of Right-to-Left (RTL) text.

Sometimes, extraction of RTL text from PDF creates text in which the line is reversed, character-by-character. We can detect reversals automatically by checking the words against a dictionary or word-frequency list to derive a percentage of unknown words. We then compare that percent unknown against the typical score for text from that language. If the percent unknown is supsiciously high, we can use a program to character-reverse the line, and repeat the dictionary check; a better score on the reversed line confirms the reversal error. In correcting reversed lines, we need to be careful how we handle digits, which run left-to-right within right-to-left text in many Arabic-script langauges.

2.4 Automatic sentence alignment errors

When parallel sentences are aligned, typically via automated means, mistakes in sentence alignment lead to mis-aligned sentence pairs that do not represent mutual translations. Many parallel corpora are naturally aligned at the document level: A human translator translates a source document into a target language. However, most statistical methods that make use of parallel data require alignment at the sentence level, and automated sentence aligners may make errors.

Various automated techniques have been proposed to minimize the problem of mis-aligned sentences. Gale and Church (1991) proposed an automated length-based sentence alignment technique that compared the number of words in source and target sentences. Proposed extensions to length-based approaches include the use of cognate frequency (Simard et al., 1992) or other lexical cues (Wu, 1994). Structural tags (such as HTML elements) have also been proposed as an aid to guide sentence alignment (Resnik, 1998).

2.5 Segment-Internal Repetition and Chunking Errors

Processing errors may cause a sentence or sub-sentential fragment to be improperly duplicated within a given line. In many cases, such repetition can be automatically detected and corrected; examination of the corresponding parallel sentence can assist in this process.

The IWSLT 2014 data, for example, contain substantial cases of repetition errors, especially for certain language pairs (see Tables 2 and 3 on the preceding page). An example of a repetition error is shown in Figure 2 below:

Last year I showed these two slides so that demonstrate that the arctic ice cap, *which for most of the last three million years has been the size of the lower 48 states,* has shrunk by 40 percent.
L'année dernière, je vous ai présenté ces deux diapositives qui montraient que la calotte glacière arctique, *qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska,* ***qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska,*** avait diminué de 40%.

Figure 2: Example of repeated phrase in English-French TED data. Within the French sentence, the words in ***bold italics*** represent an erroneous copy of the words in *italics*.

While some cases of repetition are not errors (the TED Talks in particular may contain repetition for rhetorical effect), the presence of high amounts of repetition errors in training data and development data can degrade machine translation quality; correcting the large number of

repetition errors in the IWSLT 2014 Farsi test file improved Farsi-to-English performance by +1.53 BLEU.

Chunking errors occur when sub-sentential segments are automatically combined without the necessary spacing. For example, a small number of files in the QED Corpus provided to the IWSLT 2016 competition (Abdelali et al., 2014) exhibit a chunking error, in which each line has run-together words in the middle of the line (see Figure 3). This is probably an error in assembly. The QED Corpus derives from the AMARA website, which enables crowd-sourced transcription of video; the AMARA interface presents the worker with 4-second segments of video to transcribe, and these are subsequently assembled into a larger text (Zukerman, 2013). We found 57 files with mid-line chunking, out of 19K total English files.

Chunking errors create unknown words for machine translation. A human looking at these files can analyze the problem easily, based on what is reasonable to expect in the sentence, but automatic, rule-based correction faces some difficulties. A spell checker like `aspell` can be applied to detect and correct run-together words, but we have to protect named entities and technical terms which may not appear in `aspell's` dictionary. We also have to be careful to split the words in the correct place. Initially, we simply split the unknown word into progressively longer sections of first word vs. second word, until we found two known words. This led to unfortunate splits like *thoughtsand* > *thought sand* instead of *thoughts and* and *monkeysin* > *monkey sin* instead of *monkeys in*. A word frequency list could be applied to select the best split. Alternatively, language modeling could determine which split creates the most reasonable sentence.

It's the difference between divergent **thinkingand** convergent thinking. You have to separate
the two so that you can diverge your **thoughtsand** come up with this great collection of
ideas, and then once you have this great **collectionof** ideas, you focus on the convergent thinking.

Figure 3: An example of chunking errors in the QED Corpus.

2.6 Harvested Machine Translations

When parallel corpora are harvested from the web, there is a danger that some of the parallel content was created by means of machine translation, rather than human translation. Attempts have been made to automatically identify machine-translated content using various machine learning techniques, including decision tree classifiers (Corston-Oliver et al., 2001), SVM classifiers (Gamon et al., 2005), maximum entropy classifiers (Rarrick et al., 2011), watermarking (Venugopal et al., 2011), and identifying the presence of characteristic MT errors (Antonova and Misyurev, 2011). The extent to which the inclusion of machine-translated content in MT training data harms translation quality of the trained system may depend largely on the quality of the harvested machine translations (Simard, 2014).

3 *Zizania ex homine*: Weeds of human origin

In this section we survey weeds of human origin that show up in translated text from online sources. In general, *zizania ex homine* are harder to correct than *zizania ex machina*, but some

automatic correction is possible.

3.1 Mixed Alphabets

Words with mixed alphabets visually resemble correctly spelled words, but are treated as separate tokens in the machine translation process. Such words can be automatically detected and corrected, converting characters to the majority alphabet for that word when they have visually similar counterparts.

Word	Latin (L) or Cyrillic (C)	Meaning
она	LCL	she
сейчас	LCCCCC	now
MP3-плеер	LLL-CCCCC	MP3-player
MP3плеер	LLLCCCCC	MP3player
амазон.com	CCCCCC.LLL	amazon.com
іпациент	LCCCCCCC	iPatient

Figure 4: Examples of Mixed-Alphabet words. In the center column, we annotate each character of the corresponding Russian word as either Latin (L) or Cyrillic (C). For example, in the first row, the Russian word она is encoded such that the Latin characters *o* and *a* are used instead of the more appropriate (but visually indistinguishable) Cyrillic equivalents.

We have encountered mixed alphabet words in the Russian sections of the Russian-English Common Crawl and in the Russian transcriptions of TED Talks. This occurs when the translator uses a combination of Latin and Cyrillic characters to write a Russian word. The reason for these mixed spellings is unknown; perhaps it is due to limitations of the translator's input method, or perhaps it is influenced by typing both English and Russian words. For example, although the first letter and last letter in the word сейчас appear visually indistinguishable, in this instance we find that the former is U+0063 LATIN SMALL LETTER C and the latter is U+0441 CYRILLIC SMALL LETTER ES. We even find the Russian word она written with U+006F LATIN SMALL LETTER O and U+0061 LATIN SMALL LETTER A instead of the appropriate Cyrillic counterparts (U+043E and U+0430); this word is harder to correct, since the majority favors the wrong alphabet.

Some mixed alphabet spellings are deliberate, combining a borrowed English word with a Russian word. Figure 4 above shows examples of this behavior. Converting punctuated words on a part-by-part basis can protect some but not all of these deliberate mixed spellings from automatic conversion.

In addition to the mixed alphabet spellings in Russian, we find creative spellings in many languages that borrow from other character sets, or repurpose characters within the source alphabet, particularly for punctuation. Some examples are given in Figure 5 on the next page. Determining how to correct such creative spellings generally requires human intervention.

3.2 Mixed Morphology

When a translator brings in a borrowed word through transliteration, he or she may choose to inflect the borrowed word using target language morphology. For example, in Urdu text we

Language	Character Written		Character Intended	
Urdu	U+002D -	Latin Hyphen	U+06D4 ۔	Urdu Full Stop
French	U+00A8 ¨	Latin Diaeresis	U+0022 "	Latin Quotation Mark
Russian	U+0431 б	Cyrillic Small Letter Be	U+0036 6	Latin Digit Six
English	U+006F o	Latin Small Letter O	U+00B0 °	Latin Degree Sign

Figure 5: Examples of Creative Spelling.

find the borrowed English word *leader* with the plural suffix /-wn̲/, creating لیڈروں /lyd̲rwn̲/, as well as the borrowed word with the original English plural form (*leaders*), لیڈرز /lyd̲rz/. Names in particular are subject to variation in the application of target language morphology. An examination of names borrowed into Russian from English in the TED Talk data showed this range of behavior: a) first and last name both uninflected, b) first and last name both inflected, c) last name only inflected. Examples are shown in Figure 6; all three examples are possessive structures which should occur with genitive case.

Russian Text	Phonemes	English Text	Annotation Type
песню Уитни Хьюстон	/uitni x'yuston/	a Whitney Houston song	a) neither name inflected
закон Артура Кларка	/artur+a klark+a/	Arthur Clarke's law	b) both names in genitive case
Книга Эл Гора	/ɛl gor+a/	The Al Gore book	c) last name in genitive case

Figure 6: Examples of Mixed Morphology.

Inflected borrowed words often show up as out-of-vocabulary (OOV) words in MT output. If OOV words are going to be transliterated (see §3.3), it is useful to first apply a stemmer to remove any inflectional endings. Lexical approximation can sometimes rehabilitate inflected borrowed words and allow them to be translated (Mermer et al., 2007). Alternatively, Schwartz et al. (2014) identify inflected OOV words at the start of the decoding process, and replace them with variant inflected forms from the phrase table.

3.3 Transliteration of Names and Borrowings

Borrowed words and names may occur in transliteration, with the original sounds mapped into the characters of the new language. While such coinages are not errors, they are subject to variation that creates problems when an MT system attempts to relate them to the original forms.

Statistical methods may be applied to deal with this variation in transliteration, as for example in Durrani et al. (2014). Our work focuses instead on improving rule-based transliteration, which maps characters into their typical sound values. Because there can be variation in the character-to-sound mapping in both languages, the output of rule-based transliteration is often faulty. This output can be improved by constraining the results to actual English spellings. In particular, we address the recovery of named entities that persist as OOV words in the output of machine translation. We describe two ways to constrain the results of named entity (NE) transliteration into English, one using an English pronunciation dictionary, and another using parallel training data to create a transliteration-based map of NE pairs.

3.3.1 Recovering Names via Transliteration in Conjunction with an English Pronunciation Dictionary

Rule-based transliteration can be improved by leveraging a target language pronunciation dictionary. We adapt the CMU English pronunciation dictionary[3] to guide transliteration from Russian into English. Because vowel spellings may be variable, we create a fall-back representation for each word in which all vowels are converted to a placeholder character, @. We derive a word frequency count from the training data and record the frequency count for each dictionary entry. We also supplement the pronunciation dictionary by noting any words in the WMT 2014 Russian data (Bojar et al., 2014) that are not listed in the CMU dictionary, and deriving their phonetic forms via `Sonic` (Pellom and Hacioglu, 2001).

When we run our transliteration program, we first map the Cyrillic characters into their typical sounds, recording multiple possibilities where appropriate. Next, we compare these phonetic mappings to the phonetic entries in the English pronunciation dictionary. We try to find words which match the sound pattern for both consonants and vowels; failing that, we use the vowel placeholder representations and allow @ to match any vowel or sequence of vowels. If there are multiple candidate words, we select the word with the highest word frequency count. We output the English spelling of the chosen word.

3.3.2 Recovering Names via a List of Transliterated NE Pairs

We apply transliteration and NE tagging to create a list of NE pairs from parallel Russian-English text; this list can subsequently be used to either pre-translate NEs, or to recover OOV names in the MT output. First, we apply the `mystem`[4] morphological analyzer to tag NE in the Russian text. For each NE, we then use rule-based transliteration to get a phonetic form, from which we identify possible matches in the English sentence. We record the best match along with the Levenshtein edit distance between the phonetic form and the English spelling, normalized for word length. NE pairs with a distance score below 0.66 are stored in a NE list that can be used to translate Russian NEs. When applied to the Russian-English WMT 2014 training data, this method generated a list of 216K potential NE pairs.

3.3.3 Third Language Mappings

Automatic transliteration processes can stumble when dealing with words that derive from languages other than the source or target. In English, for example, the letter *j* usually indicates the affricate sound [dʒ], but in words of Spanish origin, it may represent [h]. This presence of a third-language sound pattern complicates the use of transliteration. Hagiwara and Sekine (2011) and Li et al. (2007) suggest ways to detect alternate languages in statistical transliteration: Li et al. (2007) train with language-tagged word pairs; Hagiwara and Sekine (2011) introduce latent classes to model language origins. For rule-based transliteration, developing programs to detect and correct such third-language spelling differences requires examination of the sound patterns of the various languages; human intervention may be required to decide when to apply the alternate mappings.

Russian provides a particular problem for transliteration due to the presence of third-

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[4] https://api.yandex.ru/mystem/

language sound patterns from Chinese. When referring to Chinese names in Russian texts, Russian writers follow the Palladius mapping (Palladius and Popov, 1888) to transliterate Chinese names into Cyrillic. Many Cyrillic characters generated by this mapping represent different sounds than those Cyrillic characters typically represent in Russian. For example, the Cyrillic character ж typically represents /zh/, but in the Palladius mapping it represents /r/, and the combination of characters Чж is used to represent /zh/. Figure 7 illustrates how applying the typical Russian-to-English transliteration for OOV Russian words will cause errors for Chinese names, unless we first reverse this Palladius mapping (Young et al., 2012).

(**a**) 翟志刚
(**b**) Чжай Чжиган
(**c**) Chzhay Chzhigan
(**d**) Zhai Zhigang

Figure 7: Chinese name (**a**), with transliterations into Cyrillic (**b**) and Latin using normal Cyrillic-to-Latin transliteration (**c**) and reverse Palladius transliteration (**d**). The output in (**d**) is correct.

3.4 Under-achieving Translation

We use the phrase *under-achieving translation* to designate weeds that result from a lack of attention by the human translator. Sometimes translators leave a word untranslated; this kind of error can be detected by methods discussed above in §2.1, including the detection of out-of-range characters if the languages have different alphabets. More subtle weeds can occur when the translator chooses transliteration in place of translation, as the appropriateness of transliteration depends on context.

3.4.1 Transliteration in Place of Translation

Sometimes the human translator simply transliterates the source word, even when an appropriate translation exists in the target language. This may represent a translator's decision to preserve the original form in a named entity, or it may reflect a careless translation. For example, the English word *review* has various Russian translations, such as журнал (review, journal) and рецензия (review, critique). However, in the IWSLT 2014 training data we find *review* transliterated in the phrase, *Harvard Business Review*, Гарвард Бизнес Ревью /garvard biznes rev'ju/. This choice preserves the title of the publication; translating *review* to журнал /ʒurnal/ would have introduced confusion with the English word *journal*. In the Common Crawl, on the other hand, we find an inappropriate transliteration of *review*, in the phrase *Awards and Reviews*, which becomes Награды и Ревью /nagradi i rev'ju/. This instance should probably have been translated. For unfamiliar words, a translator may resort to letter-by-letter spelling, as in the Russian spelling опоссум for *opossum*, which reflects the English spelling rather than the pronunciation [pasəm] or [əpasəm]. The coexistence of translation, sound-based transliteration, and letter-based transliteration creates more variation that must be addressed in machine translation.

3.4.2 Code-switching

The use of transliterated foreign words may also be driven by a form of code-switching (Myers-Scotton, 1993; Diab et al., 2014, 2016) in which the writer deliberately uses foreign words. For example, Urdu writers frequently use transliterated English words, instead of their Urdu counterparts, because the use of English exhibits a level of prestige (Upal, 2008). Hence, we may find transliterated English words in Urdu source text, as well as in Urdu text that has been translated from English. In Table 4, the Urdu writer has used English words in transliteration for four words, in place of using the Urdu words. Such transliterations complicate the machine translation of Urdu by creating variations between transliterated English words and the actual Urdu words.

English	In the top ten, India comes in the last
Urdu	اس سرٹیفکیشن کی ٹاپ ٹین میں بھارت آخری نمبر پر ہیں۔
Transliteration	as srṯyfkyšn ky ṯap ṯyn myn bhart Ajry nmbr pr byn
English words	– certification – top ten – – – number – –

Table 4: Urdu transliteration example. In this example, the author of the Urdu sentence used four English words (transliterating *certification* into *srṯyfkyšn*, *top* into *ṯap*, *ten* into *ṯyn*, and *number* into *nmbr*) instead of using the corresponding Urdu words.

3.5 Over-achieving Translation (Explicitation)

Human translators intend to communicate meaning, and so may depart from the source text in ways that improve understanding, but degrade the usefulness of the translation as parallel text. Translators may expand acronyms, add explanation of localized vocabulary, or include the actual source-language words. Translators working on informal speech may remove false starts and clean up awkward sentence structure.

We term this type of explicitation (Blum-Kulka, 1986) *over-achieving translation*, in contrast to the *under-achieving translation* of the previous section. This type of extra information is difficult to detect and modify for machine translation. If the translator has set off added material in brackets or parentheses this can be detected, but often the additional material is integrated into the translation.

The TED Talks suffer from particular problems with over-achieving translation, since they are spoken presentations supplemented by visual aids. The English transcriptions tend to follow the speaker closely, while the translations often clean up disfluency.[5] If text appears on the slides, the translators often include a translation of this material in the transcript.

Similarly, sentence alignment problems can also be caused when human translators summarize (Khadivi and Ney, 2005), engage in one-to-many, many-to-one, or many-to-many sentence translations (Gale and Church, 1991), or engage in non-literal free translations (Imamura and Sumita, 2002);[6] the resulting parallel sentences may be less useful from the perspective of

[5] Cho et al. (2014) suggest handling this issue by tightly integrating disfluency removal into the MT decoding process.

[6] Imamura and Sumita (2002) also identify as problematic to their data-driven rule-based MT technique situations where a given source phrase is translated in multiple different ways throughout the corpus. Modern statistical machine translation techniques tend to be relatively resistant to this variety of weed.

machine translation training than other more literal translation pairs. This problem may be mitigated by removing less literal translation pairs from the parallel corpus (Okita, 2009; Jiang et al., 2010), or by flagging sentence pairs which exhibit atypical length ratios for manual inspection (our tools take the latter approach).

3.6 Translation Directionality

Other researchers have noted that translated text differs in crucial ways from native text, in both general simplification (Lembersky et al., 2013) and by influence from the word order and vocabulary choice of the source language text (Fusco, 1990). Koppel and Ordan (2011) show that classifiers can be trained to distinguish the direction of translation. Translation models are typically built from parallel corpora without regard for which language of the pair is the original source language. Changing this paradigm to one where original source language is taken into account has been shown to improve translation quality (Kurokawa et al., 2009).

4 Conclusion

This work is motivated by the lack of a standardized field guide outlining best practices for curating parallel corpora, especially those harvested from the web. Even the most-well curated parallel corpus is likely to contain some problems; even Europarl (Koehn, 2005), arguably the most widely examined parallel corpus, has undergone eight distinct revisions since its release in 2005. In this work, we categorize six major types of problems that originate in automated processing of corpora, as well as six major types of problems that originate in human translator actions. In this work, we establish an initial taxonomy of weeds. While this work is by no means comprehensive of all problems extant in corpus creation, we nevertheless believe that a practical taxonomic field guide, laying out likely pitfalls awaiting corpus curators will represent an important contribution to our community.

The extent to which various types of weeds are harmful in practice is not fully established. Asia Online (2009) and others have claimed substantial positive results from weeding. Likewise, we found substantial improvement in translation quality when major repetition errors are corrected. On the other hand, Goutte et al. (2012) report that statistical MT systems may be robust to sentence alignment errors as high as 30%. In future work we plan a more thorough empirical examination exploring how sensitive various machine translation systems are to various types of weeds.

References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proc. Building and Using Comparable Corpora (BUCC'11)*.

Asia Online (2009). Study on the impact of data consolidation and sharing for statistical machine translation. Technical report, TAUS.

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In House, J. and Blum-Kulka, S., editors, *Interlingual and Intercultural Communication*. Narr, Tübingen.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proc. WMT*.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. WMT*.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proc. EAMT'12*, Trento, Italy.

Cho, E., Niehues, J., and Waibel, A. (2014). Tight integration of speech disfluency removal into SMT. In *Proc. EACL'14*, Gothenburg, Sweden.

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proc. ACL'01*, Toulouse, France.

Diab, M., Fung, P., Hirschberg, J., and Solorio, T., editors (2016). *Proceedings of the Second Workshop on Computational Approaches to Code Switching*.

Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors (2014). *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Durrani, N., Sajjad, H., Hoang, H., and Koehn, P. (2014). Integrating an unsupervised transliteration model into statistical machine translation. In *Proc. EACL'14*, Gothenburg, Sweden.

Fusco, M. (1990). Quality in conference interpreting between congnate languages: A preliminary approach to the Spanish-Italian case. *Interpreters' Newsletter*, (3).

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proc. ACL'91*, Berkeley, California.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proc. EAMT'05*, Budapest, Hungary.

Goutte, C., Carpuat, M., and Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In *Proc. AMTA'12*, San Diego, California.

Hagiwara, M. and Sekine, S. (2011). Latent class transliteration based on source language origin. In *Proc. ACL'11*, Portland, Oregon, USA.

Hellstern, A. and Marciano, J. (2014). Two sides of a coin: Machine translation and post-editing projects from the perspectives of the client and language services provider. Proc. ATA'14.

Imamura, K. and Sumita, E. (2002). Bilingual corpus cleaning focusing on translation literality. In *Proc. INTERSPEECH'02*, Denver, Colorado.

Jiang, J., Way, A., and Carson-Berndsen, J. (2010). Lattice score based data cleaning for phrase-based statistical machine translation. In *Proc. EAMT'10*, Saint-Raphaël, France.

Khadivi, S. and Ney, H. (2005). Automatic filtering of bilingual corpora for statistical machine translation. In *Proc. NLDB'05*, volume 3513 of *Lecture Notes in Computer Science*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. MT Summit X*, Phuket, Thailand.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proc. ACL'11*, Portland, Oregon.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proc. MT Summit XII*, Ontario, Canada.

Lembersky, G., Ordan, N., and Wintner, S. (2013). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, 39(4).

Li, H., Sim, K. C., Kuo, J.-S., and Dong, M. (2007). Semantic transliteration of personal names. In *Proc. ACL'07*, Prague, Czech Republic.

Lui, M., Letcher, N., Adams, O., Duong, L., Cook, P., and Baldwin, T. (2014). Exploring methods and resources for discriminating similar languages. In *Proc. Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138.

Matthew (1st century). Gospel of Matthew. Greek New Testament.

Mermer, C., Kaya, H., and Doğan, M. U. (2007). The TÜBİTAK-UEKAE statistical machine translation system for IWSLT 2007. In *Proc. IWSLT'07*, Trento, Italy.

Myers-Scotton, C. (1993). *Social Motivations for Codeswitching: Evidence from Africa*. Oxford Studies in Language Contact.

Notenbloom, L. (2009). Why do i get odd characters instead of quotes in my documents? http://askleo.com.

Okita, T. (2009). Data cleaning for word alignment. In *Proc. ACL-IJCNLP'09*, Singapore.

Palladius and Popov, P. S. (1888). *Китайско-русский словарь (Chinese-Russian Dictionary)*. Beijing, China. `https://archive.org/details/11888`.

Pellom, B. and Hacioglu, K. (2001). Sonic: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder, Colorado.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93.

Rarrick, S., Quirk, C., and Lewis, W. (2011). MT detection in web-scraped parallel corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, Xiamen, China.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proc. AMTA'98*, volume 1529 of *Lecture Notes in Artificial Intelligence*. Springer.

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proc. WMT'14*, Baltimore, Maryland.

Simard, M. (2014). Clean data for training statistical MT: The case of MT contamination. In *Proc. AMTA '14*, Vancouver, Canada.

Simard, M., Foster, G. F., and Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. In *Proc. Theoretical and Methodological Issues in Machine Translation (TMI'92)*, Montréal, Canada.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proc. ACL'13*, Sofia, Bulgaria.

Upal, M. A. (2008). Personal correspondence.

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. (2011). Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proc. EMNLP'11*, Edinburgh, Scotland.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proc. ACL'94*, Las Cruces, New Mexico.

Young, K. M., Gwinnup, J., and Reinhart, J. (2012). Reversing the Palladius mapping of Chinese names in Russian text. In *Proc. AMTA'12*, San Diego, California.

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In *Proc. Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland.

Zukerman, E. (2013). Review: Amara is a web-based service that lets anyone transcribe and translate online video. http://www.pcworld.com/article/2032787/review-amara-is-a-web-based-service-that-lets-anyone-transcribe-and-translate-online-video.html.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 4 Dec 2014. Originator reference number RH-14-113337. Case number 88ABW-2014-5534.