

PHONETICALLY-ORIENTED WORD ERROR ALIGNMENT FOR SPEECH RECOGNITION ERROR ANALYSIS IN SPEECH TRANSLATION

Nicholas Ruiz, Marcello Federico

Fondazione Bruno Kessler
Trento, Italy

ABSTRACT

We propose a variation to the commonly used Word Error Rate (WER) metric for speech recognition evaluation which incorporates the alignment of phonemes, in the absence of time boundary information. After computing the Levenshtein alignment on words in the reference and hypothesis transcripts, spans of adjacent errors are converted into phonemes with word and syllable boundaries and a phonetic Levenshtein alignment is performed. The phoneme alignment information is used to correct the word alignment labels in each error region. We demonstrate that our Phonetically-Oriented Word Error Rate (POWER) yields similar scores to WER with the added advantages of better word alignments and the ability to capture one-to-many alignments corresponding to homophonic errors in speech recognition hypotheses. These improved alignments allow us to better trace the impact of Levenshtein error types in speech recognition on downstream tasks such as speech translation.

Index Terms— automatic speech recognition, speech translation, mixed-effects models, error analysis

1. INTRODUCTION

Spoken language translation (SLT) systems are comprised by, at minimum, two components: an automatic speech recognition (ASR) system which provides audio transcripts of source language utterances, and a machine translation (MT) system that translates the transcripts. While there have been a number of efforts to construct tightly-coupled ASR and MT systems that are jointly trained and optimized, the majority of SLT systems employ a cascading approach in which ASR systems are trained and evaluated independently from the MT system [1, 2, 3]. In such a training paradigm, it is not clear how the introduction of ASR errors will affect translation quality. While there is a high correlation between ASR errors and translation quality, the impact of various ASR error types is still an open research problem.

ASR performance is typically evaluated using the Word Error Rate (WER) metric, which labels errors as word-level substitutions, deletions, or insertions, based on the Levenshtein word alignment between a reference transcript and an ASR hypothesis. Since the Levenshtein aligner often must decide between several “optimal” labeling sequences according

to its objective function, it may select an alignment sequence that does not adhere to phonetic or linguistic relationships between the word types. Although some evaluation toolkits use timestamp information to guide the alignment process, the problem persists. For example, the mapping of *a*→*doctor* in Fig. 1 has the side-effects of aligning *Dr.*→*brahmin* and deleting *Stanford*, thereby misaligning three content words. Additionally, WER is not capable of identifying homophonic errors across word spans, such as *anatomy*→*and that to me*¹.

These weaknesses in the word-level Levenshtein alignments used by WER inhibit the use of linguistically annotated ASR errors in assessing the quality of downstream speech-centric tasks, such as speech translation. In response, we introduce an additional step in the alignment process, which computes phonetically-oriented word alignments across adjacent word errors that were predicted by WER. We employ the text analysis component of a text-to-speech (TTS) engine, which dictates written text based on a pronunciation dictionary, letter-to-sound rules, and context-dependent pronunciation rules for numbers, ordinals, and acronyms.

In this paper, we describe the application of phonetically-oriented word alignment on spans of adjacent errors and show that these corrected alignments significantly alter the distribution of ASR word error types. We demonstrate its utility in text normalization as a pre-ASR evaluation step and additionally apply the re-aligned error types to a SLT error analysis experiment that measures the impact of speech recognition errors on SLT quality.

2. PHONETICALLY-ORIENTED WORD ALIGNMENT

The ambiguity in the word-level Levenshtein aligner is centered around the placement of substitution errors in an alignment sequence. As shown in Fig. 1, the error spans contain at least one substitution error and a number of insertion or deletion errors.

Our phonetically-oriented word alignment algorithm is divided into two stages. First, we capture error spans whose error labels are likely to be ambiguous. The reference and hypothesis words in each span are transcribed into phonemes by a TTS analyzer. Each phoneme is treated as an independent

¹We use the term “homophonic” to indicate groups of word sequences that are phonetically similar, but not necessarily identical, to one another.

WER					POWER				
<i>traditional way of learning human anatomy</i>					<i>traditional way of learning human anatomy</i>				
traditional way of loaning human and that to me					traditional way of loaning human "and that to me"				
	S		I	I	S				SS (1:4)
<i>we developed with a Dr. Brown in Stanford</i>					<i>we developed with a Dr. "Brown in" Stanford</i>				
we developed with doctor brahmin stamp or					we developed with doctor brahmin "stamp or"				
	S	S	S	S	D	S	SS (2:1)	SS (1:2)	

Fig. 1: Error alignment differences between WER and POWER. POWER aligns homophonic errors such as *anatomy* → *and that to me*, while WER rate only aligns single words (e.g. *anatomy*→*me*).

token and word and syllable boundary tokens are introduced. The reference and hypothesis tokens are aligned using a variant of the Levenshtein alignment algorithm that introduces the following constraints:

1. Boundary tokens may not be substituted.
2. Vowel phonemes can only be aligned to other vowels (including r-colored vowels, but not semivowels).
3. Consonant phonemes can only be aligned to other consonants (including semivowels).

The boundary tokens provide an implicit distance constraint, penalizing adjacent phonemes within the same syllable when they are aligned far from one another.

In the second stage, we recombine the phonetic alignments into word alignments by performing a left-to-right scan of the alignment sequence. Substitution alignments are identified by considering the words covered by the aligned phonemes contained between two “correct”-aligned word boundary markers in the reference and hypothesis. Single word substitutions (S) are distinguished from substitution spans (SS) containing multiple words in the reference or the hypothesis. If a sequence of reference phonemes are terminated with a word boundary, but no hypothesis words have been scanned, the reference word is marked as a deletion (D). Likewise, a hypothesis word with no aligned reference word is marked as an insertion (I).

Returning to Fig. 1, the Levenshtein aligner used in WER could have alternatively aligned the reference word *anatomy* to any one of the hypothesis words currently marked as insertion errors. However, *anatomy* is pronounced similarly to the entire sequence of the four hypothesis words in the error span. The phonetically-oriented alignment in Fig. 2 captures this phenomenon by aligning the smallest word boundary closure across the entire span of reference and hypothesis words, thereby identifying *anatomy*→*and that to me* as a substitution span and provides the alignment on the right-hand side of Fig. 1. Likewise, while WER may have considered slightly better word alignments like *Brown*→*brahmin* and *Stanford*→*or*, it is incapable of capturing relationships such as *Stanford*→*stamp or*.

ax n # ae t # ax m # iy
ae n d # dh ae t # t ax # m iy
S I I I I I D I I

Fig. 2: Phonetically-oriented alignment of *anatomy* and *and that to me*, with word (||) and syllable (#) boundaries.

	# ao l # ae t
0	← -3 ← -6 ← -9 ← -12 ← -15 ← -18 ← -21 ← -24 ← -27
↑	
3	0 ← -3 ← -6 ← -9 ← -12 ← -15 ← -18 ← -21 ← -24
↑	
# 6	3 0 ← -3 ← -6 ← -9 ← -12 ← -15 ← -18 ← -21
↑	
ao 9	6 3 0 ← -3 ← -6 ← -9 ← -12 ← -15 ← -18
↑	
r 12	9 6 3 4 ← -1 ← -4 ← -7 ← -10 ← -13 ← -16 ← -19
↑	
15	12 9 6 7 4 ← 0 ← 7 ← 10 ← 13 ← 16

# ao l # ae t	# ao l # ae t	# ao l # ae t			
# ao	D D D D S	# ao r	S D D D D	# ao r	S D D D D
	incorrect		incorrect		correct

Fig. 3: POWER alignments for *all at*→*or*. The Levenshtein backtrack matrix shows three alignments with the same edit distance scores. The third and correct alignment (highlighted in the backtrack matrix) compactly aligns *all*→*or*, while the others greedily align *or* to multiple reference words.

2.1. Word alignment heuristics

While the phonetically-oriented alignments provide better phonetically-grounded alignments, its underlying Levenshtein alignment algorithm must also decide between multiple equally-weighted best paths.

In particular, for alignments with large differences in the number of reference and hypothesis syllables, our implementation tends to align the first and last word boundaries close to the beginning and end of the alignment sequence. For example, Fig. 3 shows three candidate alignments for the error span *all at*→*or* that minimize the edit distance. Two out of three alignments attempt to align *or* to the entire two-syllable reference. However, only *all* should align to *or* as a substitution, and *at* should be considered a deletion error. We resolve ambiguities like these by finding the alignment that minimizes the number of alignment gaps between the first and last word boundaries in both the reference and hypothesis. In practice, we do this by encoding the best paths in the Levenshtein backtrack matrix into an edge-weighted graph and use Dijkstra’s algorithm to find the best path.

Since there still remains some noise in the phonetic alignments, we introduce a couple of heuristics to prevent the aligner from overzealously marking single-syllable words as members of a substitution span, when in reality they do not have a phonetic correspondence on the other side. When annotating a substitution span, we keep a record of the number

of reference and hypothesis syllables. If there is an extra syllable in the reference or hypothesis, we check if it is the first syllable of a new word. If so, we mark this word as a deletion or insertion error, respectively.

2.2. Scoring

Our Phonetically-Oriented Word Error Rate (POWER) score is defined nearly identically to WER as:

$$\text{POWER} = \frac{S + D + I + SS}{L},$$

$$SS = \sum_{span} \max(|span_{ref}|, |span_{hyp}|), \quad (1)$$

where L is the length of the reference and S , D , and I are the number of word-level substitution, deletion, and insertion labels, respectively. SS is the count of substitution spans, weighted by the maximum number of words in each span. These one-to-many or many-to-many word alignments indicate phonetic confusability as the cause of the error.

3. EXPERIMENTS

Following the experimental framework of [4], we perform our experiments on an intersection of the ASR and MT results of the IWSLT 2013 evaluation campaign [5], which focused on the translation of TED talks. These 580 utterances map a subset of the ASR hypotheses provided by 8 ASR systems to the corresponding MT inputs in the English-French MT track. The unpunctuated MT input serves as the ASR reference data. Eight French human post-edited references serve as the MT references in the SLT analysis.

In order to minimize the effects of formatting issues on our experimental results, the ASR hypotheses are evaluated, normalized, recased, and punctuated according to the MT input and are translated by a baseline English-French Moses SMT system, corresponding to the WIT³ data from 2014 [6]. Since we desire to keep the ASR reference intact, we apply oracle-based text normalization and punctuation insertion techniques similar to that of [4], instead of applying a general .glm normalization file to the both the ASR reference and hypothesis. We use POWER to align and normalize hypothesis words with respect to the reference. POWER uses the Festival TTS system with the CMU English pronunciation dictionary [7] to convert words into phonemes. Prior to other normalization steps, we use the VARCON tool from SCOWL² to convert British English words in the ASR hypotheses to American English. We also use libraries from NLTK [8] to annotate ASR errors with part-of-speech and word class information, as well as lemmatization for morphological analysis.

We conduct two sets of experiments. First, we analyze the ASR error annotations given by the WER and POWER alignments to measure the effects of Levenshtein alignment heuristics on the reported results. In the second set of experiments,

²<https://github.com/kevina/wordlist>

System	Data set				ASR WER ↓		SLT	
	tokens	open	closed	ratio	orig	norm	BLEU ↑	TER ↓
fbk	10095	5581	4514	1.24	21.3	16.5	51.9	38.5
kit	10141	5571	4570	1.22	15.1	10.1	55.4	35.2
mitll	10147	5594	4553	1.23	16.3	11.4	55.0	35.8
naist	10076	5571	4505	1.24	15.6	10.5	55.1	35.3
nict	10165	5595	4570	1.22	14.4	9.2	56.5	34.3
prke	10106	5545	4561	1.22	21.2	16.5	52.1	38.4
rwth	10160	5563	4597	1.21	16.4	11.6	54.3	36.2
uedin	10151	5592	4559	1.23	17.1	12.3	54.6	36.1
gold	10158	5614	4544	1.24	0.0	0.0	62.9	29.1

Table 1: Statistics for each ASR system on the ratio of open to closed class words by ASR system; ASR WER scores before and after text normalization; and English-French translation scores for normalized and punctuated ASR hypotheses, compared to the ASR reference (gold).

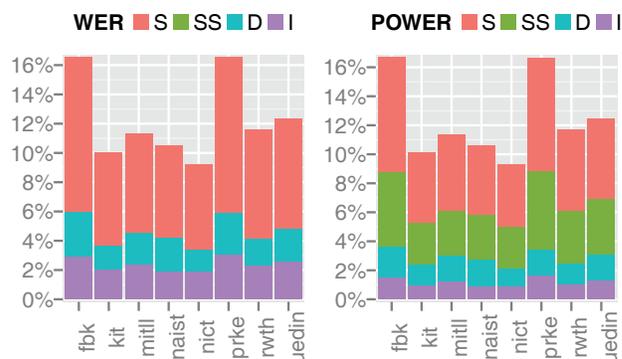


Fig. 4: Distribution of error types for WER (left) and POWER (right) for each IWSLT 2013 ASR evaluation participant.

we construct several mixed-effects models [9] that measure the contribution of various ASR error types on MT errors. Table 1 provides summary statistics on the words in each ASR hypothesis, the WER scores before and after text normalization for each ASR output, and scores from two translation metrics. In particular, the POWER-driven text normalization reduces each system’s WER scores by 5%.

4. ASR ERROR ANALYSIS

In this set of experiments, we observe the contribution of particular error types to the global WER and POWER scores for each ASR system. We outline the shortcomings of WER’s statistics due to the erratic behavior of Levenshtein aligners.

4.1. Basic Levenshtein error types

We begin by looking at the basic ASR error types (S, D, I, and SS), which implicitly contain no linguistic information. Fig. 4 shows the contribution of the basic Levenshtein error types toward the error rate score for each ASR system. According to WER, substitutions intuitively make up the majority of error types (62.3% ± 0.7%). Across all ASR systems, WER suggests that the number of deletions are slightly lower

SysID	SS.ref	SS.hyp	SS: ref>1	SS: hyp>1	SS: ref>1 & hyp>1
fbk	0.036	0.040	0.450	0.615	0.065
kit	0.017	0.025	0.254	0.800	0.054
mitll	0.019	0.026	0.321	0.714	0.036
naist	0.019	0.026	0.336	0.715	0.051
nict	0.018	0.025	0.305	0.763	0.069
prke	0.039	0.042	0.488	0.585	0.073
rwth	0.023	0.032	0.310	0.737	0.047
uedin	0.025	0.032	0.317	0.700	0.017

Table 2: Left: Percentage of reference/hypothesis words appearing in a substitution span. Right: Percentage of substitution spans containing multiple reference words, multiple hypothesis words, or both.

than the number of insertion errors ($17.9\% \pm 0.7\%$ deletions and $19.8\% \pm 0.5\%$ insertions).

However, POWER suggests that roughly half of these alleged insertion errors ($10.0\% \pm 0.3\%$) are instances where a larger reference word is being hypothesized as a homophonic sequence of shorter words. Likewise, a portion of “deletion” errors are instances where multiple reference words were hypothesized as a longer homophonic word ($4.1\% \pm 0.5\%$). Since these substitution span errors are typically cases of one-to-many alignments, the number of reported word-level substitution errors are reduced. As such, POWER claims that $30.0\% (\pm 0.7\%)$ of the errors are substitution spans involving homophony, leaving $13.8\% (\pm 0.8\%)$ of the remaining errors as deletions and only $9.8\% (\pm 0.3\%)$ as insertions whose pronunciations do not align to any words – both measures are substantially lower than those reported by WER. The remaining $46.4\% (\pm 0.5\%)$ are word-level substitutions.

We can corroborate this by observing in Table 2 that, across all ASR systems, $70.4\% (\pm 2.5\%)$ of the substitution spans involve multiple hypothesis words, while only $34.8\% (\pm 2.8\%)$ contain multiple reference words. The first figure may be explained by the presence of out-of-vocabulary words in the ASR reference, as well as the effects of domain variation in the evaluation data. The alignment of multiple reference words to a single hypothesis word may be indicative of mispronunciations and/or underarticulation by the speaker. These hypotheses should be explored in future work.

4.2. Word classes and morphology

Given the inconsistent error labeling in WER, which types of errors are actually being skewed by false alignments? To answer this question, we annotate the reference and hypothesis words by their word class and observe their alignment statistics. We additionally apply lemmatization to distinguish morphological errors from other substitution types. According to the word statistics in Table 1, the ratio of open to closed class words remains the same across each ASR hypothesis and the reference (gold). The proportion of errors associated with each ASR error type is shown in Table 3.

Word-level substitution errors. Both WER and POWER report that the majority of substitution errors are within the same word class. While the proportion of closed-closed class substitutions remain the same, POWER reports 8% fewer

ErrorType	WER	POWER	WER Rank	POWER Rank
S.open_open	0.299	0.219	1	1
SS.open_span		0.186		2
S.closed_closed	0.148	0.140	2	3
D.closed	0.112	0.097	4	4
S.open_closed	0.107	0.069	4	6
SS.span_open		0.069		6
I.closed	0.101	0.065	6	7
D.open	0.067	0.041	8	8
S.closed_open	0.069	0.036	8	8
I.open	0.096	0.033	6	10
SS.span_closed		0.019		12
SS.span_span		0.016		12
SS.closed_span		0.010		13

Table 3: Proportion of ASR error types by word class, averaged across all ASR systems and ranked by importance. Substitution labels (S, SS) show the alignment from reference class to hypothesis class. Substitution spans (SS) contain a span of words aligned either to a single word or another span.

open-open class substitution errors, which are often instances of substitution error spans containing a word-level substitution error and one or more short function words (e.g. *Brown in*→*brahmin* from Fig. 3). Of the open-open class substitution errors, 5.4% are morphological errors. POWER likewise reports 7% fewer cross-class substitution errors, many of which are attributed to the correction of misalignments.

Deletions and Insertions. According to WER, deletion and insertion errors account for $37.7\% (\pm 0.7\%)$ of all errors. WER marks nearly as many open class insertions as closed class insertions, but suggests that closed class deletions are more prominent than open class ones ($6.7\% \pm 0.4\%$ open versus $11.2\% \pm 0.5\%$ closed class deletions). However, with POWER, deletion and insertion errors only account for $23.6\% (\pm 0.8\%)$ of all errors, with the majority of the reduction attributed to fewer open class insertion errors ($3.3\% \pm 0.1\%$). An example of a corrected open class “deletion” is *Stanford*→*stamp or* from Fig. 3.

Substitution spans. The majority of substitution spans have a single open class reference word ($18.6\% \pm 0.7\%$), such as *anatomy*→*and that to me* in Fig. 3; these represent the second most common POWER error type. Likewise, the presence of a substitution span in the ASR reference indicates that the hypothesis word is likely to be a content word ($6.9\% \pm 0.8\%$). Closed class function words are unlikely to be aligned to substitution spans ($2.9\% \pm 0.3\%$), since most have few syllables that cannot easily be mistaken for multiple words. Instead, as shown in Table 3, closed class words are more likely to be deletion or insertion errors.

Table 4 provides confusion pair examples from FBK’s ASR system output that demonstrate the utility of POWER. Word confusion pairs such as *a*→*today* are likely errors induced by an ASR language model that biases the acoustic model to artificially recognize non-existent phonemes. Likewise, POWER is able to provide insight that *crude*→*crudely* is not a morphological error, but rather another language model-induced bias that considers *leaf* an unlikely successor to *crude*. Other confusion pairs include word normalizations, affix errors, and phonetic confusions.

WER		POWER	
Reference	Hypothesis	Reference	Hypothesis
a	today	a day	today
ascending	and	ascending	and sending
anesthetize	and	anesthetize	and decent size
butchering	the	butchering	maturing
centigrade	cents	centigrade	cents a great
crude	crudely	crude leaf	crudely
cyclones	soy	cyclones	soy clones
face-to-face	face	face-to-face	face to face
of	obama	of anatomic	obama panic

Table 4: Confusion pair examples using WER and POWER.

5. SLT ERROR ANALYSIS

Given that POWER yields a significantly different distribution of error types, how can it be leveraged to understand the impact of ASR errors on downstream natural language processing tasks? We turn our attention to the translation of TED talks from English to French. Similar to [4], we measure the impact of utterance-level ASR errors on their associated translation score by measuring the increase in translation error rate (ΔTER) [10] against a gold standard translation which contains no ASR errors. We use linear mixed-effects regression models to measure the importance of each ASR error type, taking into consideration random effects caused by an ASR system and the particular features of each ASR utterance. We use the R [11] implementation of mixed-effects models in the *lme4* library [12]. All of our models are fit using maximum likelihood and incorporate random intercepts for each ASR utterance (labeled as *UttrID*) and ASR system (labeled as *SysID*), as well as a random slope by the WER score. We use the repeated observations of our 580 speech utterances by eight ASR systems, yielding a total of 4,640 observations. As fixed effects, we normalize the counts of each ASR error type by the length of the ASR reference for each utterance in order to consider its contribution toward the utterance-level WER score. In each model, *SysID* was not significant, with a standard deviation near zero.

5.1. WER versus POWER features

Our first experiments consider the ASR error labels provided by WER and POWER. Our baseline considers a single error feature, corresponding to the WER score for each utterance. We compare it to two mixed-effects models and report their coefficients in Table 5. *WER.basic* is trained with WER’s basic substitution (WER.S), deletion (WER.D), and insertion (WER.I) labels; *POWER.basic* is trained additionally with POWER’s substitution span labels (WER.SS). Both sets of features are normalized by the reference length in order to be a decomposition of the WER metric.

As in [4], we observe a significant difference between *WER.basic* and the baseline, rejecting the null hypothesis that each basic ASR error type contributes equally to translation quality, in terms of ΔTER ($\chi^2(2) = 16.922, p < 2.12 \times 10^{-4}$). We additionally observe a significant difference between the standard WER-aligned error types (*WER.basic*) and the POWER-aligned error types (*POWER.basic*) that include

Table 5: Fixed effects coefficients and 95% confidence intervals for the first three mixed-effects models, which measure the effect of ASR error types on ΔTER for English-French SLT. The baseline encapsulates all error types in a single WER measure, while the subsequent models use WER and POWER-aligned error types.

	WER (1)	WER.basic (2)	POWER.basic (3)
WER	0.630*** (0.586,0.674)		
WER.D		0.564*** (0.506,0.622)	0.615*** (0.556,0.674)
WER.I		0.707*** (0.642,0.772)	0.829*** (0.753,0.906)
WER.S		0.624*** (0.578,0.671)	0.649*** (0.601,0.696)
WER.SS			0.535*** (0.487,0.584)
Constant	0.001 (-0.003,0.004)	0.001 (-0.002,0.004)	-0.0001 (-0.003,0.003)
Observations	4,640	4,640	4,640
Log Likelihood	6,172.170	6,180.631	6,194.288
Akaike Inf. Crit.	-12,330.340	-12,343.260	-12,368.580
Bayesian Inf. Crit.	-12,285.240	-12,285.280	-12,304.150

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

substitution spans ($\chi^2(1) = 27.314, p = 1.73 \times 10^{-7}$), indicating that substitution spans are a significant predictor of translation quality. As shown in Table 5, while the impact of substitution errors remains in principle the same, the impact of insertions increase sharply, both due to the higher quality of the error labels and their lower frequency. *POWER.basic* indicates that an utterance with a WER (or equivalently, POWER) score of 10% as insertion errors would expect an increase in TER by $0.1 \times 0.829 - 0.0001 = 8.3\%$, while 10% in substitution errors would correspond to a TER increase of $0.1 \times 0.649 - 0.0001 = 6.49\%$.

5.2. Frequency-weighted ASR errors

While the coefficients in Table 5 show the expected increase in TER for each percentage of WER associated with a particular error type, an error type with a high coefficient but a low frequency may not be important from an error correction standpoint. Ideally, we wish to measure which ASR errors are particularly problematic for a given SLT task.

We construct an additional mixed-effects model using the word class-annotated error types from Section 4.2. We compute frequency-weighted scores for each error type, based on the observations in our data set. Considering the fixed and random effect scores on each utterance, we measure the average weighted contribution of each ASR error type toward the ΔTER measure. In other words, if we observe one ASR error of a particular type, how much is it expected to degrade the translation quality? By doing so, we seek to rank the importance of each error type. Table 6 reports the mean and standard error for each weighted error type, using the word class-annotated error types provided by POWER. We observe

ErrorType	coef	weight-mean	weight-se	SLT Rank
WER.S.open_open	0.687	0.0175	0.0006	1
WER.S.closed_closed	0.585	0.0132	0.0005	2
WER.SS.open_span	0.723	0.0123	0.0006	3
WER.D.closed	0.451	0.0069	0.0004	4
WER.I.closed	0.548	0.0059	0.0004	6
WER.S.open_closed	0.663	0.0057	0.0003	6
WER.D.open	0.546	0.0048	0.0005	7
WER.I.open	0.553	0.0044	0.0004	9
WER.S.closed_open	0.590	0.0036	0.0002	9
WER.SS.span_open	0.802	0.0038	0.0003	9
WER.SS.span_closed	0.757	0.0016	0.0002	12
WER.SS.span_span	1.036	0.0015	0.0003	12
WER.SS.closed_span	0.713	0.0011	0.0002	13
(Intercept)	0.000	0.0002	0.0003	14

Table 6: Mixed-effects coefficients (coef) for POWER ASR error types with word class annotations, and their mean frequency-weighted contributions toward translation Δ TER (weight-mean).

that, similar to the ASR-only experiments, within-class substitution errors have the highest frequency-weighted contribution toward Δ TER. While substitution spans containing open class reference words have a high weighted score, substitution spans with open class hypothesis words have a substantially lower weighted score.

6. DISCUSSION

Based on the error statistics provided above and recorded in Table 3, we identify the following error types as interesting to focus on when constructing models to cope with ASR errors. 16.2% ($\pm 0.6\%$) of the ASR errors are either insertions or deletions on closed class words. These types are also ranked highly in our SLT experiments. While a recovery model to insert or delete hypothesis words is non-trivial, we consider closed class words to be low-hanging fruit, since the number of alternative words are small and a language model would likely have statistics that support their inclusion or removal. While closed class words can be under-articulated, they receive the majority of their support from the language model due to the large amount of observations. The advantage of using POWER to model errors is that we have more confidence that the words we mark as deletions or insertions during training/development are really unaligned words.

Likewise, it is useful to consider the effects of substitution spans, as they are among the most frequent errors caused by ASR systems. By identifying substitution error spans, we are able to capture a consecutive string of words that can significantly alter the meaning of a sentence. However, oftentimes they are due to homophonic errors where an ASR system may have reasonable confidence in the phonemes detected, but due to the interaction between the acoustic and language models, a shorter sequence of similar-sounding words was selected. It would be worthwhile to identify common phonetic error patterns to either rescore ASR hypotheses or carry forward the ambiguity of a span of words in the hypothesis to allow the downstream process to decide which similar-sounding alternative makes the most sense.

7. RELATED WORK

Mixed-effects models were first used in ASR error analysis in [13] to analyze the effects of lexical, prosodic, contextual, and disfluency features of individual words on WER. They show that of the various disfluency types, word fragments, non-final repetitions, and words preceding fragments have a significant impact on WER. Our work proposes a phonetically-oriented word alignment process that is more successful in aligning words of the same word class. Such an alignment process would alter the individual WER measure proposed in [13], which could provide more reliable results.

A related area of work is ASR confidence estimation, which seeks to label erroneous words in an ASR hypothesis. [14] uses the comparison of phones in a strong ASR system and a weak ASR system without a language model as features for error detection. Regions where the difference is large indicate a higher likelihood of errors. Other approaches include using ASR consensus votes as well as recurrent neural networks to capture longer contexts [15].

On downstream tasks such as speech translation, [16, 17] propose ASR channel modeling techniques that rely on the concept of phonetic confusability to convert error-free source language phrases into ASR-like outputs in order to model ASR errors during machine translation model training. Phonetically-oriented alignments could be used in either approach to identify error regions during training to focus the channel model on confusable words. Our use of a TTS analyzer to generate pronunciation sequences on ASR references and hypotheses is based on [17].

8. CONCLUSION

We have developed a phonetically-oriented word alignment pipeline as an extension to Word Error Rate’s Levenshtein aligner. Spans of adjacent Levenshtein errors containing minimally one substitution error are converted into phonemes with word and syllable boundaries. A second Levenshtein alignment process on phonemes is carried out and the alignment information is used to guide the word alignment process. We demonstrate that our phonetically-oriented word alignments generate virtually the same error rate score as WER, with the added benefit of more reliable substitution error tags, and a reduction of erroneous deletion and insertion error labels on open class words. We demonstrate that the use of phonetically-oriented error labels significantly alters the statistics gathered from error analyses on ASR outputs. Additionally, for speech translation error analysis tasks, we demonstrate that our phonetically-oriented word error alignments result in better error models in mixed-effects modeling. We demonstrate that homophonic error spans comprise a significant portion of ASR errors with a large impact on speech translation quality and deserve to be considered as an additional substitution error type in error recovery efforts. Our POWER software is available as open source software for the research community at <https://github.com/NickRuiz/power-asr>.

9. REFERENCES

- [1] E. Matusov, S. Kanthak, and H. Ney, “Integrating speech recognition and machine translation: Where do we stand?,” in *Proceedings of ICASSP*, Toulouse, France, 2006, pp. 1217–1220.
- [2] N. Bertoldi, R. Zens, and M. Federico, “Speech translation by confusion network decoding,” in *Proceedings of ICASSP*, Honolulu, HA, 2007, pp. 1297–1300.
- [3] Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal, “Recent efforts in spoken language processing,” *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, May 2008.
- [4] Nicholas Ruiz and Marcello Federico, “Assessing the Impact of Speech Recognition Errors on Machine Translation Quality,” in *Association for Machine Translation in the Americas (AMTA)*, Vancouver, Canada, 2014, pp. 261–274.
- [5] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico, “Report on the 10th IWSLT Evaluation Campaign,” in *Proc. of the International Workshop on Spoken Language Translation*, December 2013.
- [6] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico, “Report on the 11th IWSLT Evaluation Campaign,” in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.
- [7] Alan W. Black and Paul A. Taylor, “The Festival Speech Synthesis System: System documentation,” Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997, Available at <http://www.cstr.ed.ac.uk/projects/festival.html>.
- [8] Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*, O’Reilly Media, Inc., 1st edition, 2009.
- [9] S. R. Searle, “Prediction, mixed models, and variance components,” Tech. Rep. BU-468-M, Biometrics Unit, Cornell University, June 1973.
- [10] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul, “A study of translation edit rate with targeted human annotation,” in *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August 2006.
- [11] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [12] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker, *lme4: Linear mixed-effects models using Eigen and S4*, 2014, R package version 1.1-6.
- [13] Sharon Goldwater, Daniel Jurafsky, and Christopher D. Manning, “Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates,” *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.
- [14] Christopher White, Geoffrey Zweig, Lukas Burget, Petr Schwarz, and Hynek Hermansky, “Confidence estimation, oov detection and language id using phone-to-word transduction and phone-level alignments,” in *Proceedings of ICASSP*, 2008.
- [15] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang, “ASR error detection using recurrent neural network language model and complementary ASR,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 2312–2316, IEEE.
- [16] Yulia Tsvetkov, Florian Metze, and Chris Dyer, “Augmenting translation models with simulated acoustic confusions for improved spoken language translation,” in *EACL*, 2014, pp. 616–625.
- [17] Nicholas Ruiz, Qin Gao, William Lewis, and Marcello Federico, “Adapting Machine Translation Models toward Misrecognized Speech with Text-to-Speech Pronunciation Rules and Acoustic Confusability,” in *Proceedings of Interspeech*, Dresden, Germany, September 2015, ISCA.