PARSING-BASED OBJECTIVE FUNCTIONS FOR SPEECH RECOGNITION IN TRANSLATION APPLICATIONS

D. Hillard†, M. Hwang†, M. Harper‡, M. Ostendorf†,

†University of Washington, Electrical Engineering Dept., Seattle, WA
‡University of Maryland, Computer Science Dept., College Park, MD

ABSTRACT

This paper looks at a parsing-based alternative to word error rate (WER) for optimizing recognition, SParseval, hypothesizing that it may be a better objective for applications such as translation. We find that SParseval is more correlated than WER with human measures of subsequent translation performance, but that optimizing explicitly for SParseval does not give a significant reduction in translation error as measured by automatic methods based on a single translation reference. However, anecdotal examples indicate that SParseval does improve automatic speech recognition (ASR) results, leaving open the possibility that it may be more useful in the future or for other language processing tasks.

Index Terms— speech translation, speech recognition objective, parsing

1. INTRODUCTION

Recent work on machine translation (MT) of speech has provided mixed results on the impact of speech recognition errors. One study shows that recognition errors, source, and domain-mismatch are important variables in predicting translation errors [1]. But other researchers report that improvements in ASR error do not consistently lead to gains in translation performance. Certainly large ASR improvements are likely to benefit translation, but could it be the case that smaller gains (e.g., 10% reduction in error) simply won't matter until the performance of MT improves? Or could it be that word error rate (WER) is simply not the right objective?

The main problem with WER as an objective for speech recognition when the end goal is language processing (whether machine translation, information extraction, or other types of processing) is that WER counts all errors equally. Intuitively, it seems clear that all words are not equally important – filled pauses being an extreme example. Scoring methods sometimes account for such problems with "allowable errors," but typically automatic performance optimization is based on simpler metrics that do not make these distinctions. In addition, it is probably not the case that all error types are equal. For example, a deletion is probably much worse than an insertion or substitution, since all information is lost when a word is deleted but at least some phonetic cues are present with other error types. Ide-

ally, one would use a weighted word error rate, where the weighting function was appropriate for the target application.

A problem with learning a weighting function is that weights dependent only on error types are probably not sufficiently rich and vocabulary-dependent weights yield too many free parameters. The alternative would be to use word-based weights that are motivated by the task but not automatically learned. For example, one might use an information-based measure, such as the inverse-document frequency weights often used in information retrieval and topic clustering. The approach explored here is a parsing-based word string scoring function, specifically SParseval. As described in Section 2, the head-dependency scoring method in SParseval effectively puts more weight on syntactic head words, although it is not strictly a weighted word error rate. An advantage of a parsing-based measure is that parses are the first stage of processing for many more complex tasks, such as information extraction and in some MT systems. As will be shown in Section 3, even when the MT system does not explicitly use syntax, the SParseval score is more correlated with MT quality than is character error rate (CER), a more commonly used measure than WER for Mandarin. This motivated us to exploit parsing-based ASR objectives in discriminative score combination for speech recognition (Section 4) and investigate the impact on MT (Section 5). Analyzing MT performance in terms of automatic metrics unfortunately did not show significant performance gains, and we speculate on the implications in Section 6.

2. SPARSEVAL

While spoken language processing can benefit from parsing as much as written language processing, the standard Parseval metrics [2] and their canonical implementation (EVALB)[3] are undefined when the words input to the parser do not match the reference words in the gold standard parse trees exactly. To address the fact that output from an ASR system is likely to contain word errors and the segmentation of these words into sentences is likely to differ from that used in the gold standard parses, Roark et al. [4] have developed the publicly available SParseval tool for scoring spoken language parses against gold standard parses. This tool supports both bracket and head dependency scoring and produces recall, precision, and F-score for each method; here we use dependency scoring. The process of dependency scoring maps each word in a sentence (for both the hypothesize words+parse and the reference) to a triple that includes the word, the head word that it is dependent on, and the type of dependency. Scoring is based on the number of matches of these triples between the reference and hypothesized parse. (As in EVALB, the labels can optionally be ignored.) The more links there are to a headword, the more triples in which it will be present, and hence an error on that word will contribute to multiple errors for the triples, effec-

tively putting a higher weight on these headwords. The tool provides dependency scores based on all of the head dependencies extracted from the trees, as well as a more focused set of open class dependencies, which involve open class content words. Dependency scoring utilizes a head percolation table, and dependency scores can be computed either with or without the word-level alignment constraints needed to calculate bracket scores.

In the experiments discussed in this paper, we compute head dependency scores based on a head table that we developed for the LDC Chinese Treebank 6.0. When there is a mismatch between the sentence segmentations of the reference and ASR transcripts, as will be the case in the correlation studies reported in Section 3, we use alignment-based head-dependency scores calculated over the spoken document. On the other hand, when reference and ASR hypotheses cover the same time span, as in the experiments reported in Section 5, we utilize head-dependency scores without alignment constraints. Although alignment adds an extra match constraint that can slightly reduce dependency scores, the reduction is negligible when scoring the parses on short segments.

3. CORRELATION STUDY

We first assessed the usefulness of SParseval as an ASR objective function in speech translation applications by computing the correlation of ASR scoring criteria on Mandarin speech (CER and SParseval) with error measures on the English translations. Our main interest was in improving HTER, a human-based error measure that computes the translation error rate (TER) between the MT hypothesis and a human-edited version that reflects the same meaning as a reference translation with a minimal number of edits [5]. However, since HTER is costly to obtain, we also compare to TER, which can be automatically computed (on the single available reference).

The study was done on the broadcast news (BN) portion of the Mandarin GALE [6] 2007 evaluation test set, which consists of 66 documents for which HTER results were available. The documents are typically a news story or portion of a story that include a few sentences. We also looked at a subset of 30 documents for which the average CER is less than 3%, since one would expect the ASR differences to matter less in this range.

To obtain the SParseval results, both the Chinese reference transcript and the ASR hypothesis were parsed automatically, as described in Section 5.2, and head dependency scores were computed using the SParseval tool. Since the parser used human-annotated sentence segmentations for the reference and automatic segmentation for the ASR hypotheses, scoring was based on document-level alignment to handle sentence segmentation mismatch.

The results are reported in Table 1, which shows that SParseval is substantially more correlated with both HTER and TER than WER. Even for the low error rate subset, where one would expect less benefit from ASR improvements and hence a lower correlation, there is a big difference in the correlations. Of course, in no cases are the correlations large, due to the fact that MT modeling factors cause many errors (separate from errors that can be attributed to ASR).

We investigated whether automatic sentence segmentation hurts the usefulness of the SParseval objective, since it is known that parse scores degrade considerably with segmentation errors. Comparing results using automatic vs. oracle segmentation, we found that the correlation was in fact higher for the automatic case. We hypothesize that SParseval is implicitly incorporating segmentation error into the score, which is useful for predicting MT performance since it too is sensitive to segmentation error. We also experimented with broadcast conversations (BC). In initial studies using a parser trained

Table 1. Correlation between two ASR scores (CER and SParseval) and two MT scores (HTER and TER).

Test Set	ASR Score	MT Score	
		HTER	TER
Eval07-BN	CER	0.32	0.46
	SParseval	0.44	0.61
CER < 3% subset	CER	0.19	0.26
	SParseval	0.38	0.47

on BN, correlation with HTER was much lower for SParseval compared to CER. After BC Treebanked data became available and the parser was retrained, the correlation of SParseval with HTER improved substantially on the BC data, though it was still slightly lower than that of CER with HTER (0.26 vs. 0.37).

4. ASR OBJECTIVES FOR N-BEST RESCORING

The ASR objective function (or score) impacts the system at the N-best rescoring stage, where weights associated with different knowledge sources (acoustic model, language model(s), word count) are trained to optimize the score of the top ranking hypothesis based on the weighted combination. We use the weight optimization function in the SRILM toolkit [7], specifically `nbest-optimize`, which uses a simplex-based "Amoeba" search on the objective function [8]. The optimal parameters returned by the search are then used in ASR decoding. In a 2-system combination framework, weights are optimized separately for N-best lists of the two systems, and the two N-best lists are then combined at the character-level via confusion network combination [9] with the posterior probability computed by applying the optimized weighting parameters.

The typical approach to ASR is to optimize weights to minimize CER (or WER) with respect to the correct transcription. Building on the framework of [10], we introduce an alternative parse-based optimization criterion that specifically maximizes dependency-pair F-score by minimizing $\hat{e} = L \times (1 - F)$. We include L, the number of words in the reference segment, to avoid over-weighting short segments. The dependency F-score for a hypothesis is based on the reference transcription with an automatically generated parse.

In combining different knowledge sources, a problem arises when there are cases where there is no score. For example, an utterance with only laughter or noise and no words will have no parse score. We add another "knowledge source" that is a simple indicator of these conditions so that we can learn a compensating weight, similar to the word insertion penalty.

5. MT EXPERIMENTS

5.1. ASR System

The ASR system adopted in this paper is the one used in our GALE 2007 evaluation [11], except that there is no cross adaptation with RWTH Aachen University. In brief, two acoustic models (AMs) were trained on 870 hours of speech data, one based on PLP+pitch features, the other MFCC+pitch+MLP (multi-layer perceptron based phoneme posterior features). Maximum-likelihood based word segmentation on the Chinese training text was used, based on 60,000 Chinese lexical words, and n-gram (up to 4-gram) language models were trained on over 1 billion words of text. Each testing show was automatically segmented into "utterances," based on long pauses

and automatic speaker boundaries prior to recognition. The two AM systems cross adapted each other and produced 1000 best hypotheses each, for each testing utterance, and the two N-best lists were combined via a confusion network at the character-level. The best character sequence was then re-segmented into words with the same word segmenter used during training.

5.2. Mandarin Parser

The Mandarin parser is based on a modification of the Berkeley unlexicalized parser [12]. This parser uses a new approach for learning that begins with a PCFG grammar derived from a raw Treebank, and then iteratively refines the grammar. During each stage, all symbols are first split in two (e.g., NN may become NN-1 and NN-2) and a refined grammar is estimated using a variant of the forward-backward algorithm; next less helpful symbol splits are retracted based on a likelihood gain approximation; and finally a simple smoothing strategy is applied. This method can learn to distinguish alternative uses of words and phrases, thereby producing higher quality parses. The original Berkeley parser achieved a bracketing F-score of 82.4%[1] on the Chinese Treebank 5.2, which exceeds the performance of state-of-the-art parsers on Chinese (typical F-scores are between 79% and 81%). Beginning with this capability, we updated the parser in a number of ways. We addressed unknown words by using all characters of an unknown word to estimate word probability (building on our Mandarin part-of-speech (POS) tagging work [13]), improving the F-score to 82.8%. We also considered differences in how Mandarin was treebanked as an avenue for improving parse accuracy [14], removing rarely invoked unary rules from the trees prior to training and obtaining an F-score of 84.6%. We also added parent annotations to the Treebank prior to training, since the Berkeley parser uses little explicit context of a symbol, resulting in F-score of 84.93%. Finally, we added training data from the recent release of CTB6.0, including Broadcast News trees. Using this larger training set together with our other enhancements, the F-score of our parser on the same test set is 86.5%.

For experiments here, we trained the parser on a text-normalized version of CTB6.0 (i.e., all Arabic digits were replaced by verbal tokens in the tree) with punctuation removed to better match the conditions to which the parser would be applied. We parsed both the Chinese reference transcription and ASR hypotheses using the same parser with MAX-RULE-PRODUCT decoding.

5.3. Annotation

The ASR system outputs the best sequence of words (with time stamps) corresponding to the specified testing segments in the input shows. This information is then sent to our annotation module before translation. The annotation module consists of many functions, including speaker diarization, sentence unit (SU) detection, POS tagging, punctuation prediction, inverse text normalization (from spoken numbers to written Arabic digits) and named-entity identification. The output of the annotation module in this paper was simply a sequence of Chinese sentences with sentence boundaries (but not punctuation) and with numbers in digit form.

5.4. MT System

For automatic translation, we used the state-of-the-art phrase-based statistical machine translation system built by RWTH Aachen Uni-

versity [15]. It was trained using the LDC bilingual training corpora, consisting of 7 million pairs of Chinese-English sentences. The lexicon model is bi-directional, i.e., both $p(e|f)$ and $p(f|e)$ were trained and interpolated. The target language model $p(\mathbf{e})$ was a 6-gram English LM trained on the English part of the bilingual training corpora and additional monolingual English data from the Gigaword corpus. The total amount of training text for this 6-gram LM was about 600M words. Note that this LM is different from the ASR n-gram; it contains punctuation and digits, among other things. The system also used a phrase reordering model [15], along with other penalty models. The total translation score is a linear combination of the log likelihoods of the individual models:

$$\mathbf{e}^* = \operatorname*{argmax}_{\mathbf{e}} \sum_m \lambda_m \log P_m(\mathbf{f}, \mathbf{e})$$

MT takes the output from the annotation module and translates sentence by sentence, as defined by the SU annotation module. Punctuation is automatically derived during MT decoding with the Chinese-to-English phrase table, which was trained by removing punctuation marks from the Chinese training sentences while keeping them in the corresponding English training sentences.

5.5. Data Description

We used the GALE Mandarin 2007 audio development set as our testbed. Our goal is to translate from Mandarin speech to English text. Since our initial Chinese parser was trained exclusively on newswire data, we used only the broadcast news part of this set, denoted as dev07-bn here. Dev07-bn comes from 40 different Chinese shows aired in November 2006, consisting of 54 different documents, for a total of 108 minutes and 19,000 Chinese characters. There are 524 sentences in the English gold translations, each with only a single gold translation.

5.6. Results

To better understand the impact of the two objective functions (CER and SParseval), we compare them in both automatic and oracle conditions, where "automatic" involves N-best rescoring with the discriminatively trained weights. We also try to improve results for the SParseval objective by including an additional parse confidence knowledge source (log parse probabilities) with the acoustic and language model scores. This leads to five experimental conditions:

1. CER: This is the baseline system.
2. SParseval dependency F-score: We use the "error" $\hat{e}$, described in Section 4 with the standard knowledge sources.
3. SParseval F-score+Confidence: We use $\hat{e}$ as above, but adding the parse confidence knowledge source.
4. SParseval Oracle (S-oracle): As an oracle comparison, we select the hypothesis from the N-best lists with the best F-score.
5. CER-oracle: The comparable oracle for CER is the hypothesis that has the minimum CER among all hypotheses.

In Table 2, we report ASR CER, MT TER [5] and MT BLEU [16] scores for each experiment, together with the number of SUs determined by the annotation module. While computing MT errors, we ignore differences in case and punctuation for simplicity. MT errors are computed at a per-sentence basis, based on the sentence definition in the gold translation. To do that, we automatically segment the machine-translation output (now in English per annotated

---

[1]This is slightly lower than reported in [12]; their score ignored two long sentences that returned null parses parses in an earlier version of the parser.

SU sentence) into the same number of sentences as the gold translation, by minimizing the word-alignment errors [17]. MT errors are high, in part due to the fact that there is only one gold translation per sentence. Unfortunately, we did not get improvements in translation by optimizing directly for SParseval in recognition, but we also found little margin for improvement in BLEU and HTER between the baseline condition and the best case oracle condition.

Table 2. ASR scores and MT scores on dev07-bn.

Experiment	#SU	CER	TER	BLEU
(1) CER	905	3.4%	70.4%	18.9
(2) F-score	904	3.5%	70.4%	18.8
(3) F-score+Conf	905	3.4%	70.3%	18.9
(4) S-Oracle	904	1.2%	69.7%	19.1
(5) CER-Oracle	903	0.9%	69.5%	19.3

The example below illustrates the types of translation errors that the MT system made given different input. It first shows the reference transcription in the source language and the reference translation in the target language, followed by the source-target hypotheses of Experiment (1) and (2). Mis-recognitions/mis-translations are underlined.

REF: 这 种 防止 翻车 的 新 系统 相信 每 年 可以 在 美 国 挽救 至少 上万 条 人命
REF: It is believed that the new rollover-prevention system for automobiles can save at least 10,000 lives each year in the United States.

Expt (1): CER
HYP: 这 种 防止 翻车 的 新 系统 将 信 每 年 可 以 在 美国 挽救 至少 上万 条 人命
HYP: this new system will prevent ” derailed by letter each year can save at least 10,000 lives in the united states.

Expt (2): F-score
HYP: 这 种 防止 翻车 的 新 系统 相信 每 年 可以 在 美 国 挽救 至少 上万 条 人命
HYP: this new system to prevent nobody-was believe that each year in the united states can save at least 10,000 lives .

In (1), the verb "is believed" was misrecognized as two characters that do not correspond to any common meaningful Chinese verb and therefore led to translation errors. In (2), we were able to pick the perfect Chinese word sequence because it could be parsed easily. Unfortunately, our MT had so much difficulty in translating the word 翻车 (automobile rollover) that it did not yield a better translation.

In subsequent experiments with an improved ASR system and genre-matched parsers, we were able to assess performance on both BN and BC. In this case, experiments were based on reference segmentations. There was a small but insignificant gain BLEU and TER for the BN data, but a larger gain for BC (increase of BLEU from 12.4 to 12.7). Further study is needed to assess the impact of segmentation.

6. SUMMARY

In summary, this work has tried to address the question of whether improvements to ASR could have a greater impact on MT if optimized in terms of a different objective function than WER (or CER for Mandarin). While we found that SParseval scoring of ASR outputs is more correlated with human evaluations of translations than CER, we did not get improvements in translation by optimizing directly for SParseval in recognition. However, we also found that there is not much margin for improvement in BLEU and TER between the baseline and best case oracle conditions. In addition, anecdotal examples indicate that SParseval does improve ASR results, and preliminary experiments with BC show some gains.

The mixed results arguably support the hypothesis that moderate ASR improvements will not have visible impact on MT until MT is further improved. However, it may be that when MT has some extent of syntax knowledge, it would benefit from ASR output with improved syntactic structure, as produced by the proposed objective. It may also be that using only one reference translation limits the sensitivity of the automatic evaluation; further study with human evaluation is needed to assess this possibility. In addition, it may be useful to consider SParseval as one of a suite of evaluation tools in assessing ASR performance for purposes of translation, and to investigate the impact of SParseval optimization on other applications.

7. REFERENCES

[1] K. Kirchhoff, O. Rambow, N. Habash, and M. Diab, "Semi-automatic error analysis for large-scale statistical machine translation," in *Proceedings of the MT Summit XI*, 2007, pp. 289–296.

[2] E. Black *et al.,* "A procedure for quantitatively comparing syntactic coverage of English grammars," in *Proceedings 4th DARPA Speech & Natural Lang. Workshop*, 1991, pp. 306–311.

[3] S. Sekine and M. J. Collins, "The evalb software," http://cs.nyu.edu/cs/projects/proteus/evalb, 1997.

[4] B. Roark *et al.,* "Sparseval: Evaluation metrics for parsing speech," in *Proceedings of Language Resource and Evaluation Conference*, Genoa, Italy, 2006.

[5] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, 2006, pp. 223–231.

[6] "Global Autonomous Language Exploitation (GALE)," http://www.darpa.mil/ipto/programs/gale/.

[7] A. Stolcke, "SRILM—an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, vol. 2, pp. 901–904.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1998.

[9] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, pp. 373–400, 2000.

[10] J. G. Kahn, D. Hillard, M. Ostendorf, and W. McNeill, "Joint optimization of parsing and word recognition with automatic segmentation," Tech. Rep., University of Washington, EE Dept., 2007.

[11] M.Y. Hwang, G. Peng, W. Wang, A. Faria, A. Heidel, and M. Ostendorf, "Building a highly accurate Mandarin speech recognizer," in *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 490-495, 2007.

[12] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *Proceedings of HLT/NAACL*, 2007, pp. 404–411.

[13] Z. Huang, M. Harper, and W. Wang, "Mandarin part-of-speech tagging and discriminative reranking," in *Proceedings of EMNLP-CoNLL*, 2007, pp. 1093–1102.

[14] N. Xue and F. Xia, *The Bracketing Guidelines for the Penn Chinese Treebank*, 2000.

[15] A. Mauser and et al., "The RWTH statistical machine translation system for the IWSLT 2006 evaluation," in *IWSLT*, 2006, pp. 103–110.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *40th Annual Meeting of the Assoc. for Computational Linguistics*, 2002, pp. 311–318.

[17] E. Matusov, G. Leusch, and O. Bender ad H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *IWSLT*, 2005, pp. 148–154.