QUALITY ESTIMATION FOR ASR K-BEST LIST RESCORING IN SPOKEN LANGUAGE TRANSLATION

Raymond W. M. Ng, Kashif Shah, Wilker Aziz, Lucia Specia, Thomas Hain

Department of Computer Science, University of Sheffield, United Kingdom
{wm.ng,kashif.shah,w.aziz,l.specia,t.hain}@sheffield.ac.uk

ABSTRACT

Spoken language translation (SLT) combines automatic speech recognition (ASR) and machine translation (MT). During the decoding stage, the best hypothesis produced by the ASR system may not be the best input candidate to the MT system, but making use of multiple sub-optimal ASR results in SLT has been shown to be too complex computationally. This paper presents a method to rescore the k-best ASR output such as to improve translation quality. A translation quality estimation model is trained on a large number of features which aim to capture complementary information from both ASR and MT on translation difficulty and adequacy, as well as syntactic properties of the SLT inputs and outputs. Based on the predicted quality score, the ASR hypotheses are rescored before they are fed to the MT system. ASR confidence is found to be crucial in guiding the rescoring step. In an English-to-French speech-to-text translation task, the coupling of ASR and MT systems led to an increase of 0.5 BLEU points in translation quality.

Index Terms— Spoken language translation, Quality estimation, System integration

1. INTRODUCTION

Spoken language translation (SLT) combines automatic speech recognition (ASR) and machine translation (MT). State-of-the-art SLT systems adopt a pipeline approach where the 1-best ASR output (or analysis) is directly fed to the MT system [1]. ASR models are stochastic in nature, thus there is a degree of uncertainty inherently associated with every candidate analysis. This uncertainty (or ambiguity) is largely neglected in most SLT settings, leading to error propagation, and poor empirical results in utilising the sub-optimal ASR results beyond the 1-best solution [2, 3, 4, 5]. Propagating uncertainty into SLT models is challenging, both with respect to MT modelling and decoding.

In lattice and confusion network decoding, the input is generalised into an acyclic finite-state automaton [6]. This allows to propagate a rich set of options from the ASR into the MT system avoiding premature decisions. However, MT decoding complexity grows exponentially with the number of states of the input automaton (lattice or confusion network) requiring more aggressive pruning. For pruning to be effective, the analyses in the input automaton must be weighted adequately, otherwise decoding resources are wasted on poor hypotheses, ultimately degrading translation quality. Several techniques have been designed to deal with the peculiarities of converting (a meaningful part of) the ASR hypothesis spaces into word lattices or confusion networks for MT [7] . Another approach to tackle the problem, which we follow in this work, is to rely solely on the single best ASR decoding, but to introduce an independent modelling and rescoring step after ASR and before MT. The model learns to adjust the ASR decision based on predicted translation quality. This approach avoids the largely increased complexity of lattice decoding.

We incorporate a wide range of information from both the ASR and the MT hypothesis spaces into a translation quality estimation (QE) model [8, 9]. The QE model is a module independent from the core ASR and MT systems, which predicts the quality of the translation of competing ASR analyses. Based on the predicted quality score, the ASR hypotheses, in the form of k-best list, are rescored (or reranked) before they are fed to the MT system to produce the final translation. The ASR confidence level is found to be crucial in guiding the rescoring step. We report valuable BLEU improvements from 31.33 to 31.87 on an English-speech-to-French-text translation task. The results reported are based on very competitive ASR and MT systems operating on real-life data (video lectures).

2. RELATION TO PRIOR WORK

There have been extensive efforts towards a tighter integration in SLT systems. Coupling frameworks have been proposed to incorporate the scores from the ASR and MT [10]. Weighted finite-state transducers are popularly used [4, 11]. On the input side, k-best lists, confusion networks or lattices can be employed [12, 13, 14, 15] to keep alternative ASR hypotheses during decoding in the translation engine.

Few studies directly address the use of features that go beyond scores from ASR systems in SLT. Part-of-speech information was shown to be useful in deriving a confidence measure in a speech translation system [16]. In [17], the vector

Table 1. Summary of 117 features for the quality estimation system

Type	Description	#Feat	Type	Description	#Feat
ASR	Acoustic model & Language model score	6	LM♯	Source/Target sentence LM probability/perplexity	6
	Inverse document frequency	1	Punct♯	Counts and % of punctuations	7
	Binary features for the identity of k in k-best	10		Absolute difference in punctuations between source and target sentences	14
	Number of words and its normalised variants	4			
Count♯	Counts of tokens / brackets / quotation marks	8	POS♯	% of nouns / verbs / content words	12
	Average number of translations per source word as given by IBM 1 model	16	Glassbox♯	Global score of the MT system	1
				Model features	15
LM♯	1-3 gram counts and statistics in different frequency quartiles in source language	16	Pseudo ref♯	Evaluation of target language with pseudo reference from third-party translation system	1

♯: MT-based features

space representation of words was exploited for the extension and refinement of translation tables.

This study is similar to [18], where the k-best ASR outputs were reranked based on a small number of ASR features. However, ours is the first attempt to show complementarity by coupling both speech- and translation-related features, and in total more than 100 features of various types were used. In addition, our QE model was not trained towards a minimum ASR risk as in previous work, but directly towards a metric of translation quality. Finally, our experiments were conducted on a natural setting with TED talks covering different domains.

3. QUALITY ESTIMATION SYSTEM

The QE system used takes into consideration a wide range of features, which we summarise in Table 1. These include:

3.1. ASR-based features

21 features are extracted from the ASR system output. The decoder scores from the acoustic and language models, the difference of scores compared with their 1-best counterparts, and the normalised variants of the difference form 6 features. An inverse document frequency feature is generated from the in-domain training corpus with 734 documents (§5.1). These features are first extracted at phone/word levels, then averaged at sentence level. 10 binary features indicate the order of the ASR k-best ($k \leq 10$) using one-hot encoding. Another feature counts the number of words in the ASR no-case no-punctuation format. The count features is mean- and variance-normalised. Its difference with the 1-best count, and the normalisation of the difference are also extracted.

3.2. MT-based features

The MT-based features are extracted using the open source toolkit QUEST[1] [8, 9]. These features sum up to 96 and are based on source segments (difficulty of translation), target segments (translation fluency), comparison between the source and target segments (translation adequacy), and the MT system confidence. Among the first three categories, a large number of features are explored. In Table 1, they are grouped into 4 classes – count, language model (LM), punctuation, and part-of-speech (POS).

The count features include 8 features which describe the number of tokens, brackets and quotation marks in the source, the target sentences, and the ratio between the two. 16 features describe the average number of translations per word in the source language using an IBM 1 word alignment model with probabilities thresholded in different ways. The LM features include 6 features describing the LM probability and perplexity of source and target texts, and 16 features denoting the counts of n-grams in four frequency quartiles. The punctuation features comprise 7 features counting the occurrences of different punctuations (such as : ; ?) in source or target sentences, and 14 features describing the absolute difference of designated punctuation marks in source and target sentences. Numbers and accented characters (non a-z) are also taken into accounts. Using POS tags, 12 features describe the percentage of content words, nouns, verbs in source and target sentences.

In addition to the statistics- and comparison-based features above, 16 "glassbox" features from the MT system are extracted. They provide an indication of the confidence of the MT system in the translation it produces. They include the features of the MT linear model as well as the total score of the hypothesis as given by the decoder. Finally, a special feature – the **pseudo-reference** – is extracted by evaluating the translation by taking the output of a held-out MT system (in our case, Google Translate) as reference. The evaluation metric is the geometric mean of smoothed 1-4-gram precision score.

3.3. Algorithm

The quality estimation system learns the relationship between the features and the translation quality with support vector regression (SVR) machines [19]. Given a multi-dimensional features $\boldsymbol{x}$, a trained SVR computes the value of the predicted targets, which is given by:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - {\alpha_i}^*) K(\boldsymbol{x}_i, \boldsymbol{x}) \quad (1)$$

$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ... , $\boldsymbol{x}_N$ are the N support vectors. α_i and α_i^* are the Lagrangian multipliers in the primal problem. $K(\cdot, \cdot)$ is

[1] http://www.quest.dcs.shef.ac.uk/quest_files/features_{blackbox|glassbox}

the kernel function. The SVR looks for a training data subset as the support vectors, and infers optimal values of α_i and α_i^* which minimise the prediction error of the regression function. In this experiment, we tried different ASR and MT features (i.e. $\boldsymbol{x}$ with different dimensions). Sentence-based METEOR scores [20] were used as the learning target.

4. DATA

All results reported in this paper are based on TED talks [21]. TED organises and records English spoken short lectures by important figures of the public. These lectures are subtitled by professional human transcribers. In addition, translation into different languages is provided by a community of volunteers.

Apart from TED data (in-domain), acoustic data from a similar domain (i.e. lectures) was included in the training of the ASR acoustic model. Out-of-domain data was used in the training of language models and the translation systems. The use of these datasets is detailed in §5.

The IWSLT2010 development and evaluation sets were used for tuning the various parameters in the SLT system. These included grammar scale factor, insertion penalty, pruning beam-width in the ASR decoder; the linear model combining different MT components (e.g. translation, reordering and language models); and language model interpolation in ASR and MT. The IWSLT2011 evaluation set was used to train the quality estimation (QE) system. Empirical results on the IWSLT 2012 evaluation set are reported [22].

5. SYSTEM SETUP

5.1. ASR system

The ASR system was a multi-pass system. Different optimisation technologies were used. The first pass used deep neural networks (DNNs) in a tandem configuration [23]. The 26-dimensional bottleneck-layer features were concatenated with PLP features and derivatives, to a dimensions of 65, on which decision-tree-generated clustered tri-phone models were trained [24]. The second pass used VTLN, where the frequency warping factors were estimated from the first-pass output. The acoustic model was trained with the MPE criterion. CMLLR and MLLR transforms, both with 16 regression classes, were learnt in an unsupervised manner [25].

For acoustic modelling, training data comprised 734 TED talks published before 31 Dec 2010. Lecture data from the e-corner corpus and the LLC corpus were also included. After resegmentation and silence removal, the total duration of the TED, e-corner and LLC data were 132, 60 and 106 hours, respectively. For language modelling, a 4-gram interpolated LM with standard ARPA format was built on 5 corpora. TED with 3.17 million words served as the in-domain (ID) corpus. In addition, 4 out-of-domain (OOD) corpora were used: Europarl, commoncrawl, giga-word, news-commentary and UN-doc databases. All OOD data underwent data selection based on cross-entropy difference. The total number of words were 322.12 million. The size of vocabulary was $60,568$.

ASR decoding followed the multi-pass regime as mentioned above. In the final pass, a lattice was constructed for acoustic and language model rescoring, after which 10-best ASR hypotheses were extracted. The ASR system was tested on the IWSLT 2012 English data with $1,224$ sentences ($19,075$ words). For the 1-best ASR output, the word error rate was 14.3%. Detailed descriptions of the system can be found in [24].

5.2. MT system

A phrase-based model was trained in a standard setting using MOSES [26]. For phrase extraction all of the TED data (3.17M words) was used. Following previous findings [7], data selection via a cross-entropy criterion was used to select about 5% of the OOD data (30.58M words). The phrase length was limited to 5. Lexicalised reordering models were trained using the same data. For language modelling, we used the complete sets of OOD data (i.e. no data selection). 5-gram LMs were trained using LMPLZ [27]. 100-best MIRA tuning was employed [28]. Cube pruning [29] was performed in both tuning and testing.[2] To restore the correct case of the output we employed the truecasing heuristic.

In SLT, the input to the MT system was ASR output, which typically lack casing and punctuation. Following previous work [7], a monolingual translation model was trained to recover casing and punctuation from the ASR output, thus producing source sentences which are more adequate for translation. Our baseline system achieved 40.91 BLEU on gold-standard transcripts (i.e. without ASR).

5.3. QE and re-reanking system

QE features were extracted from the 10-best ASR outputs (i.e., English source text with punctuation and casing recovered via monolingual translation), and the 10 corresponding 1-best translations from the MT system (i.e., French target text). As previously mentioned, the learning target was the sentence-based METEOR score computed for each machine translation against a human translation. QE training was conducted on IWSLT2011 evaluation set, which covers 818 sentences, with a total of $8,180$ training samples (818×10). For the SVR, we experimented with linear and radial basis function (RBF) kernels. Parameters in the RBF kernel were tuned via cross-validation. We only report results for linear SVR, as the RBF kernel did not lead to better results.

Tests were performed on the IWSLT2012 eval set ($1,124$ sentences). A translation quality METEOR score was predicted for each of the 10-best ASR outputs (and its 1-best translation). Based on the predicted quality, the 10-best ASR

[2] Decoding was done with the minimum Bayes risk criterion and reordering over punctuations was forbidden.

Table 2. Performance on individual k-best ASR & translation

k	WER	BLEU	k	WER	BLEU
1	14.3%	31.33	6	18.6%	28.40
2	16.3%	30.05	7	18.7%	28.64
3	17.3%	29.28	8	18.9%	28.82
4	17.8%	29.13	9	18.9%	29.11
5	17.9%	28.96	10	18.9%	29.00

hypotheses were reranked and the translation performance of the entire test set was evaluated using BLEU.

6. RESULTS

6.1. Baseline and oracle performance

Table 2 gives the baseline performance with the 10-best ASR output on IWSLT 2012 evaluation data set in terms of WER (for ASR) and BLEU (for MT). For ASR, there is a 4.6% absolute WER increase moving from 1^{st}-best to 10^{th}-best output. The corresponding drop in BLEU is 2.3. The degradation of performance is small between the 7^{th}- and 10^{th}-best.

The oracle WER and BLEU are also computed by reranking ASR analyses according to the true sentence-based METEOR score of their corresponding translations. The WER drops significantly to 13.7% in oracle 2-best reranking, then slowly decreases along k, with 13.4% in oracle 10-best. Oracle BLEU consistently increases with the size of k, where oracle reranking 10-best leads to a BLEU score of 35.52. The 2^{nd}- to 10^{th}-best ASR output gives moderately worse results than 1^{st}-best, but the quality is not unacceptably poor and there is a huge potential of translation improvements.

6.2. Quality estimation with various feature sets

Table 3 shows the reranked results based on different QE feature sets. It can be seen that both ASR- and MT-based features, when used alone, bring a performance drop. However, the combined features yield 31.51 BLEU. This is strong evidence for complementarity between the ASR- and the MT-based features. The 21 ASR-based features were augmented incrementally to form 6 different feature sets. The first set (with 45 features) did not give a robust enough quality estimate for translation improvement. More features in the QE system led to considerably better translation quality. Augmenting the MT-based features in different orders may give different results. This could be further studied in future.

Based on the result analysis, it was found that rescoring (or reranking) was generally more effective when the 1-best sentence had low ASR confidence. Therefore, a follow-up experiment was conducted. ASR confidence was computed by averaging the word confidence in every ASR 1-best hypothesis. A confidence threshold was set, and QE-informed rescoring was performed only if the confidence in 1-best ASR was below threshold. As such, a hybrid data set, which comprises the rescored 1-best and the original 1-best sentences, was created. Figure 1 shows that the translation performance on this hybrid data set varies with different proportions of

Table 3. BLEU with different QE-informed settings

Features (#Feat.)	BLEU
No rerank (0)	31.33
MT-based features (95)	30.90
ASR-based features (21)	31.05
+ MT count and word-word alignment (45)	31.24
+ MT LM (67)	31.42
+ MT punctuation (88)	31.43
+ MT POS (100)	31.44
+ MT glassbox (116)	31.47
+ pseudo reference (117)	31.51
+ 117 feat. + ASR confidence-informed partial rescoring	31.87

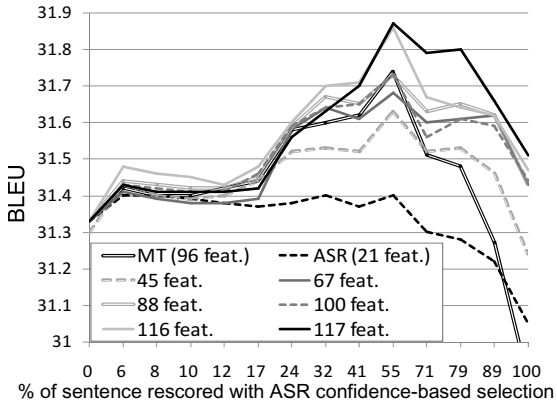

Fig. 1. BLEU against different settings of partial rescoring

rescored sentences, as a result of changing the ASR confidence threshold. Across different feature set sizes, the same optimal threshold point was found when 55% of the sentences were rescored. With this threshold, the addition of features contributed to a monotonic increase of BLEU. The best BLEU score achieved is 31.87, which is 0.54 above the baseline without reranking. The WER at the corresponding point is 15.0%.

Despite the observed BLEU increase, the type of translation errors being corrected through the re-ranking mechanism is not clearly known. Previous work on error analysis tried to tackle this question [30]. It is an interesting topic to investivate in future studies.

7. CONCLUSION

This paper presented a rescoring strategy based on a quality estimation method considering a large number of features from the ASR and MT outputs and systems. The two types of features were shown to be complementary. Cumulative improvements in translation performance were observed with the addition of more features. Reranking the ASR hypotheses where the 1-best ASR confidence is low led to further improvements. The quality estimation system is flexible in terms of feature combination, learning target, and machine learning algorithms. Therefore, it could be extended in different ways to further improve the final translation performance.

8. REFERENCES

[1] M. Cettolo, J. Neihues, S. Stüker, L. Bentivogli, and M. Federcio, "Report on the 10th IWSLT evaluation campaign," in *Proc. IWSLT*, 2012, pp. 15–32.

[2] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Proc. ICASSP*. IEEE, May 2011.

[3] K. Sudoh, T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "NTT statistical machine translation system for IWSLT 2008," in *Proc. IWSLT*, 2008, pp. 92–97.

[4] A. Pérez, M. I. Torres, and F. Casascuberta, "Potential scope of a fully-integrated architecture for speech translation," in *Proc. EAMT*, 2010.

[5] P. R. Dixon, A. Finch, C. Hori, and H. Kashioka, "Investigation on the effects of ASR tuning on speech translation performance," in *Proc. IWSLT*, 2011, pp. 167–174.

[6] C. Dyer, S. Muresan, and P. Resnik, "Generalizing word lattice translation," in *Proceedings of ACL-08: HLT*, June 2008, pp. 1012–1020.

[7] A. Birch, N. Durrani, and P. Koehn, "Edinburgh SLT and MT system description for the IWSLT 2013 evaluation," in *Proc. IWSLT*, 2013.

[8] L. Specia, K. Shah, J. G. C. d. Souza, and T. Cohn, "QuEst - A translation quality estimation framework," in *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics: Demo Session*, 2013, p. 794.

[9] K. Shah, E. Avramidis, E. Biçici, and L. Specia, "Quest - design, implementation and extensions of a framework for machine translation quality estimation," *Prague Bull. Math. Linguistics*, vol. 100, pp. 19–30, 2013.

[10] H. Ney, "Speech translation: coupling of recognition and translation," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, 1999, pp. 517–520 vol.1.

[11] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. Interspeech*, 2005, pp. 3177–3180.

[12] N. Bertoldi, R. Zens, M. Federico, and W. Shen, "Efficient speech translation through confusion network decoding," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1696–1705, 2008.

[13] E. Matusov and H. Ney, "Lattice-based asr-mt interface for speech translation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 721–732, 2011.

[14] G. A. Saon and M. A. Picheny, "Lattice-based Viterbi decoding techniques for speech translation," in *Proc. ASRU*, 2007, pp. 386–389.

[15] V. H. Quan, M. Federico, and M. Cettolo, "Integrated N-best re-ranking for spoken language translation," in *Proc. Eurospeech*, 2005.

[16] N. F. Ayan, A. Madnal, M. Frandsen, J. Zheng, P. Blasco, A. Kathol, F. Becchet, B. Favre, A. Martin, T. Kwiatkowski, M. Ostendorf, L. Zettlemoyer, P. Salletmayr, J. Hirschberg, and S. Stoyanchev, "Can you give me another word for "Hyperbaric?": Improving speech translation using targeted clarification questions," in *Proc. ICASSP*, 2013, pp. 8392–8395.

[17] T. Mikolov, Q. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," in *Tech. report, arXiv*, 2013.

[18] C.-H. Li, N. Duan, Y. Zhao, S. Liu, and L. Cui, "The MSRA machine translation system for IWSLT 2010," in *Proc. IWSLT*, 2010, pp. 135–138.

[19] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," 1998.

[20] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of WMT14*, 2014.

[21] TED, "Technology entertainment design," http://www.ted.com, 2006.

[22] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. IWSLT*, 2012, pp. 12–33.

[23] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000.

[24] R. W. N. Ng, M. Doulaty, R. Doddipatla, O. Saz, M. Hasan, T. Hain, W. Aziz, K. Shaf, and L. Specia, "The USFD spoken language translation system for IWSLT 2014," *Proc. IWSLT*, pp. 86–91, 2014.

[25] R. Doddipatla, M. Hasan, and T. Hain, "Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition," 2014.

[26] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007, pp. 177–180.

[27] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, August 2013, pp. 690–696.

[28] C. Cherry and G. Foster, "Batch tuning strategies for statistical machine translation," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 427–436.

[29] L. Huang and D. Chiang, "Forest rescoring: Faster decoding with integrated language models," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 144–151.

[30] M. Popović and H. Ney, "Towards automatic error analysis of machine translation output," *Comput. Linguist.*, vol. 37, no. 4, pp. 657–688, Dec. 2011.