A study on the stability and effectiveness of features in quality estimation for spoken language translation

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

Department of Computer Science, University of Sheffield, United Kingdom

{wm.ng,kashif.shah,l.specia,t.hain}@sheffield.ac.uk

Abstract

A quality estimation (QE) approach informed with machine translation (MT) and speech recognition (ASR) features has recently shown to improve the performance of a spoken language translation (SLT) system in an in-domain scenario. When domain mismatch is progressively introduced in the MT and ASR systems, the SLT system's performance naturally degrades. The use of QE to improve SLT performance has not been studied in this context. In this paper we investigate the effectiveness of QE under this setting. Our experiments showed that across moderate levels of domain mismatches, QE led to consistent translation improvements of around 0.4 in BLEU score. The QE system relies on 116 features derived from the ASR and MT system input and output. Feature analysis was conducted to understand the information sources contributing the most to performance improvements. LDA dimension reduction was used to summarise effective features into sets as small as 3 without affecting the SLT performance. By inspecting the principal components, eight features including the acoustic model scores and count-based word statistics on the bilingual text were found to be critically important, leading to a further boost of around 0.1 BLEU score over the full set of features. These findings provide interesting possibilities for further work by incorporating the effective QE features in SLT system training or decoding.

Index Terms: Spoken language translation, quality estimation, system robustness

1. Introduction

Quality estimation (QE) is a popular research topic in machine translation (MT). The idea behind QE is to make use of the input, output and optionally internal scores from the MT systems to build a link between these information and the translation quality. One possible application is to use this to guide the inference process in MT towards optimal performance in translation. In spoken language translation (SLT), most systems adopt a pipelined approach whereby the automatic speech recognition (ASR) and MT components are trained independently. It has already been suggested, however, that this type of approach where ASR systems are trained independently of their final application is suboptimal in the context of speech translation [1]. Analogously, the component MT system is trained on human texts, which are significantly different from the actual runtime input (ASR hypotheses). Given these mismatches, additional information from QE models can be very useful to improve an SLT system. In our recent study, QE was used to predict the translation quality of the k-best hypotheses from the ASR system output. Based on these predictions, ASR k-best list rescoring was performed before the actual machine translation. This way of considering additional information in SLT proved to be efficient in a benchmark SLT task for IWSLT [2].

Given the positive results in the benchmark SLT task, two natural questions arise. First, the stability of the system performance based on QE information. It is unknown how QE performance changes under various sources of variability in the ASR models, MT models and also the different SLT inputs. Second, given the use of over 100 features in QE, it would be beneficial to gain a deeper insight into the relative contribution of different features. To answer the first question, we conduct a stability study by replicating the QE experiments and progressively introducing moderate domain mismatch in the ASR and MT systems. For the second question, linear discriminant analysis (LDA) was conducted on the features.

This work represents a substantial extension of our recent work [2]. Here we tested and demonstrated the stability of the QE approach in SLT by using three domain mismatch scenarios. Another contribution of this work is from feature factorisation and analysis. A straightforward linear transformation technique served to reduce the feature dimensionality, as well as to reveal eight important features which gave rise to extra gains over the full feature set. In the following, §2 introduces the technical details of the QE and LDA approached used in this study study. These are followed by a description of the data and setup in §3, 4. The main results are in §5, and conclusions in §6.

2. Quality estimation for SLT

2.1. Features for QE

The QE system takes into consideration a wide range of features. A total of 116 features were used in this paper. The features were almost identical to the feature inventory used in [2]. The only change was to exclude the pseudo reference features. This was due to the expensive computation for the feature, and the very small performance gain it gave.

The 116 features are summarised in Table 1. They can be classified into three big classes. 21 features were extracted from the ASR system output. These features describe the decoder scores from the acoustic and the language models, the ASR k-best rank information and other count statistics. 79 are translation "blackbox" features. They were extracted based on source segments (difficulty of translation), target segments (translation fluency), and the comparison between the source and target segments (translation adequacy). 16 features are MT system-dependent, the so called "glassbox" features. They describe the confidence of the MT system, such as the global model score. The blackbox and glassbox features were extracted using the open source toolkit QUEST (`http://www.quest.dcs.shef.ac.uk`). More detailed descriptions on the features can be found in [2, 3, 4].

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology) and Google.

Table 1: Summary of 116 features for the quality estimation system

Type	Description	#Feat
ASR	Acoustic model & Language model score	6
	Inverse document frequency	1
	Binary features for the identity of k in k-best	10
	Number of words and its normalised variants	4
Blackbox$^\sharp$	Counts of tokens / brackets / quotation marks	8
	Average number of translations per source word as given by IBM 1 model	16
	Source/Target sentence LM probability/perplexity	6
Blackbox (con'd)$^\sharp$	1-3 gram LM counts and statistics in different frequency quartiles in source language	16
	Counts and % of punctuations	7
	Absolute difference in punctuations between source and target sentences	14
	% of nouns / verbs / content words	12
Glassbox$^\sharp$	Global score of the MT system	1
	Model features	15

$^\sharp$: MT-based features

2.2. Dimension reduction of features

The nature of the 116 features suggests that they may be highly redundant, which may introduce noise affecting QE model learning. Previous studies suggested the usefulness of feature selection techniques to select a subset of features for QE in MT. For instance, [5, 6] employed Gaussian Process to discriminatively rank the features and showed that a model built based on the top selected $10 - 25$ features outperforms a model trained on the full feature set.

In SLT, linear modelling of QE features was shown to give the best results [2]. For this reason, linear discriminant analysis (LDA) was carried out. Let $[\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N-1}] \forall \boldsymbol{x}_n \in \mathbb{R}^D$ represents a feature set with N samples and D dimesnions, $[y_0, y_1, \cdots, y_{N-1}], \forall y_n \in [0, 1, \cdots, C - 1]$ represents the class labels. LDA aims to find a projection $\mathbf{A}$ to apply on x_n that maximises the Fisher criterion, which is defined as the ratio of between-class covariance to within-class covariance,

$$\mathbf{A} = \arg\max_{\mathbf{A}} \frac{|\mathbf{A}\mathbf{S}_b\mathbf{A}^T|}{|\mathbf{A}\mathbf{S}_w\mathbf{A}^T|}, \quad (1)$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the covariance matrices of x_n in the original space. In this study, the number of samples N and the dimensions of features D are 8133 and 116, respectively. Following the conventional LDA methodology one would easily run into the problem of degeneracy, i.e. having covariance matrices which are non-invertible. To avoid this problem, an approach with principal-component-analysis (PCA) followed by LDA is used for dimensionality reduction of features [7]. In the first stage, PCA is used to project the features from the original space to a class space, where the dimensionality is reduced and $\mathbf{S}_w$ no longer degenerates. In the second stage, LDA transformation is used to produce the most discriminating feature set.

QE in SLT has been traditionally addressed as a regression problem, where a continuous target value (in this case, a METEOR quality score [8]) is predicted from $\boldsymbol{x}_n$. In order to use LDA, a scalar quantisation step was conducted to discretise the target variable into C classes. In this experiment, we experimented with varying C from 3 up to 10.

Table 2: *Data involved in SLT training/testing*

Model	Training data set (duration / # words)
ASR_1-AM	TED(132h), LLC(106h), ECRN(60h)
ASR_2-AM	TED(112h), AMI+AMIDA+ICSI(165h), ECRN(60h).
$\text{ASR}_{1/2}$-LM	TED(3.17M)$^\flat$, News commentary(4.0M), Commoncrawl(70.7M), Gigaword(575.7M), Europarl (50.3M).
ASR-LM-dev	Dev(17K) & Test(27K) sets from IWSLT 2010
MT-TM	TED (3.17M)$^\#$, News commentary(0.7M), Commoncrawl(10.8M), Gigaword(14.9M), Europarl (1.9M).
MT-LM	TED (3.25M)$^\#$, News commentary(4.7M), Commoncrawl(76.7M), Gigaword(145.6M), Europarl(52.5M), News-test-2013(187.2M), UNdoc(90.4M)
MT-tune	Dev(17K) & Test(27K) sets from IWSLT 2010

$^\flat$ # In-domain data (TED) removed in training ASR_2-LM / mimatched MT models

2.3. Stability of quality estimation in SLT

The stability of a QE system is defined as its capability to maintain a fairly constant performance across different scenarios. The QE system depends on a number of features. The concept of stability is thus essentially connected to whether or not the features used are reliable across scenarios. This question was addressed by a variability study, where the experiments were replicated in different settings. In particular, it is known that SLT's performance can degrade with training and test domain mismatches. Therefore we investigate the interesting problem of whether QE can retain a stable performance in those cases.

3. Data

Table 2 shows the data used for ASR and MT system training, which mostly followed the settings in our IWSLT 2014 submission [9]. Two ASR models, ASR_1 and ASR_2 were trained. For the acoustic models, both ASR_1 and ASR_2 were trained on TED data (i.e. in-domain data). The AM data for ASR_1 was augmented by the lecture archives from the liberated learning consortium (LLC) and the Stanford University's entrepreneurship corner (ECRN) [10, 11]. In ASR_2, besides ECRN 165 hours of meeting data from the AMI, AMIDA and ICSI corpora were added so the ASR model would also reflect generic domains other than only lectures [12]. ASR language models were trained on the same dataset except that TED data was removed for ASR_2.

The text data for language and translation models training were mostly taken from WMT14 [13], supplemented with the official in-domain TED data in IWSLT evaluations [14]. Data other than TED talks came from news commentaries and parliamentary minutes and they are considered to be out-of-domain.

Language model adaptations and MT system tuning were performed on the IWSLT 2010 development and test data (44K words). These data were the held-out set from the TED training data mentioned above.

The QE system was trained on features extracted from SLT system input and output. In the training phase, SLT was run on IWSLT 2011 test data. It comprises 818 segments with 1.1 hours of length in English speech and $13K$ words in French text. The QE system was tested on IWSLT 2012 test data, with 1124 sentences (1.8 hours in English speech, 20K words in French text).

4. Experimental setup

4.1. ASR and MT

The SLT task reported in this paper is an English-speech-to-French-text translation task on TED talks data [15]. The SLT system comprises an English ASR and an English-to-French MT system, which follows the setup presented in [9].

The ASR system was a multi-pass system comprising DNN acoustic models with tandem configurations, VTLN wrapped features, MPE trained HMM models with CMLLR and MLLR transformation and 4-gram language model rescoring. Two variants of the ASR system were considered: ASR_1 and ASR_2. ASR_1 resembles an in-domain system setting and ASR_2 was built to introduce domain mismatches. Compared with our previously reported system, cross CMLLR and MLLR adaptations using ASR_1 hypotheses to adapt ASR_2 (and vice versa) were substituted by self adaptations. This is to make a clear distinction between the matched- and mismatched-domain ASR runs. On the IWSLT 2012 test data, the ASR_1 and ASR_2 variants reached WERs of 14.3% and 16.4% respectively, i.e. a 15% relative increase of errors when out of domain data became dominant in the training.

Two variants of phrase-based MT system were trained on a standard setting [16]. In-domain MT models were trained and tuned on the data listed in Table 2, i.e. human written text in the source language, taking no account of issues such as punctuation deletions, case and word errors in the ASR outputs. A monolingual translation model was used to recover casing and punctuation from the ASR output, producing source sentences which are more adequate for translation. Domain mismatched MT models were built by removing the TED data from all training components (those components marked $^{\#}$ in Table 2).

In our experiments, all translation results were scored on true-cased output, which gives around 0.55 BLEU score increase on the results we had previously reported.

4.2. QE system setup

The QE-informed ASR k-best list rescoring described in [2] was conducted. In brief, the ASR and MT systems were run on the QE training and test data (§3). The top 10 ASR and their 1-best MT results were generated. 116 sentence-based features were then derived from the ASR and MT system input / output and system internal scores. METEOR score was computed on every decoded sentence in the QE training set (IWSLT 2011 test) based on human reference translations as the learning target. The QE models were generated to learn the relationship between the features and the target using support vector regression (SVR) machines [17]. Afterwards, METEOR scores on the translations for the 10-best ASR outputs in the QE test set (IWSLT 2012 test) were predicted, based on which this ASR 10-best list was rescored, before the corresponding translation hypothesis was decided. An ASR confidence-informed heuristic was applied such that rescoring was only conducted on sentences with lower ASR confidence [2]. This heuristic was applied in the experiments reported throughout this paper.

4.3. QE tests in various domain mismatch scenarios

To investigate the robustness of QE in different domain mismatch scenarios, three SLT system settings were tested as illustrated in Table 3. Setting A is the best possible scenario, where both ASR and MT were trained on in-domain data. Domain mismatch was introduced in Setting B by excluding TED data from the MT training data. In Setting C we tested the extreme case where in-domain TED data was excluded from MT and a general-domain ASR model was used.

QE experiments were replicated in these three settings. The mismatch in ASR led to a relative increase of 15% in WER. In MT, the mismatch brought about 1.4 BLEU score reduction.

Table 3: *Different system training data in three scenarios*

Setting A:	ASR_1	Contains TED
Setting B:	ASR_1	Excludes TED
Setting C:	ASR_2	Excludes TED

4.4. Factorisation of features

Linear transformation of features using LDA was carried out (§2.2). With the features in the projected space, the steps of SVR learning, METEOR score prediction and ASR 10-best list rescoring described in §4.2 were replicated to generate corresponding SLT results. In the experiments reported in this paper, the size of the projected dimensions was tied to the number of target classes. We also tried other combinations where the number of target classes was higher than the projected dimensions. However, this did not lead to better results.

5. Results

5.1. Behaviour of QE under domain mismatch

The SLT systems in the three scenarios were applied on the QE training and test data. Table 4 summarises the differences in the IWSLT 2012 test set generated for each scenario (i.e., different English source segments and different translations). The translations of the 10-best ASR output were concatenated into a long list and an edit distance metric (TER [18]) was computed by pairwise comparisons of the three sets. On the English side, Settings A and B share the same ASR input so the TER is zero. The highest TER is observed between Settings A and C in French (41.7%), which clearly reflects the changes caused by different training data in both ASR and MT systems. The same level of TER is observed for the QE training data (IWSLT 2011 test). These numbers suggest that Settings A, B and C are generating significantly different input for QE (especially on the target language side). Therefore, any observed trends in the QE results across the three settings can be interpreted as consistent results across very different data.

Table 4: *Difference between IWSLT 2012 test output with different systems (Settings A, B, C) in terms of TER*

		Setting B	Setting C	French
		20.8%	41.7%	Setting A
Setting B	0.0%		33.3%	Setting B
Setting C	21.3%	21.3%		
English	Setting A	Setting B		

Table 5 shows the BLEU scores with ASR 10-best list rescoring on the IWSLT 2012 test data using the QE model learnt on IWSLT 2011 data. The three columns represent different domain mismatch levels (Settings A, B and C). The top row shows the baseline results, where the first-best ASR was translated as is. Rescoring with the full set of 116 features is reported in the bottom row. Six different groupings of features were made according the feature types in Table 1 and their results are also shown. It was found that across Settings A, B and C, the BLEU improvements with all features are stable (0.41, 0.44 and 0.53). The absolute BLEU improvement is smaller compared with the 0.54 reported in [2]. This is because the translation model and target language model used in this study were trained on 2.3M and 3.9M fewer words respectively. Post-experimental studies showed that the new models led to small differences ($<$ 0.1 BLEU) for the first-, second- and third-best ASR output, but a reduction of up to 0.3 BLEU points for the ninth- and tenth-best ASR output. This limited the improvement of QE rescoring overall.

We have also run extra experiments to introduce variability in SLT performance by replacing the 5-gram target language models with 2-gram ones. This significantly reduced the baseline BLEU score by 3-4 points (A: 29.05, B: 26.68, C: 25.49).

By running QE on these data sets, the corresponding BLEU score increase is 0.21, 0.32 and 0.24 respectively. The expressive power of both blackbox and glassbox features was weakened by the use of a weaker target language model, thereby also affecting the overall results of QE.

Table 5: *Translation results (BLEU) by ASR 10-best list rescoring with different features in different scenarios*

	Setting		
Features (# features)	A	B	C
Baseline (0)	32.03	30.64	29.41
ASR (21)	31.97	30.48	29.56
Glassbox (16)	32.41	31.06	29.93
Blackbox (79)	32.32	30.80	29.95
ASR + Glassbox (37)	32.45	30.96	30.02
Blackbox + Glassbox (95)	32.51	30.99	30.04
ASR + Blackbox (100)	32.37	31.03	30.04
ASR + Blackbox + Glassbox (116)	32.44	31.08	29.94

Test data: IWSLT 2012 test

5.2. Dimensionality reduction of features

Table 5 shows that, in different scenarios, different subsets of QE features for rescoring lead to better results than the full set of 116 features. For Set A, the best result was attained with blackbox and glassbox features. The full feature sets worked best for Set B. For Set C, either the ASR or the glassbox features had to be excluded. These results are included in the second row in Table 6. Only limited groupings were explored, but it was infeasible to try out all possible groupings given the large number of combinations. LDA essentially acted as a soft selection tool of features by optimally combining them. Table 6 shows the performance with LDA. In Settings A and B, LDA projections down to 3 dimensions led to an even higher increase in BLEU than using the full set (0.5 and 0.48 over the baseline). For Setting C, the highest increase in BLEU was obtained with 10 dimensional features (0.67 over the baseline).

Table 6: *QE results with LDA dimensionality reduction*

	Setting		
Features	A	B	C
Full set	32.44	31.08	29.94
Best grouping (from Table 5)	32.51	31.08	30.04
LDA with projected space dimensions:			
3	32.53	31.12	29.85
4	32.48	31.02	30.05
5	32.43	31.01	30.00
6	32.47	31.05	29.97
7	32.48	30.06	30.01
8	32.36	30.99	30.00
9	32.42	30.97	30.03
10	32.49	31.03	30.08

The extra 0.1 BLEU score gain over the use of the full feature set is an interesting result. We further inspected the LDA transformation matrices. It was discovered that in all settings, the weights were fairly equally distributed among features. Nevertheless, eight features in particular had a weight magnitude at least ten times higher than the average of the others. These eight features are listed in Table 7.

To further analyse the contributions of these eight features, a control experiment was performed. We removed the eight features beforehand, and followed exactly the same LDA-QE pipeline. The resulting BLEU scores were 32.47, 31.03 and 29.96 for Settings A, B and C, respectively, i.e. very similar to the full feature set results. Also, without those features, the projection weights no longer showed a skewed distribution.

Table 7: *List of critically important features*

Type	Feature description (Dimension)
ASR	Normalised sentence-based score from the acoustic model (1), Difference of the above feature to its 1-best counterpart (1)
Blackbox	Average number of translations per source word in the source sentence (4), The above feature weighted by the source words frequency (2)

The LDA experiment showed that the 116 features optimally combined in a simple manner to give 0.4 BLEU improvement. Emphasising the eight features would give a further boost of 0.1 BLEU score. The feature of acoustic model scores is a collapsed scalar metric which averages out all the component phoneme statistics in the sentence. Its design for QE may be reconsidered so they can be fully exploited. The 6 remaining blackbox features are known to perform well for QE in general. They reflect the ambiguity of the source words in terms of the number of possible translations they have. It is known that the more ambiguous a word is, the higher the chances it can be incorrectly translated.

Table 8: *Best LDA results in relation to 8 important features*

Include 8 critically	Settings		
important features	A	B	C
Yes	32.53	31.12	30.08
No	32.47	31.03	29.96

Finally, we present some preliminary observations on the transfer capabilities of the features (and the LDA projected features) across different scenarios. We found that under certain criteria, features from different systems (i.e. settings) may be shared. More empirical analysis is needed to check this hypothesis. However, this finding bares significant implications. Although a QE system is not expensive to train, the generation of the QE training data involves the use of ASR and MT systems. Most of the data was used in training the systems, leaving only a small amount of held-out data for QE training. The ability to transfer the QE model from one scenario to another would thus greatly improve the efficiency of model training.

6. Conclusions

In this paper we demonstrated the stability of QE system and proposed an effective linear factorisation for the QE features. A constant BLEU score improvement of 0.4 to 0.5 was achieved across different scenarios where domain mismatches are progressively introduced in the training data of ASR and MT systems. LDA was used to project the over 100 QE features into very small dimensions without reducing the QE performance. Moreover, 8 features were found to be critically important and gave a further 0.1 boost in BLEU score across various scenarios. These findings provide promising directions for further improvements of SLT by incorporating effective QE features into SLT system training or decoding.

7. Data Access Statement

Data used in this paper was obtained from these sources: ICSI Meetings corpus (LDC# LDC2004S02), AMI corpus (DOI# 10.1007/11677482_3), TedTalks, E-corner and MT training data (harvested from www.ted.com, ecorner.stanford.edu, and www.statmt.zorg/wmt14). Specific file lists used in the experiments, as well as result files can be downloaded from http://mini.dcs.shef.ac.uk/publications/papers/is15-ng.

8. References

[1] 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.

[2] R. W. M. Ng, K. Shah, W. Aziz, L. Specia, and T. Hain, "Quality estimation for ASR K-best list rescoring in spoken language translation," in *Proc. of ICASSP*, 2015.

[3] 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.

[4] 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.

[5] K. Shah, T. Cohn, and L. Specia, "An investigation on the effectiveness of features for translation quality estimation," *Proceedings of MT Summit XIV, Nice, France*, 2013.

[6] ——, "A bayesian non-linear method for feature selection in machine translation quality estimation," *Machine Translation*, pp. 1–25, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10590-014-9164-x

[7] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[8] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems," in *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 2011.

[9] 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.

[10] LLC, "Liberated learning consortium," http://liberatedlearning.com.

[11] ECRN, "Stanford university's entrepreneurship corner," http://ecorner.stanford.edu.

[12] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincoln, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *Proc. MLMI*. Springer, 2007.

[13] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 workshop on statistical machine translation," in *Proceedings of WMT14*, 2014, pp. 12–58.

[14] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

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

[16] 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.

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

[18] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, 2006, pp. 223–231.