Proceedings Paper:
Ng, R.W.M., Shah, K., Specia, L. et al. (2016) Groupwise learning for ASR k-best list reranking in spoken language translation. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 20-25 Mar 2016, Shanghai. http://dx.doi.org/10.1109/ICASSP.2016.7472853, 2016-M. , pp. 6120-6124. ISBN: 9781479999880. ISSN: 1520-6149.

https://doi.org/10.1109/ICASSP.2016.7472853

GROUPWISE LEARNING FOR ASR K-BEST LIST RERANKING IN 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

This paper studies the enhancement of spoken language translation (SLT) with groupwise learning. Groupwise features were constructed by grouping pairs, triplets or M-plets of the ASR k-best outputs. Regression and classification models were learnt and a straightforward score combination strategy was used to capture the ranking relationship. Groupwise learning with pairwise regression models give the biggest gain over simple support vector regression models. Groupwise learning is robust to sentences with different ASR-confidence, meaning that the confidence threshold heuristics in reranking are no longer needed. This technique is also complementary to linear discriminant analysis feature projection. Altogether a BLEU score improvement of 0.80 was achieved in an in-domain English-to-French SLT task.

***Index Terms*—** groupwise learning, ordinal regression, spoken language translation, support vector regression

1. INTRODUCTION

Spoken language translation (SLT) involves automatic speech recognition (ASR) and machine translation (MT) systems trained on different data and objective criteria. There have been extensive efforts in SLT system enhancements. Format and character conversion minimise the model mismatch between ASR and MT model trained in independent environment [1]. Incorporating ASR transcript or its simulation in MT system training also reduces system mismatch [1, 2]. With the goal of tighter system integration, coupling frameworks have been proposed to incorporate the scores from the ASR and MT [3]. Weighted finite-state transducers are popularly used [4, 5].

ASR and MT systems are usually large and complex. Considerable efforts are necessary to adapt or integrate system components. Without readapting the models, we could re-prioritise the search result during the decode stage. k-best lists, confusion networks or lattices can be employed [6, 7, 8, 9] to keep alternative ASR hypotheses during decoding in the translation engine.

This research is partially funded by Google.

Distinctive features derived from ASR and MT could be used to inform an optimal SLT results [9, 10, 11]. In our previous study, a quality estimation model was used to predict the translation performance of a sentence based on a comprehensive set of features. According to the predicted quality, reranking was performed on a 10-best ASR subject to optimal SLT performance [12].

In the above work, a global model is learnt to generate a score for one single hypothesis at a time. In this study we look at SLT enhancement as a groupwise learning problems, where pairs (or groups) of the ASR k-best outputs are compared. We show that groupwise learning plus a straightforward score combination strategy effectively capture the ranking relationship better than the quality prediction model for single hypothesis, and outperforms the latter in SLT tasks.

2. FEATURES FOR QUALITY ESTIMATION

In the SLT quality estimation problem, a D-dimensional feature vector $\boldsymbol{x}_t$ is extracted for every sentence t to represent its property. In our experiments, the feature vector contains 116 features and 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`). The list of features and the way they are extracted were identical as described in [12]. More details could be found in [12, 13, 14].

3. MODEL CONSTRUCTION

We define a quality estimation (QE) problem where the SLT performance metric y_t of a sentence t is predicted based on

the D-dimensional feature vector $\boldsymbol{x}_t$. y is METEOR score [15], which is an automatic translation quality metric with continuous range. A regression model was used in prediction. In this study, it was realised by support vector regression [16],

$$\hat{y}_t = f(\boldsymbol{x}_t) = \sum_{i=1}^{N} (\alpha_i - \alpha_i{}^*) Ker(\boldsymbol{x}_i, \boldsymbol{x}_t), \qquad (1)$$

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ... , $\boldsymbol{x}_N$ are the N support vectors from the training data collection. α_i and α_i^* are the Lagrangian multipliers in the primal problem. $Ker(\cdot, \cdot)$ is the kernel function.

Assume a set of feature vectors $\boldsymbol{x}_{(t,1)}$, ... , $\boldsymbol{x}_{(t,k)}$, ... , $\boldsymbol{x}_{(t,K)}$ which represents the ASR K-best candidates of a particular test sentence t. The rank $\hat{y}_{(t,k)}$ among k is more important than their absolute values. The same problem was studied in handwriting recognition [17], face detection [18] and in biology for protein sequence detection [19]. The main idea is to focus on the relative comparison within groups of two or more samples in the training data collection and learn a distance metric for each of these groups. With this intuition, a pairwise and a M-plet feature construction are proposed in a regression and a classification setup respectively to contrast the vector-based model in Eq.(1). The control setting, which considers only one single hypothesis at a time, is hereinafter known as the *single SVR* model.

3.1. Pairwise regression

In pairwise regression, ordered pairs of features are constructed by concatenating different K-best candidates of the same sentence t to form $(\boldsymbol{x}_{(t,k)}, \boldsymbol{x}_{(t,l)}) \forall l \neq k$. Identical operation were performed on the training data collection. The feature vector in Eq.(1) is augmented and a new target $d_{((t,k),(t,l))} = y_{(t,k)} - y_{(t,l)}$ is learnt. The difference of METEOR scores y. It represents the relative translation quality of (t, k) with respect to (t, l).

Finally, a new metric for $z_{(t,k)}$ is computed by averaging all relevant predictions $\hat{d}_{((t,k),\cdot)}$,

$$z_{(t,k)} = \frac{1}{L} \sum_{l \neq k} \hat{d}_{((t,k),(t,l))}. \qquad (2)$$

In this study, pairwise regression with varying degrees of K from 3 to 10 will be tried.

3.2. Binary classification with M-plet

The method of pairwise feature concatenation can be extended to ordered triplets, ordered quadruplets and ultimately ordered M-plet where M is equal to the order of K-best. The augmented feature vector is thus in the form $(\boldsymbol{x}_{(t,k)}, \boldsymbol{x}_{(t,l_1)}, \cdots, \boldsymbol{x}_{(t,l_{(M-1)})})$, $\forall [l_1, \cdots, l_{(M-1)}] \neq k$. A long feature vector with $M > 2$ potentially correspond to comparison of k^{th}-best with other $M - 1$ candidates. The support vector regression formulation above, which captures the difference of scores of 2 candidates, can no longer be used to model this kind of relationship. Thus, a binary classification task is formulated as follows,

$$b_{((t,k),(t,l_1),\cdots,(t,l_{(M-1)}))} = \begin{cases} 1, & \text{if } k = \arg\max_l y_{(t,l)}, \\ -1, & \text{otherwise.} \end{cases} \qquad (3)$$

In SLT quality estimation, a soft estimate of $\hat{b}_{((t,k),(t,l_1),\cdots,(t,l_{(M-1)}))}$ was computed and they are averaged to give $z_{(t,k)}$ in the same way as averaging d to give z in (2). In this study, M-plet classification with M varying from 2 up to K would be tested for different K-best settings. The number of training samples (combinations of M-plets) duplicates $\frac{K!}{(K-M)!}$ times, which is an exponential factor of M. For quadruplets of 8-best, this means a 1680 times increase of training size. In this experiment, K varied from 3 to 10. For each K, different M where the duplication factor < 100 would be tested.

3.3. Comparison to other methods

From the literature, pairwise and M-plet feature constructions accompany with customised kernel functions to reduce the space complexity of the very high dimensional features [17, 18]. This is not necessary in our experimental setup.

The above formulation is also related to ordinal regression. It covers problems in social science and information retrieval where the target labels are mostly generated by human and are not continuous [20, 21]. It can be readily modelled with rank SVM [22]. However, in the SLT reranking problem, y (METEOR) is continuous and has a higher granularity. The fine details of information in y is retained in the regression setup, while the classification setup simulates a rank SVM.

4. DATA

The ASR and MT systems in SLT were trained on large amount of data. For ASR, the acoustic models were trained on TED data, augmented by the lecture archives from the liberated learning consortium (LLC) and the Stanford University's entrepreneurship corner (ECRN) [23, 24], with a total duration of 298 hours. ASR language models were trained on TED data (3.17M words) augmented with broadcast news transcripts and parliamentary minutes from News commentary, Commoncrawl, Gigaword and Europarl with data selection, leading to a total of 703.9M words. For MT, the text data for language and translation models training were mostly taken from WMT14 [25], supplemented with the official in-domain TED data in IWSLT evaluations [26]. The training text for language and translation models contain 560.35M and 31.47M words respectively. Language model adaptations and MT system tuning were performed on the IWSLT 2010 development and test data (44K words).

The quality estimation (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).

5. EXPERIMENTAL SETUP

5.1. ASR and MT

The SLT task reported in this paper is an English-speech-to-French-text translation task on TED talks data [27].

The English 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.

The English-to-French MT system was a phrase-based system with standard setting [28]. The phrase length in translation model and order of N-gram in language model is 5. An English monolingual translation model frontend was used to recover casing and punctuation from the ASR output.

5.2. Reranking with groupwise learning

The quality estimation (QE)-informed ASR k-best list reranking described in [29] was conducted. In brief, the SLT system was applied on the QE training and test data (§4). The top K ASR and their 1-best MT results were generated. For each of the K-best candidates $(t,1)$, ... , (t,k) , ... , (t,K) in sentence t, a feature vector $\boldsymbol{x}(t,k)$ with 116 dimensions as described in §2 were extracted. A QE model was trained and it was used to predict the sentence translation quality to rerank the K-best sentences.

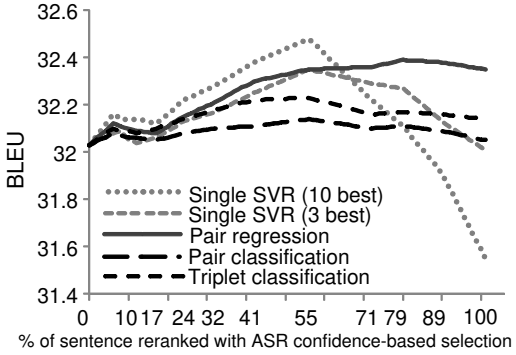
To test the proposed groupwise learning models, two types of feature concatenation following §3.1 and §3.2 respectively were carried out. For each method, pairs, triplets, or M-plet features (different sizes of groups, M) would be tested under different ASR K-best scopes (different K values). For each regression/classification setting with particular K and M values, new models were trained and quality metrics $z_{(t,k)}$ were computed to replace the prediction with the single SVR model (Eq. (1)). These predictions were used to reranked the K-best candidates and the resulting BLEU scores across different settings were compared.

The ASR confidence-informed heuristic in reranking was also revisited, where different thresholds were applied and reranking was only conducted on sentences with lower average word confidence reported from on the 1st-best ASR [29].

To illustrate the stability of performance, the whole experiment were replicated in two extra settings with progressive introduction of domain mismatch [12]. The default setting was labelled as Setting A, where both ASR and MT systems are in-domain. The MT system in setting B were slightly off-domain and further domain mismatch in ASR system were introduced in setting C. In summary, SLT performance degraded from Setting A to B to C. Details of these settings can be referred to in [12].

Based on the performance, one optimal groupwise learning configuration was chosen and linear discrimination analysis (LDA) was carried out on the features. LDA aims to find a projection of the feature vector to a low dimensional space subject to the Fisher criterion, and was shown to give an extra 0.04-0.11 BLEU score increase in previous SLT enhancement experiment [12].

Fig. 1. 3-best reranking with in-domain ASR and MT



6. RESULTS

6.1. Groupwise learning with 3-best candidates

To gain an insight to the effectiveness of groupwise learning, the reranking case with 3-best ASR and their 1-best MT hypotheses (i.e. K=3) with in-domain ASR and MT was studied. Figure 1 shows three groupwise learning models (one regression with M=2; two classifications with M=2[pair] and M=3[triplet]) compared with two control *single SVR* models with K=3 and K=10.

The vertical axis shows the BLEU score from using different models. The horizontal axis shows the increasing percentage of sentences being reranked using the confidence-informed heuristics. The baseline performance is 32.03 (where 0% of sentences were reranked). From 10-best single SVR to 3-best single SVR, the best performance dropped from 32.48 to 32.35 (with 55% sentences reranked). This is because of the reduced scope of potential improvement with lower-order K-best.

When focusing on the groupwise learning models, the pair regression model was found to give the same performance as the 3-best single SVR (32.35) at the 55% data selection point. The two classification models give 32.14 and 32.09 BLEU scores respectively. The two single SVR models require data filtering, as illustrated by the significant drop of BLEU beyond 55% sentence selection. The three groupwise learning models are more robust in reranking sentences with high ASR confidence. In the following experiments, two thresholds on average word confidence would be used (i) 0.96, this is the empirical optimal threshold from previous experiment (ii) 1.00, reranking is only skipped for sentences with average word ASR confidence equal 1.00, this corresponds to roughly 10% of the sentences.

6.2. Groupwise learning up to 10-best

In this Section, the groupwise learning models with regression and classification were explored with varying orders of

Table 1. BLEU score with groupwise learning under different K, M, confidence selections and domain mismatch settings

K-best order		3		4			5			6		7	8	9	10
Size of group (M)		2	3	2	3	4	2	3	4	2	3	2	2	2	2
Setting A (In-domain ASR, In-domain MT, Baseline: 32.03)															
Regression	(55%)	32.35	–	32.55	–	–	32.59	–	–	32.50	–	32.53	32.56	32.57	32.66
(% selected)	(89%)	32.38	–	32.58	–	–	32.63	–	–	32.55	–	32.60	32.59	32.58	**32.72**
Classification	(55%)	32.14	32.23	32.29	32.23	32.06	32.24	32.13	32.22	32.39	32.03	32.35	32.45	32.57	32.60
(% selected)	(89%)	32.09	32.16	32.18	32.20	32.09	32.21	32.13	32.25	32.32	32.03	32.26	32.36	32.51	32.51
Setting B (In-domain ASR, Out-of-domain MT, Baseline: 30.64)															
Regression	(55%)	31.02	–	31.10	–	–	31.07	–	–	31.04	–	31.09	31.07	31.18	31.16
(% selected)	(89%)	31.06	–	31.13	–	–	31.02	–	–	30.96	–	31.06	31.06	**31.23**	31.19
Classification	(55%)	30.81	30.83	30.86	30.76	30.70	30.92	30.64	30.87	30.77	30.64	30.86	30.90	30.91	30.97
(% selected)	(89%)	30.77	30.79	30.87	30.74	30.65	30.89	30.64	30.78	30.72	30.64	30.87	30.95	31.02	31.10
Setting C (Out-of-domain ASR, Out-of-domain MT, Baseline: 29.41)															
Regression	(59%)	30.02	–	29.95	–	–	30.09	–	–	30.17	–	30.14	30.22	30.30	30.25
(% selected)	(90%)	30.17	–	30.15	–	–	30.22	–	–	30.30	–	30.31	30.36	**30.48**	30.43
Classification	(59%)	29.67	29.62	29.75	29.63	29.61	29.88	29.68	29.82	29.96	29.46	29.94	30.04	30.12	30.23
(% selected)	(90%)	29.68	29.65	29.82	29.67	29.69	29.99	29.74	29.92	30.00	29.46	30.04	30.23	30.29	30.42

ASR K-best (K) and sizes of groups (M) under the three domain mismatch settings.

Table 1 summarises the performance in terms of BLEU. The regression models learnt from pairwise features so M always had a value of 2. The classification models had the values of M varied from 2 up to K. From $K = 5$ onwards, the growing space complexity limits the upper bound of M to be tried. Three settings with increasing domain mismatch, with baseline BLEU score equal to 32.03, 30.64 and 29.41 were tested. Following the previous experiments with 3-best, reranking with two ASR confidence thresholds were reported. These thresholds roughly correspond to $55\% - 59\%$ and 90% of data being reranked in the three settings.

In general, performance improves with K because of the larger potential scopes with longer K-best lists. Across different settings, the regression models were better than the classification models across all K, while the performance gaps are closing up when $K \geq 9$. There is not a conclusive trend observed with the increase of group size M. The use of ASR-confidence threshold (selecting 55% of the sentences to rerank) seems to be necessary only in groupwise classfication with Setting A. Even for this particular setting, missing out sentence selection only brings < 0.1 BLEU degradations.

The best performance for setting A, B and C are marked with bold fonts and underlined in Table 1. They all using groupwise regression model with $K = 9$ or 10 and 90% sentences were reranked. For consistency, the configuration with $K = 10$ was for further experiments and result comparison.

Table 2. BLEU with all techniques in 3 settings

	A	B	C
Baseline	32.03	30.64	29.41
Single SVR [12]	32.44	31.08	29.94
Single SVR + LDA [12]	32.53	31.12	30.08
Groupwise	32.72	31.19	30.43
Groupwise + LDA	32.83	31.26	30.62

For this setting, the BLEU score for setting A, B and C are 32.72, 31.19 and 30.43 respectively.

Table 2 showed the performance comparison with different techniques. Compared with the single SVR method, groupwise learning contribute 0.28, 0.11 and 0.49 BLEU increase.

6.3. Groupwise learning with LDA

In the final experiment, LDA was applied on the specified groupwise learning condition discussed above. The dimension of projection varied from 3 to 10 and the optimal results were included in Table 2. LDA on top of groupwise learning brings additional 0.11, 0.07 and 0.19 BLEU score increase to Settings A, B and C respectively. The optimal LDA projection dimensions for these these settings are 3, 5 and 4 respectively.

7. CONCLUSION

In this paper, a groupwise learning strategy was proposed for the SLT reranking problem. Groups of 2 up to K sentences from the ASR K-best list are grouped together and vector-based regression and classification models were used to learn a likelihood metric used for re-ranking. Compared with learning with individual samples, groupwise learning gives 0.11 to 0.49 additional increase to BLEU in three settings. Groupwise learning is complementary to the previously proposed LDA feature projection method, allowing further performance improvement. Space complexity is an issue. Unlike conventional vector-based classification problem where special kernels and operations are needed for the high dimension, in the formulation of groupwise learning the number of samples grow exponentially. Research in support vector regression like primal training should help [30]. Moreover, the technique could be extended to other non-SLT problems where information are incorporated to redirect a search.

8. REFERENCES

[1] Natalia Segal, Hlne Bonneau-Maynard, Quoc Khanh Do, Alexandre Allauzen, Jean-Luc Gauvain, Lori Lamel, and Franois Yvon, "LIMSI English-French Speech Translation System," in *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, 2014.

[2] 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 *Proc. Interspeech*, 2015, pp. 2247–2251.

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

[4] Alicia Pérez, M. In/'es Torres, and Francisco Casascuberta, "Potential scope of a fully-integrated architecture for speech translation," in *Proc. EAMT*, 2010.

[5] Evgeny Matusov, Stephan Kanthak, and Hermann Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. Interspeech*, 2005, pp. 3177–3180.

[6] Nicola Bertoldi, Richard Zens, Marcello Federico, and Wade Shen, "Efficient speech translation through confusion network decoding," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1696–1705, 2008.

[7] Evgeny Matusov and Hermann 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.

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

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

[10] Chi-Ho Li, Nan Duan, Yinggong Zhao, Shujie Liu, and Lei Cui, "The MSRA machine translation system for IWSLT 2010," in *Proc. IWSLT*, 2010, pp. 135–138.

[11] Ruiqiang Zhang, Genichio Kikui, Hirofumi Yamamoto, Taro Watanabe, Frank Soong, and Wai Kit Lo, "A unified approach in speech-to-speech translation: Integrating features of speech recognition and machine translation," in *Proc. COLING*, 2004.

[12] Raymond W. M. Ng, Kashif Shah, Lucia Specia, and Thomas Hain, "A study on the stability and effectiveness of features in quality estimation for spoken langauge translation," in *Proc. Interspeech*, 2015.

[13] Lucia Specia, Kashif Shah, José G. C. de Souza, and Trevor Cohn, "QuEst - A translation quality estimation framework," in *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics: Demo Session*, 2013, p. 794.

[14] 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.

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

[16] Alex J. Smola and Bernhard Schölkopf, "A tutorial on support vector regression," 1998.

[17] Faqiang Wang, Wangmeng Zuo, Lei Zhang, Deyu Meng, and David Zhang, "A kernel classification framework for metric learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1950–1962, 2015.

[18] Carl Brunner, Andreas Fischer, Klaus Luig, and Thorsten Thies, "Pairwise support vector machines and their application to large scale problems," *Journal of Machine Learning Research*, vol. 13, pp. 2279–2292.

[19] Li Liao and William Stafford Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of Computational Biology*, vol. 10, no. 6, pp. 857–868, 2003.

[20] Wei Chu and S. Sathiya Keerthi, "Support vector ordinal regression," *Neural Computation*, vol. 19, no. 3, pp. 792–815, March 2007.

[21] Eyke Hllermeier, Johannes Frnkranz, Weiwei Cheng, and Klaus Brinker, "Label ranking by learning pairwise preferences," *Artificial Intelligence*, vol. 172, no. 167, pp. 1897 – 1916, 2008.

[22] Thorsten Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2006, KDD '06, pp. 217–226, ACM.

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

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

[25] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna, "Findings of the 2014 workshop on statistical machine translation," in *Proceedings of WMT14*, 2014, pp. 12–58.

[26] Mauro Cettolo, Christian Girardi, and Marcello 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.

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

[28] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan 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.

[29] Raymond W. M. Ng, Kashif Shah, Wilker Aziz, Lucia Specia, and Thomas Hain, "Quality estimation for ASR K-best list rescoring in spoken language translation," in *Proc. of ICASSP*, 2015.

[30] Olivier Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, no. 5, pp. 1155–1178, 2007.