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

Quality estimation models are used to predict the quality of the output from a spoken language translation (SLT) system. When these scores are used to rerank a k-best list, the rank of the scores is more important than their absolute values. This paper proposes groupwise learning to model this rank. Groupwise features were constructed by grouping pairs, triplets or M-plets among the ASR k-best outputs of the same sentence. Regression and classification models were learnt and a score combination strategy was used to predict the rank among the k-best list. Regression models with pairwise features give a bigger gain over other model and feature constructions. Groupwise learning is robust to sentences with different ASR-confidence. This technique is also complementary to linear discriminant analysis feature projection. An overall BLEU score improvement of 0.80 was achieved on an in-domain English-to-French SLT task.

***Index Terms*—** groupwise learning, spoken language translation

1. INTRODUCTION

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

ASR and MT systems are usually large and complex. Considerable efforts are necessary to adapt or integrate system components. An alternative to adapting the models is to rerank/rescore the search hypotheses obtained in the decoding 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 work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology) and Google.

Distinctive features derived from ASR and MT could be used to identify optimal SLT models [10] or results [9, 11, 12]. 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. Based on the predicted quality, the 10-best ASR hypotheses were reranked subject to optimal SLT performance [13].

In the above work, a global model was learnt to generate a score for a single hypothesis at a time. In this study we look at SLT enhancement as a groupwise learning problem, where pairs (or groups) of the ASR k-best outputs are compared. We show that groupwise learning followed by a score combination strategy can effectively capture the ranking relationship between good and bad translations.

2. FEATURES FOR QUALITY ESTIMATION

The objective of this work is to make best use of quality estimation features for the purpose of restoring the optimal hypotheses in an SLT system. We use 116 features to represent the property of a sentence, mostly related to its performance in the ASR and MT systems. 21 features were extracted from the ASR system output. These features describe the decoder scores from the acoustic and language models, the ASR k-best rank information and other count statistics. 95 features related to MT was extracted using an open source toolkit QUEST (`http://www.quest.dcs.shef.ac.uk`) [14, 15]. Among those, 79 are called "blackbox" features. These 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, which are called "glassbox" features. They describe the confidence of the MT system, such as the global MT model score. The full list of features were detailed in [16].

3. GROUPWISE LEARNING

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 the METEOR score [17], which is an automatic translation quality metric with continuous range. A support vector regression (SVR) model is used to predict the score [18]:

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

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_I$ are the I support vectors from the training data collection, and α_i and α_i^* are the Lagrangian multipliers in the primal problem. $Ker(\cdot,\cdot)$ denotes the kernel function.

Considering the ASR K-best candidates of a particular test sentence t, represented by multiple feature vectors $\boldsymbol{x}_{(t,1)}$, $\ldots$, $\boldsymbol{x}_{(t,k)}$, $\ldots$, $\boldsymbol{x}_{(t,K)}$ that are inter-related. The ranking of $\hat{y}_{(t,k)}$ among k is more important than their absolute values. Hence the objective is to determine the ranking of entries rather than their scores. Similar ordering problems have been studied in handwriting recognition [19], face detection [20] and for protein sequence detection [21]. The main idea is to build models that focus on groups of two or more samples in the training data collection and learn to determine their ranking automatically. Following this intuition, this work proposes to use a pairwise and an M-plet feature combination in the regression or classification setup. This is to be contrasted with the vector-based model in Eq.(1). The control setting, which considers only one single hypothesis at a time, is hereinafter referred to as the *single SVR* model.

3.1. Pairwise regression

For pairwise regression ordered pairs of features are assembled by concatenating two of the K-best candidates of the same sentence t to form $(\boldsymbol{x}_{(t,k)}, \boldsymbol{x}_{(t,l)}) \forall l \neq k$. The feature vectors on the RHS in Eq.(1) thus have doubled their dimensions. A new target $d_{((t,k),(t,l))} = y_{(t,k)} - y_{(t,l)}$ is learnt. d is the difference in METEOR score between (t,k) and (t,l), therefore representing the relative translation quality between k and l. In the testing phase the predicted value $\hat{d}$ is computed. A total rank score $z_{(t,k)}$ is computed by averaging all pairwise metrics $\hat{d}$ related to t,

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

In this study, pairwise regression with varying degrees of K from 3 to 10 was explored.

3.2. Binary classification with M-plets

The method of pairwise feature concatenation can be extended to ordered triplets, ordered quadruplets and ultimately ordered M-plets where M is equal to the size of K-best list. The augmented feature vector thus has 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 high dimensional feature vector with $M > 2$ potentially corresponds to the comparison of the k^{th}-best with other $M-1$ candidates. The support vector regression as formulated above, which captures the difference of scores of 2 candidates, can no longer be used to model this kind of relationship. A binary classification task is formulated as follows,

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

In the testing phase, a soft estimate of $\hat{b}_{((t,k),(t,l_1),\cdots,(t,l_{(M-1)}))}$ was computed. The total rank score $z_{(t,k)}$ is computed by averaging all M-plet metrics $\hat{b}$ related to t, in a way similar to Eq.(2). In this study, M-plet classification with M varying from 2 up to K will be tested for different K-best settings. The number of training samples (combinations of M-plets) grows $\frac{K!}{(K-M)!}$ times, or exponentially in M. For quadruplets of 8-best, this means a 1680-time increase of the training set size. In this experiment, K varied from 3 to 10. For each K, different M where the duplication factor< 100 will be tested.

3.3. Comparison with other methods

In the literature, pairwise and M-plet feature constructions have been used with customised kernel functions to reduce the space complexity of very high dimensional feature sets [19, 20]. This is not necessary in our experiments. The above formulation is also related to ordinal regression, which is used for problems in social science and information retrieval where the target labels are mostly generated by human and are canonical [22, 23]. It can be readily modelled with a rank SVM [24]. However, in the SLT reranking problem, y (METEOR) is continuous valued. The fine details of information in y are retained in the regression setup, while the classification setup simulates a rank SVM.

4. DATA AND EXPERIMENTAL SETUP

4.1. Data

The ASR and MT systems for SLT were built on large datasets. For ASR, acoustic models were trained on TED data, lecture archive data from the liberated learning consortium (LLC), and Stanford's entrepreneurship corner (ECRN) [25, 26], comprising a total 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, yielding a total of 703.9M words. For MT, the text data for language and translation models training were mostly taken from WMT14 [27], supplemented with the official in-domain TED data in IWSLT evaluations [28]. The training data 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 system was trained on features extracted from SLT system input and output. The SLT outputs from the official IWSLT 2011 test set were used for training the QE system. It comprises 818 segments with 1.1 hours of length in English speech and $13K$ words in French text. The QE system was then tested on IWSLT 2012 test data, with 1124 sentences (1.8 hours in English speech, 20K words in French text).

4.2. ASR and MT system

The SLT task in this paper is an English-speech-to-French-text translation task on TED talk data [29]. The English ASR system operated in multiple passes 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 [30]. The phrase length in the translation model and the LM N-gram order was 5. An English monolingual translation model frontend was used to recover casing and punctuation from the ASR output.

4.3. Reranking with groupwise learning

Quality estimation-informed ASR k-best list reranking as described in [16] was conducted. In short, the SLT system was applied on the QE training and test data (§4.1). The top K ASR and their 1-best MT results were generated. For each of the K-best candidates in sentence t, $\{(t,1)\ ,\ldots,\ (t,k)\ ,\ \ldots,\ (t,K)\}$, a feature vector $\boldsymbol{x}_{(t,k)}$ with 116 dimensions as described in §2 was extracted. A QE model was trained and used to predict the sentence translation quality to rerank the K-best sentences.

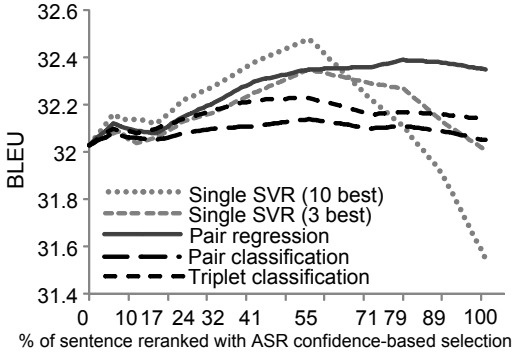
In the proposed groupwise learning, two learning strategies with regression (§3.1) and classification (§3.2). models were investigated. Pairs, triplets, or M-plet features (different sizes of groups, M) were tested under different ASR K-best scopes (different K values). For each regression/classification setting with particular K and M values, new models were trained and quality metrics $z_{(t,k)}$ (Eq. (2)) were computed to replace the prediction with the single SVR model (Eq.(1)). These predictions were used to rerank the K-best candidates. The reranking results for different settings were compared using a common BLEU [31].

Our previous work used an ASR confidence-informed heuristic in reranking, which was revisited here. Different thresholds were applied and reranking was conducted only on sentences with lower average word confidence reported from on the 1st-best ASR [16].

To illustrate the stability of the results, the experiments were replicated in two settings with progressive introduction of domain mismatch [13]. The default setting is *Setting A*, where both ASR and MT systems are in-domain. The MT system in *setting B* was slightly off-domain and further domain mismatch in ASR system was introduced in *Setting C*. In summary, SLT performance degraded from settings A to B to C. Further details of these settings can found in [13].

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 experiments [13].

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



5. RESULTS

5.1. Groupwise learning with 3-best candidates

To gain an insight into 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.

In Figure 1, 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 pairwise 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. Beyond 55% rate of sentence selection there observe significant drops of BLEU in the two single SVM models, showing the importance of the confidence-informed heuristic (§4.3). The three groupwise learning models are more robust in reranking sentences with high ASR confidence.

5.2. Groupwise learning up to 10-best

In this Section, the groupwise learning models with regression and classification were explored with varying orders of ASR K-best (K) and sizes of groups (M) under the three domain mismatch settings A, B, C (§4.3).

Table 1 summarises the performance in terms of BLEU. The regression models were learnt from pairwise features such that 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 tested. Three settings with increasing domain mismatch, with baseline BLEU score equal to 32.03,

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

30.64 and 29.41 were tested.

Concluding from the previous experiments with 3-best, reranking with two thresholds on average word confidence are used (i) 0.96, this is the empirical optimal from previous work, which corresponds to $55-59\%$ of the sentences; (ii) 1.00, reranking is performed on all sentences unless the average word ASR confidence equal 1.00. This corresponds to 90% of the sentences.

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 decrease when $K \geq 9$. No conclusive trend was 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 settings 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 used for further experiments and result comparison. With groupwise learning, the BLEU score for settings 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 contributed 0.28, 0.11 and 0.49 BLEU increase.

Table 2. BLEU with all techniques in 3 settings

	A	B	C
Baseline	32.03	30.64	29.41
Single SVR [13]	32.44	31.08	29.94
Single SVR + LDA [13]	32.53	31.12	30.08
Groupwise	32.72	31.19	30.43
Groupwise + LDA	32.83	31.26	30.62

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.

6. 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 put 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 grows exponentially. Research in support vector regression like primal training could help [32]. Moreover, the technique could be extended to other non-SLT problems where information are incorporated to redirect a search.

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/icassp16-ng.

8. REFERENCES

[1] N. Segal, H. Bonneau-Maynard, Q. Do, A. Allauzen, J.-L. Gauvain, L. Lamel, and F. Yvon, "LIMSI English-French Speech Translation System," in *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, 2014.

[2] N. Ruiz, Q. Gao, W. Lewis, and M. 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] H. 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] A. Pérez, M. I. Torres, and F. Casascuberta, "Potential scope of a fully-integrated architecture for speech translation," in *Proc. EAMT*, 2010.

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

[6] 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.

[7] 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.

[8] G. A. Saon and M. 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] M. Ohgushi, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "An empirical comparison of joint optimization techniques for speech translation," in *Proc. Interspeech*, 2013, pp. 2619–2623.

[11] 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.

[12] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, and W. K. Lo, "A unified approach in speech-to-speech translation: Integrating features of speech recognition and machine translation," in *Proc. COLING*, 2004.

[13] R. W. M. Ng, K. Shah, L. Specia, and T. Hain, "A study on the stability and effectiveness of features in quality estimation for spoken langauge translation," in *Proc. Interspeech*, 2015.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1950–1962, 2015.

[20] C. Brunner, A. Fischer, K. Luig, and T. Thies, "Pairwise support vector machines and their application to large scale problems," *Journal of Machine Learning Research*, vol. 13, pp. 2279–2292.

[21] L. Liao and W. S. 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.

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

[23] E. Hllermeier, J. Frnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences," *Artificial Intelligence*, vol. 172, no. 167, pp. 1897 – 1916, 2008.

[24] T. 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.

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

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

[27] 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.

[28] 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.

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

[30] 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.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

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