The ADAPT System Description for the IWSLT 2018 Basque to English Translation Task

Alberto Poncelas, Andy Way

ADAPT Centre, School of Computing,
Dublin City University, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie

Kepa Sarasola

Ixa Group (UPV/EHU), Faculty of Informatics
University of the Basque Country
kepa.sarasola@ehu.eus

Abstract

In this paper we present the ADAPT system built for the Basque to English Low Resource MT Evaluation Campaign. Basque is a low-resourced, morphologically-rich language. This poses a challenge for Neural Machine Translation models which usually achieve better performance when trained with large sets of data.

Accordingly, we used synthetic data to improve the translation quality produced by a model built using only authentic data. Our proposal uses back-translated data to: (a) create new sentences, so the system can be trained with more data; and (b) translate sentences that are close to the test set, so the model can be fine-tuned to the document to be translated.

1. Introduction

We participated in the Basque to English Low Resource MT Evaluation Campaign as part of the International Workshop on Spoken Language Translation (IWSLT) 2018. In this task, we aimed to build an MT model to translate subtitles of TED (Technology, Entertainment, Design) talks from Basque into English.

Basque (or Euskera), which is mainly spoken in the Basque Country in Northern Spain and Southern France, is considered an isolated language. Linguistically, it is an agglutinative language, and morphologically more complex than English. Furthermore, Basque is a low resource language. Due to these characteristics, creating an MT system that deals with Basque is a challenging task.

As the MT Evaluation Campaign consists of translating subtitles from TED talks, we built our MT engines mainly from available subtitles. TED Talks[1] is an event where experts in different fields, such as education, business, science, etc. give a talk of up to 18 minutes to disseminate their ideas.

The use of subtitles as training data is potentially problematic as they may not be literal translation, causing the original and translated sentences not to be truly parallel. This is because subtitles are subjected to a great deal of adaptation. Localization strategies (adapting the text to suit consumers of a particular locale or culture), combined with the requirement to meet time constraints (where sentences in the source and target languages which have different length are supposed to appear on the screen within the same time frame), results in sentences which are comparable but not necessarily parallel [1].

Although the adaptation does not hinder human comprehension of the intended message, when these sentences are used as training data for an MT model, the translation inaccuracies become obstacles for the system to correctly learn to translate.

The system presented in this paper aims to overcome the aforementioned problems. First, the creation of synthetic data has two purposes: (i) it provides a new set of parallel sentences that mitigates the problem of Basque being a low resourced language; and (ii), artificially-created sentences tend to be more literal than usual translated subtitles. Therefore the former may constitute better training data for an MT model than the latter. Secondly, as TED Talks topics cover a wide variety of domains, we use data selection techniques to adapt an MT model to a particular test set.

The remainder of the paper is structured as follows. In Section 2, we describe related work regarding MT models that include Basque as source or target language. We also describe previous work on the use of synthetic data and data selection algorithms that are related to the systems described below. Section 3 describes the two steps (hybrid corpus creation and model adaptation) performed for building the MT system. In Section 4 we present an estimation of the performance of the models created. Finally, an overview of the system is described in Section 5.

2. Related Work

The system described in this paper is based on two main techniques: (a) incorporating synthetic sentences as training data (Section 2.2), and (b) adapting the model to the test set (Section 2.3).

2.1. Basque Machine Translation

Most of the work on MT involving Basque is based on the Basque-Spanish pair. We can find multiple MT approaches including Rule Based MT (RBMT) [2], or data-driven ap-

[1] https://www.ted.com

proaches [3] such as Example-based MT [4] or hybrid (Statistical MT and RBMT) [5] systems.

Dealing with low-resource languages is a problem for NMT approaches as they require large amounts of data in order to generate good translations. For some language pairs, SMT models can outperform NMT models when trained in limited amount of data [6]. In the work of Unanue et al. (2018) [7] they perform a comparison of Basque-English SMT and NMT models. Their finding reveals that SMT models trained with *PaCo2-EuEn* corpus in the Basque-to-English direction perform better than NMT models. In the reverse direction, however, NMT models can perform better when pre-trained embeddings (which have been trained using additional sentences from Basque Wikipedia) are given to the model.

Regarding Basque-Spanish NMT models, the most recent work is presented by Etchegoyen et al. (2018) [8] where they explored different methods of splitting words into morphemes to improve the translation.

2.2. Addition of Back-translated Sentences

As Basque is a low-resource language, the amount of available parallel data is very limited. A technique to increase the number of sentences is to artificially create sentences. Sennrich et al. (2016) [9], showed that NMT models could be boosted by adding backtranslated data.

Backtranslation in this paper designates the process of translating monolingual sentences in the target language into the source-side language. By doing this, a synthetic parallel corpus is created. Adding this corpus as training data can improve the performance of the model. In fact, models built using solely back-translated data can even achieve comparable performance to those trained with authentic or hybrid data [10].

2.3. Adaptation of the MT Model to the Test Set

There are several techniques for adapting a model to a particular domain [11], such as selecting relevant data (*data-centric* approaches), or modifying the model (*model-centric* approaches).

In the case where the test set is available, it is possible to adapt the model so it performs better in the given test. In our work, we used a combination of data-centric and model-centric approaches. First, we selected data that are relevant for the test set, and then we used fine-tuning to bias the model towards the test set.

Fine-tuning [12, 13], consists of using a pre-built NMT model (trained on general domain data), and training the last epochs on smaller amounts of in-domain data. An alternative to this technique is *gradual fine tuning* [14], which involves reducing the training data as the training proceeds.

While these fine-tuning techniques aim to adapt the NMT models towards a particular domain, Li et al. (2018) [15] proposed to use fine-tuning to adapt the model to the test set, which is closer to our approach. The main difference is that while in their work the model is adapted sentence-wise (one model for each sentence), in ours, it is adapted document-wise (one model for the document).

In order to select sentences that are closer to the test set we used Feature Decay Algorithms (FDA) [16, 17, 18]. This technique has been successfully applied in both SMT [19, 20, 21] and NMT [22].

FDA is a data selection method that not only aims to select sentences that are close to a seed (generally the test set), but also to promote the variability of the training data selected.

In order to achieve that, FDA scores each sentence s in the parallel data, and the sentence with the highest score is added to a list of selected sentences L. The score of the sentence is based on how similar it is to the seed (counting the *n*-grams shared with the seed), and how different it is to previously selected sentences (penalizing *n*-grams already contained in L), which increases the variability.

Using default values of the parameters, the score of a sentence is computed as in Equation (1):

$$score(s|L) = \frac{\sum_{ngr \in s} 0.5^{C_L(ngr)}}{length(s)} \quad (1)$$

where $C_L(ngr)$ is the count of the *n*-gram ngr in the pool of selected sentences L. The more occurrences of ngr there are in L the more penalized ngr is. The factor $0.5^{C_L(ngr)}$ in Equation (1) causes the *n*-gram to contribute less to the total score of the sentence.

3. System Description

The system built consists of two steps. First, (Section 3.2) we created a basic model using authentic and synthetic data. In the second step (Section 3.3), the model was fine-tuned to be adapted to the test set.

3.1. Basque-English Data

The Basque-English parallel data used in this work were obtained by combining the OpenSubtitles2016 (173K sentences), OpenSubtitles2018 [23] (805K sentences) and the *PaCo2-EuEn* corpus[2] [24] (130K sentences) provided in the IT domain MT Shared Task of WMT 2016 [25]. We randomly sampled 5000 sentences as our dev set and the rest (1M sentences) as the training set.

In order to build the NMT models we used OpenNMT-py, which is the Pytorch port of OpenNMT [26]. All the NMT models we built were configured with the same settings (the only difference is the training data used to build them). The value parameters were the default ones of OpenNMT-py (i.e. 2-layer LSTM with 500 hidden units, vocabulary size of 50000 words for each language). All the models were executed for 13 epochs, and we also used Byte Pair Encoding

[2] komunitatea.elhuyar.org/ig/files/2016/01/PaCo_EuEn_corpus.tgz

(BPE) [27] with 30000 merge operations, following the work of Etchegoyen et al. (2018) [8].

3.2. Addition of Synthetic Data

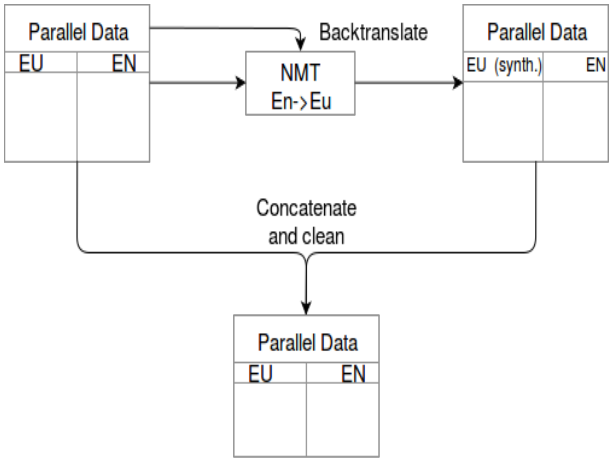

Figure 1: Creation of hybrid parallel corpus.

The first step in the construction of a baseline system is to extend the parallel corpus. In Figure 1 we present a diagram of how we built the corpus. Using an initial corpus of parallel Basque-English sentences we built an NMT model capable of translating sentences from English into Basque. Then, the English side of that parallel corpus was translated into Basque using the English to Basque NMT model.

Intuitively, translating the same sentences that were fed as training data should not be useful as it is likely to produce very similar sentences. However, the sentences produced by the model tend to be more literal translations, thereby avoiding the problems previously mentioned.

In Table 1 we show some examples of how synthetic data present Basque sentences that are closer to literal translation than a human-produced sentences. For example, in the first row, the translation for the English sentence "do I need to be there?" is "joan behar dut?", which literally means "do I have to go?". The artificially-created sentence is a more precise translation, as it uses the verb "be" ("egon") instead of the verb "go" ("joan"). In certain contexts, the use of one or another sentence does not affect the general understanding. However, using the wrong translation as training data for a model can hurt performance.

A similar effect is observed in the second row of Table 1 for the sentence "keep her steady, now.". The Basque translation of this sentence is "ez dadila mugitu." which uses the verb "mugitu" ("to move"), so it could be translated as "it shall not move" or "do not let it move". In contrast, the MT model produced the sentence, "eutsi gogor.", which used the verb "hold" ("eutsi"). Both translations are appropriate, but they belong to different contexts.

Finally, we see in the third row the English sentence "a suicide?". The corresponding sentence in Basque is "nor zen?" ("who was?"). In any other context, the two sentences have completely different meanings. The synthetic sentence by contrast is a literal translation.

	Authentic Basque	Synthetic Basque	English
1	joan behar dut?	hor egon behar dut?	do I need to be there?
2	ez dadila mugitu.	eutsi gogor.	keep her steady, now.
3	nor zen?	suizidioa?	a suicide?

Table 1: Examples of sentences in Basque (authentic), Basque (synthetic) and English translation.

Following backtranslation we obtained two parallel sets, with authentic and synthetic sentences. Next, we concatenated them as a single corpus. Note that, by doing so, the target-language sentences are duplicated.

Finally, we removed those sentences in which the length of the source and target sides differed substantially. In our work we kept a sentence pair (s, t) if $0.5 < \frac{len(s)}{len(t)} > 1.5$, in order to remove the 10% outliers. In total 255K sentences were removed (137K sentences 118K sentences from authentic and synthetic sets, respectively). The hybrid corpus contained, therefore, 1.93M sentence pairs.

We applied these criteria to both corpus of authentic, and synthetic sentences, so the potentially unaligned sentences are ignored and bad translated sentences are not considered, respectively.

3.3. Adaptation to the Test Set

The second step of building the model is to adapt it to a particular test set. The work of Poncelas et al. (2018) [22] showed that when the test set is available during training time it is possible to fine-tune a model to improve the translation of that particular test set.

In Figure 2 we show how we fine-tuned our NMT model, which requires three phases as follows:

1. Data Selection: In this phase we aimed to retrieve English sentences that were close to the test set. As the test set was in Basque, we first created an approximated translation using the NMT model built as explained in Section 3.2. This translation can be used as the seed for FDA and extract a set of sentences, from a monolingual English corpus, that were close to the pre-translated set and hence, to the test set. In this work we extracted 50,000 sentences from English training data provided in the WMT 2015 Translation Task [28].

2. Back-translation: The subset of selected English sentences were back-translated (we reused the same English-to-Basque model built as explained in Section

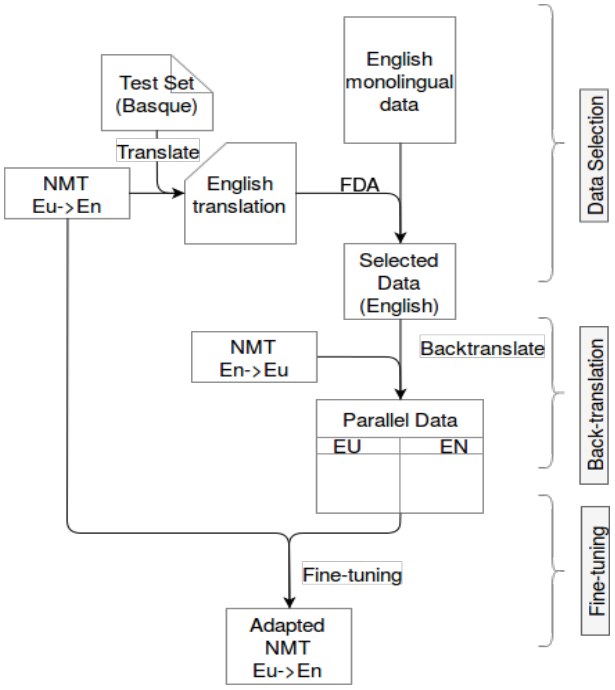

Figure 2: Fine tuning with synthetic data

2.2 to create backtranslated data) in order to build a parallel corpus.

3. Fine-tuning: The synthetic parallel corpus was used to fine-tune the MT model for one epoch. In this way, the model was tailored to the test set.

4. Experimental Results

In order to estimate the performance of the final and intermediate models described through Section 3 we evaluated them using the development set (containing 1K sentences extracted from subtitles of TED talks) provided by the organizers of the IWSLT Evaluation Campaign.

The models evaluated are: (a) the model built with only authentic data (*base model*); (b) the model built with the combination of authentic and synthetic data (*hybrid model*); and (c), the *hybrid model* adapted to the test set using FDA-retrieved data (*FDA model*).

We used several evaluation metrics to compare the outputs of the three models to a human-translated reference. In Table 2 we can see the evaluation scores for each model. The metrics we present are BLEU [29], NIST [30], TER [31], METEOR [32] and CHRF3 [33].

We also marked in bold the scores that outperform those of the *base model* (first column of Table 2) and marked with an asterisk the scores (among BLEU, TER and METEOR) that are statistically significant at level p=0.01. This was computed with multeval [34] using Bootstrap Resampling [35]. The two asterisks in column *FDA model* (METEOR row) indicate that it is statistically significant at p=0.01 when

	base model	*hybrid model*	*FDA model*
BLEU	0.1315	**0.1426***	**0.1450***
NIST	4.459	**4.683**	**4.733**
TER	0.8508	0.8576	0.8666
METEOR	0.1429	**0.1501***	**0.1528****
CHRF3	34.05	**35.92**	**36.24**
CHRF1	37.40	**38.67**	**38.81**

Table 2: Evaluation of the model built only with authentic data and using both authentic and synthetic data.

compared not only to the *base model* but also to the *hybrid model*.

As mentioned in Section 3.2, the addition of synthetic data (even if it consists of a backtranslation of the same data used for training the model) is helpful. This is verified with the results in column *hybrid model* in Table 2. As we can see, most of the *hybrid model* scores of the model are better than the model built with authentic data only (*base model* column) and according to two of the scores, the improvements are statistically significant at p=0.01. In fact, a model built using only synthetic data (Table 3) can achieve improvements over the *base model*, according to METEOR and CHRF3 evaluation metric.

	synth. model
BLEU	0.1224
NIST	4.074
TER	0.9769
METEOR	**0.1481**
CHRF3	**36.22**
CHRF1	36.40

Table 3: Evaluation of the model built with synthetic data only.

Finally, fine-tuning the *hybrid model* using sentences that are close to the test set is also beneficial. As we can see in the column *FDA model* (in Table 2), most of the scores (except TER) are better than those of the *base model* or even the *hybrid model*, and according to METEOR metric the improvement is statistically significant at p=0.01.

5. Conclusion

In this paper we have described the ADAPT system presented for the Low Resource MT Evaluation Campaign of IWSLT 2018. The system translates from Basque into English.

Basque is a morphologically rich language, which causes the task of building an MT model to be more difficult than languages such as Spanish or German. Furthermore, the available parallel Basque-English data are scarce.

Due to the limited resources of texts in Basque, we generated synthetic data that successfully boosted the performance of the MT model trained solely with authentic sentences.

Additionally, we have used a supplementary monolingual English corpus so we could retrieve sentences close to the test set and further improve our model.

6. Acknowledgments

The research leading to these results was carried out as part of the TADEEP project (Spanish Ministry of Economy and Competitiveness TIN2015-70214-P, with FEDER funding). This work has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

7. References

[1] M. Fishel, Y. Georgakopoulou, S. Penkale, V. Petukhova, M. Rojc, M. Volk, and A. Way, "From subtitles to parallel corpora," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, 2012, pp. 3–6.

[2] A. Mayor, I. Alegria, A. D. De Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola, "Matxin, an open-source rule-based machine translation system for Basque," *Machine translation*, vol. 25, no. 1, p. 53, 2011.

[3] G. Labaka, N. Stroppa, A. Way, and K. Sarasola, "Comparing rule-based and data-driven approaches to spanish-to-basque machine translation," in *Machine Translation Summit XI*, Copenhagen, Denmark, 2007, pp. 297–304.

[4] N. Stroppa, D. Groves, A. Way, and K. Sarasola, "Example-based machine translation of the basque language," pp. 232–241, 2006.

[5] G. Labaka, C. España-Bonet, L. Màrquez, and K. Sarasola, "A hybrid machine translation architecture guided by syntax," *Machine translation*, vol. 28, no. 2, pp. 91–125, 2014.

[6] M. Dowling, T. Lynn, A. Poncelas, and A. Way, "SMT versus NMT: Preliminary comparisons for Irish," in *Technologies for MT of Low Resource Languages (LoResMT 2018)*, Boston, USA, 2018, pp. 12–20.

[7] I. J. Unanue, L. G. Arratibel, E. Z. Borzeshi, and M. Piccardi, "English-Basque statistical and neural machine translation." in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018, pp. 880–885.

[8] T. Etchegoyhen, E. M. Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. C. Etxabe, A. J. Carrera, I. E. Santos, and M. M. eta Eusebi Calonge, "Neural machine translation of Basque," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*, Alacant, Spain, 2018, pp. 139–148.

[9] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016, pp. 86–96.

[10] A. Poncelas, D. Shterionov, A. Way, G. M. de Buy Wenniger, and P. Passban, "Investigating back-translation in neural machine translation," in *21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 2018, pp. 249–258.

[11] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," *arXiv preprint arXiv:1806.00258*, 2018.

[12] M.-T. Luong and C. D. Manning, "Stanford neural machine translation systems for spoken language domains," in *Proceedings of the International Workshop on Spoken Language Translation*, Da Nang, Vietnam, 2015, pp. 76–79.

[13] M. Freitag and Y. Al-Onaizan, "Fast domain adaptation for neural machine translation," *arXiv preprint arXiv:1612.06897*, 2016.

[14] M. van der Wees, A. Bisazza, and C. Monz, "Dynamic data selection for neural machine translation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 1400–1410.

[15] X. Li, J. Zhang, and C. Zong, "One Sentence One Model for Neural Machine Translation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018, pp. 910–917.

[16] E. Biçici and D. Yuret, "Instance selection for machine translation using feature decay algorithms," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, 2011, pp. 272–283.

[17] E. Biçici, Q. Liu, and A. Way, "ParFDA for fast deployment of accurate statistical machine translation systems, benchmarks, and statistics," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 2015, pp. 74–78.

[18] E. Biçici and D. Yuret, "Optimizing instance selection for statistical machine translation with feature decay algorithms," *IEEE/ACM Transactions on Audio, Speech,*

and Language Processing, vol. 23, no. 2, pp. 339–350, 2015.

[19] E. Biçici, "Feature decay algorithms for fast deployment of accurate statistical machine translation systems," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013, pp. 78–84.

[20] A. Poncelas, A. Way, and A. Toral, "Extending feature decay algorithms using alignment entropy," in *International Workshop on Future and Emerging Trends in Language Technology*, Seville, Spain, 2016, pp. 170–182.

[21] A. Poncelas, G. M. de Buy Wenniger, and A. Way, "Applying n-gram alignment entropy to improve feature decay algorithms," *The Prague Bulletin of Mathematical Linguistics*, vol. 108, no. 1, pp. 245–256, 2017.

[22] A. Poncelas, G. M. Buy Wenniger, and A. Way, "Feature decay algorithms for neural machine translation," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 2018, pp. 239–248.

[23] J. Tiedemann, "News from OPUS - A collection of multilingual parallel corpora with tools and interfaces," in *Recent Advances in Natural Language Processing*, N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, Eds. Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009, vol. V, pp. 237–248.

[24] I. San Vicente, I. Manterola, *et al.*, "PaCo2: A fully automated tool for gathering parallel corpora from the web." in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 2012, pp. 1–6.

[25] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, *et al.*, "Findings of the 2016 conference on machine translation." in *ACL 2016 First Conference on Machine Translation (WMT16)*, Berlin, Germany, 2016, pp. 131–198.

[26] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, Vancouver, Canada, 2017, pp. 67–72.

[27] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, Berlin, Germany, 2016, pp. 1715–1725.

[28] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, "Findings of the 2015 Workshop on Statistical Machine Translation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisboa, Portugal, September 2015, pp. 1–46.

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

[30] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the second international conference on Human Language Technology Research*, San Diego, CA, 2002, pp. 138–145.

[31] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, 2006, pp. 223–231.

[32] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, Ann Arbor, Michigan, 2005, pp. 65–72.

[33] M. Popovic, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 2015, pp. 392–395.

[34] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, Portland, Oregon, 2011, p. 176–181.

[35] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004, pp. 388–395.