Enhancing Clinical Information Extraction with Transferred Contextual Embeddings

Zimin Wan*, Chenchen Xu†‡, Hanna Suominen†‡§
*The Australian National University (ANU) / Canberra, ACT, Australia
†Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) / Canberra, ACT, Australia
‡University of Canberra / Canberra, ACT, Australia
§University of Turku / Turku, Finland

Abstract—The Bidirectional Encoder Representations from Transformers (BERT) model has achieved the state-of-the-art performance for many natural language processing (NLP) tasks. Yet, limited research has been contributed to studying its effectiveness when the target domain is shifted from the pre-training corpora, for example, for biomedical or clinical NLP applications. In this paper, we applied it to a widely studied hospital information extraction (IE) task and analyzed its performance under the transfer learning setting. Our application became the new state-of-the-art result by a clear margin, compared with a range of existing IE models. Specifically, on this nursing handover data set, the macro-average F1 score from our model was 0.438, whilst the previous best deep learning models had 0.416. In conclusion, we showed that BERT based pre-training models can be transferred to health-related documents under mild conditions and with a proper fine-tuning process.

Index Terms—Information extraction, Natural language processing, Medical data analysis, Transfer learning

I. INTRODUCTION

Electronic Health (eHealth) solutions support patients and healthcare workers in accessing, authoring, and sharing information. CLEF eHealth evaluation lab has introduced tasks as part of the Conference and Labs of the Evaluation Forum to construct a system that converts verbal clinical handover information into electronic structured handover forms automatically [1], [2]. This processing cascades voice recording, speech recognition (SR), information extraction (IE), and information visualization with text editing and document sign-off by nurses [3].

In this study, we focus on this IE sub-task [2], which is to pre-fill a shift-change nursing handover form by automatically extracting relevant text spans for each slot in the form (Figure 1). Although many machine learning (ML) and natural language processing (NLP) approaches have been applied to it, the performance is not always satisfactory [3]. The main reasons for this gap are the differing vocabularies and context between clinical sublanguages (e.g., nursing vs. medical wards vs. medical-ward nursing) and generic English, along with the ongoing problem of having a limited amount of healthcare-specific data available for training and evaluation; data for real-world healthcare scenarios for research and development are low-resourced [4]–[6]. Common methods in (deep) ML do not achieve a satisfying generalization to unseen data on this clinical task and consequently, the current state-of-the-art approach uses transfer learning (TL) by adapting knowledge from general via medical to clinical English to overcome data insufficiency [7]. However, this remedy relies on the selection of the source domain and requires a large amount of time to pre-train the model on the source domain.

Consequently, our aim is to analyze the adaptability to a domain shift of the timely Bidirectional Encoder Representations from Transformers (BERT) method in the nursing handover IE task. BERT has caused a stir in the ML-based NLP community by obtaining new state-of-the-art results in various tasks [8]. BERT applies the bidirectional training of the Transformer model from deep learning to language modelling. By incorporating in-sentence word-to-word semantic relations and the cross-sentence context knowledge, it addresses the gap in linguistic understanding that limited data in typical biomedical and clinical NLP data sets are unable to provide. The subsequent fine-tuning process then provides extra domain specific knowledge in these health-related data.

Our main contributions are twofold:

- Our application of BERT under a TL framework to the clinical IE task achieves the new state-of-the-art result.
- Conducted analysis to verify our framework’s capability to cope with a domain shift from the pre-training corpora to the target task with scarce data available for training and evaluation.

II. BACKGROUND

IE is a common task in NLP that aims to automatically extract structured information from free-form or semi-structured text. Template filling is a subtask in IE which is to fill out a structured form by extracting text spans. It could be seen as an ML task called sequence tagging or labelling. Its major difficulty is to build a good language model to understand language representation. ML can learn features and representations from big data, but rich, ambiguous human language is a difficult data modality to understand. A word embedding is a distributed representation of words that models a given language by taking the word context into account and embeds words into vectors [9].
A **conditional random field** (CRF) is a probabilistic sequence model, which can learn the feature representation of the words, while taking the context of words into account [10]. Linear chain CRFs are popular in NLP to solve a sequence tagging tasks, and hence, their CRF++ implementation [11] is used as the benchmark method in the handover IE task with both unigrams and bigrams (which combine predicted label for the previous word and features of the current word) and the following features [2], [3]: the previous, current, and next location; pairwise correlations of the previous and current location and the current and next location; and the combination of all features in the current location.

TL is a (deep) ML method that can learn knowledge from solving one domain and apply it to a different problem. It is becoming popular as a way to use pre-trained models as a starting point towards more task-specific solutions in NLP [12] and CV [13]. As in deep learning, many hidden layers exist between the input and output layer; through the data feed-forward, the lower layers tend to extract common and non-domain features, while the higher layers are likely to concatenate learned features and generate higher-level concepts [14], [15]. Thus, the learned weights can be reused if the structure of the higher layers are consistent.

In the handover IE task, TL has been applied to first train a linear chain CRF on the source domain as a starting point, then, the learnt weights of the source model have been used to train a 2-layer CRF to predict a target domain label, and finally, the parameters of the target domain model have been initialized by using the product of the second layer weights and the train source model weights [7]. Besides, the baseline model of TL uses different corpora as the source domain, and so far, the model with the best performance has used a general domain source corpus, called *Bolt, Beranek and Newman* (BBN) [17].
III. MATERIALS

This study used the public NICTA Synthetic Nursing Handover Data Set, which was developed for clinical SR and IE related to nursing shift-change handover from 2012 to 2016 [3]. The data set included three mutually exclusive subsets and 301 synthetic patient records in total (training, validation, and test set with 101, 100, and 100 records, respectively). In this study, our goal was to extract relevant text spans from the written free-form text documents in order to generate the written structured documents. The structured documents included annotations of five classes and each class was further divided into 38 subclasses and an additional category of not applicable (N.A.) for irrelevant information.

The IE task was challenged by the nursing notes not always being consistent with formal general written English. For example, the data contained clinical jargon and (ambiguous ad-hoc) shorthand (yrs for years and BP for Bachelor of Pharmacy, bedpan, before present, birthplace, or blood pressure). Text spans relevant to the same subclass were also sometimes scattered across the text document. Finally, for some labels, as few as two or three instances were present in the training set.

In fact, not all documents included relevant information for all form categories and some documents contained variants of the desired information multiple times.

IV. PROCESSING METHODS

Attention and Transformer mechanisms are of the significant breakthroughs in language modelling, which have greatly promoted the development of NLP [16] and CV [18]: The former was introduced to solve the bottleneck of information pass caused by the conversion of long sequences to fixed-length vectors with recurrent neural networks (RNNs). The latter uses self-attention mechanisms and completely abandons the architecture of RNNs, which can model long term dependencies among tokens in a temporal sequence more effectively (Figure 2).

As Transformer contains no recurrence and convolution layers, the position information should be injected to enable the model to take advantage of the sequence order. Thus, its input is the sum of input embeddings and the positional encodings.

Transformer can model long term dependencies of a sequence and implement parallel computing to increase training speed by using a self-attention mechanism: In the encoder stack, each of the \( N \) identical layers consists of a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Besides, each sub-layer uses the residual connection followed by a normalization layer at its output. The decoder stack is similar to the encoder stack, with \( N \) = 6 identical layers. It inserts a multi-head attention sub-layer over the output of the encoder. Similarly, each sub-layer uses the residual connection followed by a normalization layer at its output.

The BERT model is designed to pre-train deep bidirectional language representations from unlabeled text. It has become the new state-of-the-art in at least eleven NLP tasks [8], and hence, in this study, we have used it as the language representation model. Transformer is used in BERT to learn contextual relations between words in a text sequence. Because BERT aims to generate a language representation model, it only takes use of Transformer’s encoder stack. BERT pre-trains deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

The following two sizes of the pre-trained BERT model exist [8]:

1) \( \text{BERT}_{\text{BASE}} \) with the number of transformer blocks \( L = 12 \), hidden layer size \( H = 768 \), attention heads \( A = 12 \), and number of total parameters \( P = 110 \) million and
2) \( \text{BERT}_{\text{LARGE}} \) with \( L = 24 \), \( H = 1,024 \), \( A = 16 \) and \( P = 340 \) million. In this study, we have used the \( \text{BERT}_{\text{BASE}} \) model.

Generally, BERT takes either one or two sentences as an input, uses WordPiece embeddings [19] to convert the sentence into a token list, and expects special tokens \([CLS]\) and \([SEP]\) to identify the sentences. WordPiece embeddings perform textual preprocessing to convert the free-form text to a list of tokens in order to balance the flexibility of characters and efficiency of words. One of the main implementations of WordPiece is the Byte Pair Encoding (BPE) double-byte encoding. The general training process of BPE is first to divide the word into characters, second, to count the number of occurrences of the character pair within the range of the word, and third, to iteratively add the most frequent character combinations to the vocabulary. Thus, BPE can split a word further so that the vocabulary will be streamlined and the meaning of each word will be clearer. For a given token, BERT’s input representation is the sum of the corresponding token, segmentation, and position embeddings as illustrated in Figure 3. The token embeddings are generated by WordPiece, the segment embeddings are the sentence identifiers, and the position embeddings are used to show the token position within the sequence.

\[^1\text{i.e., PATIENT INTRODUCTION, MY SHIFT, APPOINTMENTS, MEDICATION, and FUTURE CARE}^1\]
\[^2\text{e.g., Clinician’s Last Name and Ward}^2\]
\[^3\text{e.g., for a patient whose given name was Timothy, also his nickname of Tim was supposed to be extracted to populate PATIENT INTRODUCTION: Given names/initials}^3\]

The output representation is the word embeddings generated by the pre-trained networks. By using BERT, they will be 768-
dimensional vectors, which can be then used as features in ML-based NLP models.

In IE and sequence labelling tasks, the model will receive a text sequence, and label each token to one of the various classes. A sequence model can be simply added on top of BERT by connecting BERT’s hidden-states output with a token classifier (Figure 4). It is trainable by feeding the output representation of each token into the classification layer to predict the corresponding label. We add a multi-class classification layer followed by sequential cross entropy loss as the optimization target. Fine-tuning all the parameters end-to-end to model the IE downstream task with BERT is straightforward. Most hyperparameters for the fine-tuning process are the same as in pre-training; and only learning rate, batch size, and the number of training epochs are chosen as according to the validation set [3].

V. EVALUATION METHODS

We have evaluated models using the Precision, Recall, and F1, together with confusion matrices. The macro-average is computed over the multiple classes to emphasize the significance of performing well in all classes, including also those with relatively rare occurrence frequency, due to the risk averse nature of clinical judgement.

Because some categories in the training set do not appear in the test set and vice versa we have calculated the performance of proposed model on the test set based on the categories that occur in the training set. This follows the conventions of the earlier studies [2], [3], [7].

VI. RESULTS

Overall, our fine-tuned model had the macro-average precision, recall, and F1 score of 48.5%, 47.7%, and 43.8%, respectively (Table I). It excelled in nine subclasses out of the 36 sub-classes with F1 ≥ 0.8 (Table I). Its performance was good for the main categories of PATIENT INTRODUCTION (F1 = 68.2%) and MY SHIFT (F1 = 60.9%). Detailed performance result per category is shown in Figure IV and our proposed model is able to achieve promising result even in minor categories which only contain minimal number of examples in the training set. Table II demonstrates the confusion matrix result from our model.

When benchmarked against the previous works, our fine-tuned resulting model outperformed the previous state-of-the-art TL method using BNN as a source domain and CRF based method. It outperformed the second best method by 2.2% for Macro F1 score, with both promising precision and recall rate (Table I).

VII. CONCLUSION

The BERT-based model became the new state-of-the-art in the clinical handover task. It performed well in IE of names, numbers, and some behaviour of the patient with their own specialized, distinguishable vocabulary but failed with complex descriptive content. TL could partially overcome the training problem of CRF with small data by learning features and reusing the weights pre-trained from other domains (i.e., general English) to adapt for the current task [7].

Our fine-tuned model outperforming the prior state-of-the-art illustrates the usefulness of TL of language representations for biomedical NLP applications. The biomedical corpora embeds unique word distributions and different semantics that are mostly not visible in the general corpora used to train initial BERT model. We show that the domain shift can be

---

5 e.g., diet and daily activities

6 e.g., the disease and problem history and chronic condition, which are rare (common) in the generic English (clinical), were difficult to classify, because BERT’s pre-training data.

---

**TABLE I** COMPARISON TO BASELINE MODELS.

| Method               | MAd Precision | MAd Recall | MAd F1 |
|----------------------|---------------|------------|--------|
| BERT-based           | 0.485         | 0.477      | 0.48   |
| TL using BBN         | **0.498**     | 0.419      | 0.416  |
| CRF                  | 0.435         | 0.233      | 0.246  |
| Random               | 0.018         | 0.028      | 0.019  |
| Majority             | 0.000         | 0.029      | 0.001  |

**TABLE II** TOP PERFORMING SUBCLASSES WITH F1 SCORE > 0.8. ACRONYMS: PATIENT INTRODUCTION (PI)

| Subclasses                       | F1  |
|----------------------------------|-----|
| PI_CurrentRoom                   | 1.0000 |
| PI_Lastname                      | 0.9950 |
| PI_GivenNames/Initials           | 0.9900 |
| PI_CurrentBed                    | 0.9645 |
| PI_Age/years                     | 0.9444 |
| MyShift_InputDiet               | 0.8618 |
| MyShift_ActivitiesOfDailyLiving  | 0.8604 |
| N.A.                             | 0.8177 |
| PI_UnderDr_Lastname              | 0.8127 |

---

**TABLE III** ANALYSIS OF SUBCLASS PERFORMANCE

---

**Fig. 4. BERT with a multi-class classifier architecture.**

---
## TABLE III

**The confusion matrix of BERT-based model on different categories.**

| CATEGORY                        | Number of Words in Test Set | Number of True Positive | Number of False Positive | Number of False Negative |
|---------------------------------|-----------------------------|-------------------------|--------------------------|--------------------------|
| A. PATIENT INTRODUCTION         | 1,611                       | 1,134                   | 579                      | 477                      |
| 1. Given Names/ Initials        | 100                         | 99                      | 1                        | 1                        |
| 2. Last Name                    | 101                         | 100                     | 0                        | 1                        |
| 3. Age in Years                 | 281                         | 280                     | 32                       | 1                        |
| 4. Gender                       | 178                         | 62                      | 0                        | 116                      |
| 5. Current Room                 | 100                         | 100                     | 0                        | 0                        |
| 6. Current Bed                  | 198                         | 190                     | 6                        | 8                        |
| 7. Under Dr: Given Names/ Initials | 53                       | 0                       | 0                        | 53                       |
| 8. Under Dr: Last Name          | 128                         | 128                     | 59                       | 0                        |
| 9. Admission Reason/ Diagnosis  | 153                         | 146                     | 341                      | 7                        |
| 10. Allergy                     | 38                          | 18                      | 1                        | 20                       |
| 11. Chronic Condition           | 73                           | 4                       | 14                       | 69                       |
| 12. Disease/ Problem History    | 206                         | 5                       | 110                      | 201                      |
| 13. Care Plan                   | 2                           | 2                       | 15                       | 0                        |
| B. MY SHIFT                     | 1,033                       | 717                     | 606                      | 316                      |
| 14. Status                      | 285                         | 234                     | 123                      | 51                       |
| 15. Contraption                 | 43                          | 7                       | 172                      | 36                       |
| 16. Input/ Diet                 | 107                         | 106                     | 33                       | 1                        |
| 17. Output/ Diuresis/ Bowel Movement | 50                       | 26                      | 32                       | 24                       |
| 18. Wounds/ Skin                | 25                          | 20                      | 34                       | 5                        |
| 19. Activities of Daily Living  | 295                         | 262                     | 52                       | 33                       |
| 20. Risk Management             | 40                          | 4                       | 4                        | 40                       |
| 21. Other Observation           | 188                         | 62                      | 156                      | 126                      |
| C. APPOINTMENTS                 | 472                         | 143                     | 227                      | 329                      |
| 22. Status                      | 79                          | 19                      | 147                      | 60                       |
| 23. Description                 | 313                         | 91                      | 44                       | 222                      |
| 24. Clinician: Given Names/ Initials | 1                       | 0                       | 0                        | 1                        |
| 25. Clinician: Last name        | 1                           | 0                       | 1                        | 1                        |
| 26. Date and Time: Day          | 34                          | 14                      | 22                       | 20                       |
| 27. Date and Time: Time         | 39                          | 17                      | 12                       | 22                       |
| 28. Date and Time: City         | 0                           | 0                       | 1                        | 0                        |
| 29. Date and Time: Ward         | 5                           | 2                       | 0                        | 3                        |
| D. MEDICATION                   | 493                         | 138                     | 95                       | 385                      |
| 30. Medicine                    | 296                         | 120                     | 65                       | 176                      |
| 31. Dosage                      | 52                          | 8                       | 18                       | 44                       |
| 32. Status                      | 145                         | 10                      | 12                       | 135                      |
| E. FUTURE CARE                  | 252                         | 88                      | 319                      | 164                      |
| 33. Alert/ Warning/ Abnormal Result | 28                       | 8                       | 8                        | 20                       |
| 34. Goal/ Task To Be Completed/ Expected Outcome | 103                 | 19                      | 279                      | 84                       |
| 35. Discharge/ Transfer Plan     | 121                         | 61                      | 32                       | 60                       |
| F. N.A.                         | 2,652                       | 2,090                   | 370                      | 562                      |
| 36. N.A.                        | 2,652                       | 2,090                   | 370                      | 562                      |

Properly mitigated by a fine-tuning process, giving superior performance without any handcrafting.

However, we believe there are still spaces left to explore on the model architecture to reduce the potential noises introduced by the pre-trained language model. For example, a CRF could be applied at the top layer of BERT instead of a fully-connected layer. This would make it possible to plug in other available clinical features, while in this study, we only used text data. In the meantime, we also propose to validate the capability of TL such models to more diverse applications in bioinformatics, medicine, health sciences, healthcare, and other health-related fields.

### References

[1] H. Suominen, L. Hanlen, L. Goeuriot, L. Kelly, and G. J. Jones, “Task 1a of the clef ehealth evaluation lab 2015: Clinical speech recognition,” in *CLEF 2015 Working notes*, 2015.

[2] H. Suominen, L. Zhou, L. Goeuriot, and L. Kelly, “Task 1 of the CLEF eHealth Evaluation Lab 2016: Handover information extraction.” in *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*, ser. CLEF2016 Working Notes, K. Balog, L. Cappellato, N. Ferro, and C. Macdonald, Eds., vol. 1609. Online: CEUR Workshop Proceedings, 2016, pp. 1–14.

[3] H. Suominen, L. Zhou, L. Hanlen, and G. Ferraro, “Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations,” *JMIR medical informatics*, vol. 3, no. 2, p. e19, 2015.

[4] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. M. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta, “Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances,” *Journal of Biomedical Informatics*, vol. 88, pp. 11–19, 2018.

[5] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1459–1469.
| A. PATIENT INTRODUCTION | 0.662 | 0.7039 | 0.6823 |
|------------------------|-------|--------|--------|
| 1. Given Names/Initials | 0.99  | 0.99  | 0.99   |
| 2. Last Name           | 1.0   | 0.9901 | 0.995  |
| 3. Age in Years        | 0.8974| 0.9964 | 0.9444 |
| 4. Gender              | 1.0   | 0.3483 | 0.5167 |
| 5. Current Room        | 1.0   | 1.0   | 1.0    |
| 6. Current Bed         | 0.9694| 0.9596 | 0.9645 |
| 7. Under Dr: Given Names/Initials | 0.0 | 0.0 | 0.0 |
| 8. Under Dr: Given Last Name | 0.6845| 1.0 | 0.8127 |
| 9. Admission Reason/ Diagnosis | 0.2998| 0.9542| 0.4562 |
| 10. Allergy            | 0.9474| 0.4737 | 0.6316 |
| 11. Chronic Condition  | 0.2222| 0.0548 | 0.0879 |
| 12. Disease/ Problem History | 0.0435| 0.0243| 0.0312 |
| 13. Care Plan          | 0.1174| 1.0   | 0.2105 |

| B. MY SHIFT            | 0.542 | 0.6941 | 0.6087 |
|------------------------|-------|--------|--------|
| 14. Status             | 0.6555| 0.8211 | 0.729  |
| 15. Contraction        | 0.0391| 0.1628 | 0.0631 |
| 16. Input/ Diet        | 0.7626| 0.9907 | 0.8618 |
| 17. Output/ Diuresis/ Bowel Movement | 0.4483| 0.52 | 0.4815 |
| 18. Wounds/ Skin       | 0.3704| 0.8   | 0.5063 |
| 19. Activities of Daily Living | 0.8344| 0.8881| 0.8604 |
| 20. Risk Management    | 0.0   | 0.0   | 0.0    |
| 21. Other Observation  | 0.2844| 0.3298 | 0.3054 |

| C. APPOINTMENTS        | 0.3865| 0.303 | 0.3397 |
|------------------------|-------|-------|--------|
| 22. Status             | 0.1145| 0.2405| 0.1551 |
| 23. Description        | 0.6741| 0.2907| 0.4062 |
| 24. Clinician: Given Names/ Initials | 0.0 | 0.0 | 0.0 |
| 25. Clinician: Last Name | 0.0 | 0.0 | 0.0 |
| 26. Day                | 0.3889| 0.4118| 0.4    |
| 27. Time               | 0.5862| 0.4359| 0.5    |
| 28. City               | 1.0   | 0.0   | 0.0    |
| 29. Ward               | 1.0   | 0.4   | 0.5714 |

| D. MEDICATION          | 0.5923| 0.2799| 0.3802 |
|------------------------|-------|-------|--------|
| 30. Medicine           | 0.6486| 0.4054| 0.499  |
| 31. Dosage             | 0.3077| 0.1538| 0.2051 |
| 32. Status             | 0.4545| 0.6089| 0.1198 |

| E. FUTURE CARE         | 0.2162| 0.3492| 0.2671 |
|------------------------|-------|-------|--------|
| 33. Alert/ Warning/ Abnormal Result | 0.6586| 0.4054| 0.499  |
| 34. Goal/ Task To Be   | 0.3077| 0.1538| 0.2051 |
| 35. Completed/ Expected Outcome | 0.4545| 0.6089| 0.1198 |

| F. N.A.                | 0.8496| 0.7881| 0.8177 |
|------------------------|-------|-------|--------|
| 36. N.A.               | 0.8496| 0.7881| 0.8177 |

| Total                  | 0.4850| 0.4770| 0.4380 |

[6] B. Zoph, D. Yuret, J. May, and K. Knight, “Transfer learning for low-resource neural machine translation,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1568–1575.

[7] L. Zhou, H. Suominen, and T. Gedeon, “Adapting state-of-the-art deep language models to clinical information extraction systems: Potentials, challenges, and solutions,” JMIR Medical Informatics, vol. 7, no. 2, p. e11499, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[11] T. Kudo, “CRF++: Yet another CRF toolkit (2005).” Available under LGPL from the following URL: [http://crfpp.sourceforge.net](http://crfpp.sourceforge.net) 2015.

[12] D. Li, C. Xu, K. Zhang, X. Yu, Y. Zhong, W. Ren, H. Suominen, and H. Li, “Arvo: Learning all-range volumetric correspondence for video deblurring,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 7721–7731.

[13] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., “Google’s neural machine translation system: Bridging the gap between human and machine translation,” arXiv preprint arXiv:1609.08144, 2016.