VANT : A Visual Analytics System for Refining Parallel Corpora in Neural Machine Translation

Sebeom Park[1 2] * Soohyun Lee[1] † Youngtaek Kim[2] ‡ Hyeon Jeon[1] §
Seokweon Jung[1] ¶ Jinwook Bok[1] ‖ Jinwook Seo[1] **

[1] Seoul National University, Seoul, Republic of Korea
[2] Samsung Electronics, Seoul, Republic of Korea

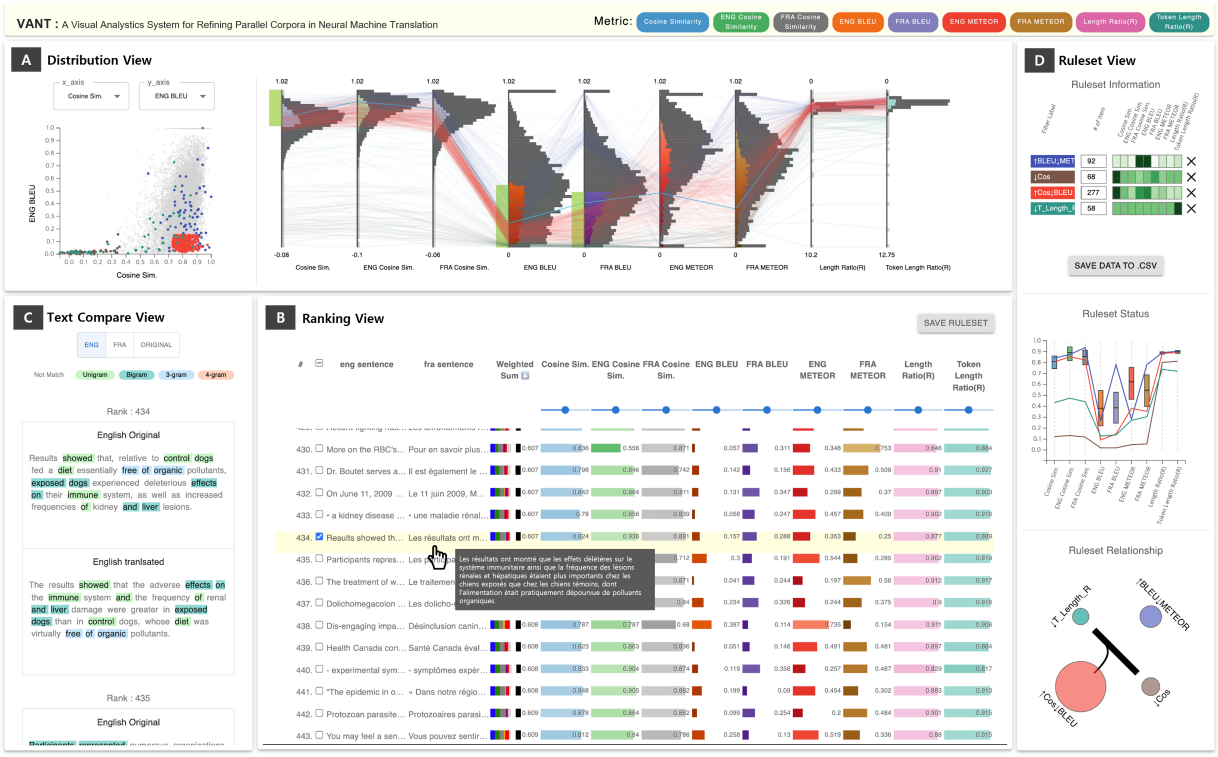

Figure 1: Overview of VANT. (A) Distribution View shows the distribution of different metrics related to the quality of parallel corpora for training a Neural Machine Translation (NMT) model. Users can identify low-quality candidate pairs (i.e., sentence pairs having low metric scores) as noise candidates, which negatively affect the model performance. (B) Ranking View depicts the ranking of noise candidates determined by a user-steerable weighted sum of metrics. Users can select and save a subset of candidates as a ruleset for refinement. (C) Text Compare View allows users to visually inspect noisy parallel corpora in their natural language form. (D) Ruleset View depicts detailed information of the selected rulesets (e.g., average metric scores, the number of common items between rulesets), so that users can analyze the status and pattern of rulesets.

Abstract

The quality of parallel corpora used to train a Neural Machine Translation (NMT) model can critically influence the model's performance. Various approaches for refining parallel corpora have been introduced, but there is still much room for improvements, such as enhancing the efficiency and the quality of refinement. We introduce VANT, a novel visual analytics system for refining parallel corpora used in training an NMT model. Our system helps users to readily detect and filter noisy parallel corpora by (1) aiding the quality estimation of individual sentence pairs within the corpora by providing diverse quality metrics (e.g., cosine similarity, BLEU, length ratio) and (2) allowing users to visually examine and manage the corpora based on the pre-computed metrics scores. Our system's effectiveness and usefulness are demonstrated through a qualitative user study with eight participants, including four domain experts with real-world datasets.

Index Terms: Human-centered computing—Visualization—Visualization system and tools—Visualization toolkits; Computing methodologies—Artificial intelligence—Natural language processing—Machine translation

1 Introduction

Training Neural Machine Translation (NMT) models requires parallel corpora, a set of sentence pairs translated into different languages. In general, parallel corpora are crawled from the web and digitized books, which often ends up with noisy parallel corpora. Such noisy corpora could lead to mistranslation [9]. Since the quality of the

*e-mail: spark@hcil.snu.ac.kr
†e-mail: shlee@hcil.snu.ac.kr
‡e-mail: ytaek.kim@samsung.com
§e-mail: hj@hcil.snu.ac.kr
¶e-mail: swjung@hcil.snu.ac.kr
‖e-mail: bok@hcil.snu.ac.kr
**e-mail: jseo@snu.ac.kr

corpora influences the performance of the NMT model, refining poor-quality(noisy) parallel corpora plays a critical role in improving the model quality itself. Therefore, improving the quality of the NMT model by detecting and removing noisy pairs in the parallel corpora has recently been an important ongoing research topic [12].

Various methods have been utilized to improve the quality of corpora. Users can manually investigate and detect noisy parallel corpora by inspecting each sentence pair within the corpora one-by-one. However, manual inspection is time-consuming, laborious, and also becomes more challenging without the linguistic background of the data. Rule-based automatic filtering techniques were also proposed, which utilized general properties of sentences such as length and word counts difference [11]. There are also model-based automatic filtering tools that exploit the semantic information (e.g., cosine similarity) of corpora [21]. However, these automatic approaches suffer from misclassification (e.g., a high quality sentence pair can be incorrectly classified as noise) [1]; Moreover, as there are diverse types of noisy parallel corpora [9], fully automated approaches may not be robust enough to deal with all types of noisy parallel corpora.

To alleviate this limitation, we present VANT, a visual analytics system for interactively refining parallel corpora in Neural Machine Translation. Our system enables users to readily understand the overall status of large parallel corpora. Users can efficiently identify noise candidates based on derived metrics to evaluate the quality of parallel corpora. By first focusing on the candidates, users can more effectively examine actual noisy parallel corpora and analyze the noise patterns with our visual encoding. We have demonstrated the usefulness and the effectiveness of our system by conducting a qualitative user study with eight participants including four domain experts who work at a major IT company.

2 Related Works

Automatic Refinement of Parallel Corpora Early works for automatic refinement of parallel corpora were based on the general properties of a sentence.

For example, Moses [11] filter sentence pairs based on the length differences between source and target sentences. Since semantic information is not utilized in these conventional approaches, several model-based filtering methods have been further presented. For instance, Xu et al. [20] and Zhang et al. [21] first convert target and source sentences into embedding vectors using a pretrained model, and filters corpora based on a similarity metrics such as cosine similarity between the vectors. However, automatic approaches can still suffer from misclassification issues [1]. VANT migrates such issues by providing an interactive visual analytics system to examine and refine parallel corpora; our system overcomes automatic approaches by combining it with a manual process, thus enables more accurate refinement.

Visual analytics for NMT Most visual analytics systems for NMT are developed for model explanation. For example, Data2Vis [6] provided interactive visualizations to understand Sequence-to-Sequence languages models. Some researches focused on visualizing attention scores [18] for detailed explanation of the translation process. Munz et al. [13] proposed a visual analytics system to help users correct erroneous translations by examining the translation result of monolingual data provided by NMT model.

However, while such visual analytics systems for a NMT model have been proposed, a system for parallel corpora is not yet introduced. Although interactive data wrangling tools with mixed-initiative interfaces can be used to clean up noisy raw data in general [8], they are not suitable for dealing with parallel corpora. Our work aims to fill such gaps by introducing an interactive visual analytics system specialized in examining and refining parallel corpora.

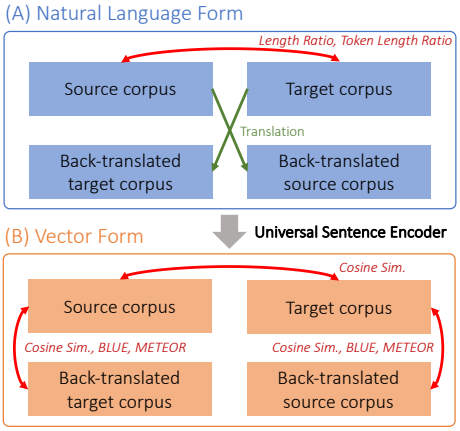

Figure 2: Illustration of the data preprocessing for VANT (Section 4). We used both the natural language form and the vector form of each corpus to widen the range of applied metrics.

3 Design Requirements

We conducted a preliminary interview with two engineers who have more than eight years of career in NMT. During the interview, we mainly discussed about (1) the necessity of detecting noisy sentence pairs and refining parallel corpora, and (2) difficulties in data filtering process. In the interview with the experts, we were able to learn current practices and difficulties in the corpora refinement task. Based on the interview results, we established four design requirements.

- **DR1: Provide a scalable overview of the quality of NMT data.** It requires too much effort for users to identify low quality sentence pairs from large parallel corpora for training NMT model. Thus, the system should show the distribution of metrics scores so that users can readily understand the overall quality of parallel corpora and find noise candidates, a subset of the corpora consisting of sentence pairs that can potentially become real noise.

- **DR2: Recommend noise candidates using multi-metric rankings with user-adjustable weights.** Each metric represents the quality of the parallel corpora from only a single perspective. For more comprehensive refinements, users need to consider several different metrics to find noise candidates from more diverse perspectives. Therefore, the system should provide several metrics and allow users to interactively adjust the weight of each metric to compute combined metric scores for sorting the noise candidates so that users can further examine the noise candidates from more diverse perspectives of their interest.

- **DR3: Enable users with low literacy to inspect the noisy corpora.** It is challenging to evaluate the correctness of the translation through the natural language form of source and target sentences if users are not fluent in both languages. The system should allow users to inspect parallel corpora in the natural language form so that they can determine noise without linguistic background.

- **DR4: Support pattern analysis of noisy parallel corpora.** Since noisy parallel corpora can exist in various forms, it is important for users to analyze the information of items which are previously identified for finding more noisy candidates. Therefore, the system should provide statistical information of actual noise such as metric scores and steerable weights, so that users can not only track the previous selection history but also analyze the characteristics of identified noise to discover patterns. Users can utilize the pattern of existing noisy parallel corpora for the next tasks.

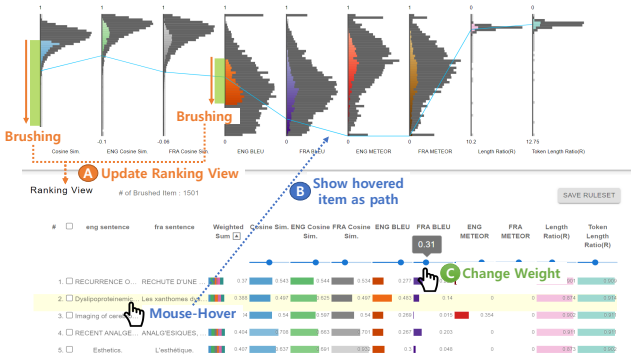

Figure 3: Linking between Distribution View and Ranking View. (A) Users can brush multiple axes on PCP to see details in Ranking View. (B) A mouse-hovered item will be displayed in PCP. (C) When metric weight is changed, the ranking will be updated.

4 Data Preprocessing for Metrics

We prepared multiple metrics that reflect the composite quality of parallel corpora to provide a scalable overview (DR1). Interactively adjusting the weights of the metrics and checking the ranked list of sentence pairs, users can effectively inspect the corpora from diverse perspectives. The preprocessing steps to extract metrics are depicted as follows.

We first extracted length ratio (target/source length) and token length ratio (tokenized target/source length) from parallel corpora that represent their general properties. However, these general properties cannot represent semantic similarity and may not be useful when the language family is different (e.g., Korean and English). To complement the limitation, we extract cosine similarity by using Universal Sentence Encoder [5] as the pre-trained encoder for encoding sentences into embedding vectors regardless of a language type, hence the metric inherently supports universal languages.

Inspired by the back-translation [19], a technique providing monolingual training data with a synthetic source sentence translated from the target sentence into the source language, we translated source and target language into target and source language, respectively, by using Google Translation API. This enabled us to apply two NMT evaluation metrics: 1) BLEU [16], presenting correspondence between a machine's output and that of a human; 2) METEOR [2], based on the harmonic means of n-gram [4] precision and recall. Note that the back-translation result is provided in Text Compare View (Section 5.3), so that the users with less expertise in either source or target language can also use our system (DR3).

In summary, the metrics provided by our system are as follows : *Cosine Similarity* (between source & target sentences, between source & back-translated target sentences, between target & back-translated source sentences), *Length Ratio*, *Token Length Ratio*, *BLEU*, and *METEOR*. The overall pipeline of our preprocessing is shown in Figure 2.

5 Visualization Design

We developed VANT, an interactive visualization system to fulfill the formulated design requirements. As shown in Figure 1, the system consists of four views: Distribution View, Ranking View, Text Compare View, and Ruleset View. The general sequence of using the system is as follows. First, select noise candidates within parallel corpora based on metric scores using Distribution View. Second, check the details of the selected candidates using Ranking View and Text Compare View. Third, select the actual noisy sentence pairs by checking them in Ranking View and save them as a ruleset. Finally, analyze the pattern of the actual noisy sentence pairs in Ruleset View.

5.1 Distribution View

In Distribution View (Figure 1A), users can identify the characteristics of the overall NMT data and obtain clues about noisy parallel corpora (DR1). We provide two visual components. 1) Parallel Coordinate Plot (PCP) to show relationships as a path between each metric score extracted in the preprocessing step; 2) Scatterplot with two user-selected metrics for x and y axes. However, as the size of parallel corpora are usually huge, visual clutter can often occur in PCP. To address this problem, we combined histograms which represent the distribution of each metrics on each axis in a different color, adopting the Parallel Histogram Plot encoding scheme [3]. In addition, the scatterplot enables users to check the correlation between two metrics in more detail. Users can select an interesting range of values of a metric to filter out noise candidates by brushing on the corresponding axis. As shown in Figure 3, the selected candidates are shown in Ranking View to help users check the details of metrics (DR1). Moreover, users can change the order of axes by dragging an axis over other axes. The order of axes is linked to all other views such as Ranking View and Ruleset View.

5.2 Ranking View

The Ranking View (Figure 1B) provides detailed information such as metrics' score and the rankings of noise candidates which are selected in Distribution View (DR2). Since the size of the parallel corpora is huge, the size of user-selected noise candidates from Distribution View may still be too big to explore one by one. Thus, we prioritize noise candidates by the weighted sum of multi-metric scores to enhance users' cognition of noise detection [10]. Determining rankings based on the weighted sum of multi-metric scores can be considered as a multi-criteria decision making (MCDM) problem. Inspired by Lineup [7], we provide a table that shows detailed information with a slider bar for adjusting the weight for each metric. Once the weights are set, Ranking View calculates the weighted sum of each candidate based on individual metric scores, then sorts the candidates by their weighted sum (DR4). Each row of Ranking View shows (1) A natural language form of the paired sentence, (2) weighted sum, (3) and individual metric scores. The weighted sum is represented as a stacked bar, and metric scores are depicted with bars in different colors. The length of the bar represents the metric score and the saturation of the bar shows the ranking of the sentence pair based on the corresponding metric. When users hover the mouse in a row in the table, Text Compare View (Figure 1C) automatically moves to the part corresponding to the hovered item for details and PCP highlights the path related to the item. By examining the candidates, users can determine whether each candidate is an actual noise or not; they can save such selected candidates as a ruleset by clicking "Save Ruleset" button. Note that when users create a ruleset, they should designate the color and the name of the ruleset.

5.3 Text Compare View

Although diverse evaluation metrics are provided in our system, the metrics may not fully reflect the actual quality of parallel corpora. It is thus necessary for users to examine the raw text of the parallel corpora. Since our design requirements considers users who are not literate in one of the source or target languages (DR3), Text Compare View (Figure 1C) offers three language selection options: `source`, `target`, and `source` ↔ `target`. If users select either `source` or `target`, the view depicts source sentence and back-translated target sentence, or target sentence and back-translated source sentence, respectively. When users select `source` ↔ `target` option, the view shows source and target sentence. If users select `source` or `target` option, the system represents the similarity between two sentences by depicting n-gram matching, so that users can more readily identify the commonalities and differences between two sentences. Common unigram, bigram, 3-gram and 4-gram within two sentences are highlighted with different text background colors.

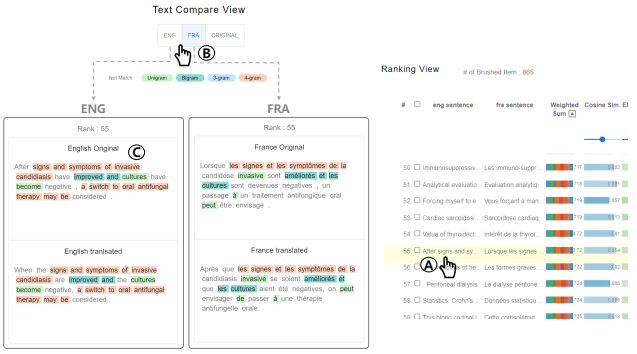

Figure 4: (A) When users mouse-hover an item in Ranking view, Text Compare View automatically moves to the hovered item. (B) Users can select their preferred language for comparison. (C) The background color of texts means n-gram words matching between a paired sentence

5.4 Ruleset View

Once a set of noisy sentence pairs are selected as a ruleset by users in Ranking View, users can analyze the pattern of the ruleset in Ruleset View (Figure 1D) to find more noisy sentence pairs. When a ruleset is created, sentence pairs within the set are highlighted in Distribution View with designated color (DR1), and thus users can easily identify their distribution. Also, to allow users to understand the characteristics of the noise of each ruleset in detail (DR4), the features of rulesets are represented in three subviews: Ruleset Information View, Ruleset Status View, and Ruleset Relationship View. In Ruleset Information View, the metadata of each ruleset—name, color, cardinality, and weight of each metric—are represented (Figure 5A) as a row. The weight of each metric is provided in a heatmap, to enable intuitive comparison between rulesets. When users click on a row, Ranking View and Text Compare View depict the sentence pairs within the ruleset, and corresponding paths and points at PCP and scatterplot in Distribution View are highlighted; this enables users to easily track the history of each ruleset. In Ruleset Status View, the average of metric scores are displayed as a line graph to help users grasp the characteristics of the noise sentence pairs within a ruleset (DR4), Note that the lines are superimposed over a boxplot which represents the statistics of the metric score of the dataset. Finally, in Ruleset Relationship view, the commonalities between rulesets are explained (Figure 5B). In this view, each ruleset is represented as a circle, where the radius of the circle represents the size of ruleset. If two rulesets have common items, they are linked with a line, where the width of the line depicts the number of common items. After users examine rulesets in Ruleset View, they can generate a new dataset where the sentence pairs within the inspected rulesets are filtered out.

6 Qualitative User Study

To demonstrate the effectiveness and usefulness of VANT, we conducted a user study with eight participants in Samsung Research. The participants consist of four domain experts (E1–E4) working in Natural Language Processing team and four professional software engineers (S1–S4) in the Software Engineering team. All participants have more than six years of experience. They are also native in Korean, fluent in English, and have no French background.

We used two real-world datasets for the evaluation: 1) English/French biomedical data from Scielo Corpus [15]; 2) Korean/English news data [17]. We prepared English/French dataset to observe how users use our system without linguistic background.

Our study was conducted in person through the following steps. First, we briefly explained the purpose of our system and the overall design for 15 minutes. We then demonstrated how to detect and filter

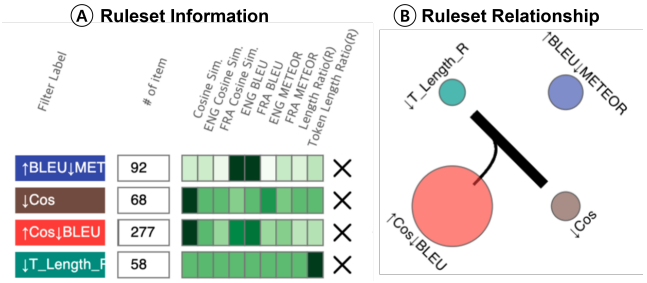

Figure 5: (A) The selected noisy paired sentences from the Ranking View are saved as a ruleset with the information of color-mapped label name, cardinality, and metric weight as a heatmap. (B) The relationship of rulesets is expressed in edge bundling. In this case, four rulesets are saved and three rulesets (↑Cos↓BLEU, ↓Cos, ↓T_Length_R) have common items.

noisy parallel corpora using our system for 20 minutes. Afterward, we asked participants to filter out noisy parallel corpora and refine the dataset with our system for 20 minutes. Lastly, we had a post-hoc interview for feedback.

6.1 Results

Detecting noise candidates In the beginning, all participants mainly investigated the distribution of metric scores using Distribution View in selecting noise candidates (DR1). The participants then selected noise candidates within a specific metric score range (e.g., low BLEU and low METEOR) by brushing on PCP and examining the noise candidates' details in Ranking View (DR2). The domain experts who have a relatively high understanding of the metrics tried to find various noise candidates through interactive exploration in Distribution View and Ranking View. For example, E1, E2, and E4 repeatedly brushed multiple metrics in PCP and adjusted the weights of the metrics in Ranking View. We observed that most experts increased the weights for BLEU and METEOR metrics and decreased the length ratio metric. In addition, some participants (E3, S2, S3) discovered that the low cosine similarity of the corpora does not guarantee their quality using Text Compare View.

Inspecting actual noise from candidates After selecting a set of paired sentences as noise candidates, the participants inspected actual noise from the candidates in Text Compare view (DR3). Most of the participants said that highlighting sentence pairs based on n-gram matching was very helpful to quickly judge whether they were noisy or not. In particular, regarding English-French data, all participants responded that they could easily and quickly compare parallel corpora utilizing back-translated English sentences from French, even though they were not literate in French.

Save rulesets and analyze their patterns Interviewees saved a subset of noise candidates as a ruleset. More than half of the participants (E1–E4, S1) mentioned the status of noisy parallel corpora sets revealed in the PCP and the scatter plot is beneficial in tracking the history of their previous selections. Besides that, E3 and E4 were interested in finding noise patterns through Ruleset View and Distribution View (DR4). E3 figured out which paired sentences were repeatedly selected from Ruleset Relationship View. E4 examined the scatter plot to find patterns while changing x and y axes.

6.2 Post-hoc Feedback

At the end of each session, we asked the participants about the usefulness of our system and possible improvements. Overall, all participants said that the distribution of each metric score represented in Distribution View was helpful in understanding the data quality. They answered that multiple views and coordinated interactions were useful for exploring noisy parallel corpora. They also mentioned that adjusting weights and showing information of a ruleset helped them

identify noise patterns. In addition, most participants (E1, E2, and S1–S4) expected that our back-translation technique in Text Compare View would reduce the time-cost for the inspection of noisy parallel corpora when handling illiterate language data. Furthermore, they said our system would be beneficial in the field. E1, E2, and E4 asked about a plan to deploy our system as a real-world application.

The participants also provided suggestions for further enhancement of our system. For example, E3 and E4 suggested adding and hiding metrics for customization. They wanted to see how noise data affects other metrics such as ROGUE and Perplexity, practically used in their actual field of work. Non-domain experts (S1, S2, S3) said it was challenging to use adjustable weights in Ranking view and Ruleset Relationship View without a detailed explanation of their purpose. Lastly, most participants commented on improving the usability of Text Compare View; they argued that the relatively small size of the view makes it hard to use. They propose to add "expand" function that can dynamically increase its size for better perception since the view is the most frequently used while inspecting noisy parallel corpora.

7 Discussion and Future Work

Extensibility As suggested in the feedback from domain experts, it can be helpful to add additional metrics. In our system, nine metrics are encoded in distinct colors. Adding more metrics could hinder users from effectively distinguishing colors [14]. Additionally, the scroll interaction would be required as more columns are added in Ranking View. Thus, we plan to provide a customization option so that users can select a small number of metrics of their interest to configure a layout accordingly.

Reliability of pre-trained model The universal sentence encoder, which we used for sentence embedding in the preprocessing step, is widely known for its good performance [5]. However, if the encoder has not learned a specific word or character in advance, the embedded vector may not have semantic meaning. Also, Google Translation used in our back-translation may mistranslate a sentence. Therefore, users should be aware of such reliability limitations.

Scalability Generally, NMT training requires a very large size of data, but our system may cause low latency while handling a huge amount of data due to the performance issue; especially when users brush subset from PCP in Distribution View and adjust weight in Ranking View. Although we tested our system was able to cover 100,000 size of parallel corpora, we should consider improving the performance of our system by serving back-end server for updating Ranking View.

8 Conclusion

We propose VANT, an interactive visual analytic system that assists users in exploring NMT data for detecting noise and refining parallel corpora. We derived various quality metrics based on machine learning techniques. The user study demonstrated its usefulness and effectiveness by showing that users can readily investigate and filter noisy sentence pairs within the corpora. We anticipate that users will be able to improve the quality of parallel corpora with our system and achieve a better performance of their own NMT model. The implementation of our system is available at `https://vant-web.github.io/demo/`.

Acknowledgments

The authors wish to thank Soyoung Eom for heartful support. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C208906213) , and in part by Samsung Electronics.

References

[1] F. Bane and A. Zaretskaya. Selecting the best data filtering method for nmt training. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pp. 89–97, 2021.

[2] 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*, pp. 65–72, 2005.

[3] J. Bok, B. Kim, and J. Seo. Augmenting parallel coordinates plots with color-coded stacked histograms. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[4] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175. Citeseer, 1994.

[5] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder, 2018.

[6] V. Dibia and Ç. Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE computer graphics and applications*, 39(5):33–46, 2019.

[7] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics*, 19(12):2277–2286, 2013.

[8] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3363–3372, 2011.

[9] H. Khayrallah and P. Koehn. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*, 2018.

[10] Y. Kim, H. Jeon, Y.-H. Kim, Y. Ki, H. Song, and J. Seo. Visualization support for multi-criteria decision making in software issue propagation. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pp. 81–85. IEEE, 2021.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177–180, 2007.

[12] P. Koehn, H. Khayrallah, K. Heafield, and M. L. Forcada. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 726–739. Association for Computational Linguistics, Belgium, Brussels, Oct. 2018. doi: 10.18653/v1/W18-6453

[13] T. Munz, D. Väth, P. Kuznecov, T. Vu, and D. Weiskopf. Visual-interactive neural machine translation. In *Graphics Interface 2021*, 2021.

[14] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[15] M. Neves, A. J. Yepes, and A. Névéol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2942–2948, 2016.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[17] J. Park, J.-P. Hong, and J.-W. Cha. Korean language resources for everyone. In *Proceedings of the 30th Pacific Asia conference on language, information and computation: Oral Papers*, pp. 49–58, 2016.

[18] M. Rikters, M. Fishel, and O. Bojar. Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(1):39, 2017.

[19] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

[20] G. Xu, Y. Ko, and J. Seo. Improving neural machine translation by filtering synthetic parallel data. *Entropy*, 21(12):1213, 2019.

[21] B. Zhang, A. Nagesh, and K. Knight. Parallel corpus filtering via pre-trained language models. *arXiv preprint arXiv:2005.06166*, 2020.