SRT's Neural Machine Translation System for WMT22 Biomedical Translation Task

Yoonjung Choi, Jiho Shin, Yonghyun Ryu, Sangha Kim
Samsung Research, Seoul, Republic of Korea
{yj0807.choi, jiho21.shin,
yonghyun.ryu,sangha01.kim}@samsung.com

Abstract

This paper describes the Samsung Research's Translation system (SRT) submitted to the WMT22 biomedical translation task in two language directions: English to Spanish and Spanish to English. To improve the overall quality, we adopt the deep transformer architecture and employ the back-translation strategy for monolingual corpus. One of the issues in the domain translation is to translate domain-specific terminologies well. To address this issue, we apply the soft-constrained terminology translation based on biomedical terminology dictionaries. In this paper, we provide the performance of our system with WMT20 and WMT21 biomedical testsets. Compared to the best model in WMT20 and WMT21, our system shows equal or better performance. According to the official evaluation results in terms of BLEU scores, our systems get the highest scores in both directions.

1 Introduction

Neural Machine Translation (NMT) has shown rapid growth with an encoder-decoder framework, especially Transformer (Vaswani et al., 2017), in recent years. Most of the research focuses on general-purpose translation models since there are a lot of parallel data available. On the other hand, domain-specific translation, which lacks relatively high-quality parallel corpus available, is one of the challenges that need to be solved in the NMT task. To address this issue, there have been several approaches such as finetuning general-purpose models with in-domain data and utilizing in-domain monolingual corpus through back-translation (Yeganova et al., 2021).

In the domain translation, one of the issues is the terminology translation. In the case of domain-specific terms, translation results are often poor because they are relatively infrequent. Yeganova et al. (2021) also mentioned that some domain-specific terms including abbreviations were not translated correctly in previous shared tasks. Moreover, when new terms are introduced such as COVID-19, it is difficult to obtain the correct translation results as they are not in the training data. To handle this issue, we adopt the soft-constrained terminology translation proposed by Molchanov et al. (2021), which provides the terminology constraints of the target language as input to our system with source sentences like a *hint*. These terminology constraints can be obtained from in-domain dictionaries.

In addition, as many domain translation studies, the back-translation strategy (Sennrich et al., 2016) is applied to generate synthetic parallel data from in-domain monolingual corpus. To improve the overall performance of our system, we also employ the Deep Transformer architecture (Bapna et al., 2018) and the ensemble strategy (Sutskever et al., 2014). Moreover, to find better translation results, noisy channel modeling (Yee et al., 2019) and discriminative reranking (Lee et al., 2021) are attempted. Our experiment shows that deep transformer and data augmentation by the back-translation strategy improve the overall performance while the performance is not improved with reranking methods.

The rest of this paper is organized as follows. Section 2 describes the training and test data used in our system, and Section 3 explains our systems including deep transformer and soft-constrained terminology translation. Section 4 describes the details of our training and experimental results of our system; Section 5 presents the official evaluation results. Section 6 is the conclusion of our work.

2 Data

In this section, we present general-domain (out-of-domain) corpus, in-domain corpus, and in-domain terminology dictionaries used as the training data

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 901–907
Abu Dhabi, December 7–8, 2022.

	En2Es	Es2En
General-domain Parallel Corpus	518M	
In-domain Parallel Corpus	3.47M	
In-domain Target-side Monolingual Corpus	2.5M	13.9M
In-domain Dictionaries	132K	
Validataion Data	4,520	
Test Data	921	897

Table 1: Data statistics of the training data, validation data, and test data used in our system.

in our system. For training, all data are tokenized by SentencePiece (Kudo and Richardson, 2018); the vocab size is 32K for each lanaugage. The validation and test data are also described in this section. The statistics of our data are listed in Table 1.

2.1 General-Domain Parallel Corpus

We collect general-domain parallel corpus for English-Spanish from several sources. Some are from WMT News translation task. The data list is as follows: ParaCrawl[1], CommonCrawl[2], Europarl[3], News Commentary[4], and Tatoeba[5].

We also consider two datasets that are provided by organizers: United Nations (UN) Parallel Corpus[6] and UFAL Medical Corpus[7]. The UN Corpus consists of official records and other parliamentary documents of the UN that are in the public domain. In UFAL Medical corpus, it contains not only medical-domain data but also general-domain data; we consider the general-domain data of UFAL as a general-domain parallel corpus in our system.

2.2 In-Domain Parallel Corpus

We use the in-domain data provided by the WMT22 biomedical task organizers.

- Medline Corpus: It contains titles and abstracts of scientific publications. They provide three groups of English-Spanish parallel data: WMT16, WMT19, and WMT22. In WMT16 and WMT19 data, all sentence pairs are already aligned, so we use them without preprocessing process. However, in WMT22 data, all sentences of one abstract are written in one line; thus, after splitting sentences with the sentence splitter provided by Moses[8], only data that matched the number of sentences in both languages are considered as in-domain parallel corpus.

- UFAL Medical Corpus: As we mentioned in Section 2.1, it consists of a general-domain and medical-domain data. The parallel data tagged as the medical-domain are considered in-domain parallel data.

- MeSpEn Corpus: It is the resource for English-Spanish Medical Machine Translation and Terminologies (Villegas et al., 2018). It provides several biomedical and clinical literature data such as IBECS, SciELO, and Pubmed. This corpus contains titles and abstracts from several records. Since all sentences of each abstract are written in one line such as WMT22 Medline corpus, we conduct the same process to extract the parallel corpus.

2.3 In-Domain Monolingual Corpus

In the in-domain parallel corpus, some data are excluded because the number of sentences is not matched between two languages as we menteiond in Section 2.2. In this paper, we use this excluded data as in-domain monolingual data.

Moreover, for the English monolingual corpus, we extract only English data from other language pairs' dataset in Medline corpus and UFAL Medical corpus.

2.4 In-Domain Terminology Dictionary

As we mentioned in Section 1, it is important to translate domain-specific terminologies well in the domain translation. So, we also collect in-domain

[1]https://paracrawl.eu/

[2]https://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz

[3]https://www.statmt.org/wmt13/training-parallel-europarl-v7.tgz

[4]https://www.statmt.org/wmt13/training-parallel-nc-v8.tgz

[5]https://tatoeba.org/en/downloads

[6]https://conferences.unite.un.org/UNCorpus

[7]https://ufal.mff.cuni.cz/ufal_medical_corpus

[8]https://github.com/moses-smt

terminology dictionaries from MeSpEn Glossaries[9] and ClinSpEn-CT[10]. Both are translated by professional medical translators. MeSpEn glossaries contain 125,645 English-Spanish term pairs and CinSpEn-CT sample set includes 7,000 term pairs. We not only utilize in-domain terminology dictionaries as the training data but also use them in the soft-constrained terminology translation.

When the dictionary data is used as the training data, all data in dictionaries is used as it is. However, for the soft-constrained terminology translation, data refinement is required since there are redundant data. It will be described in detail in Section 3.3.

2.5 Vadlidation data

For the validation data, we use the Khresmoi development data. WMT17, WMT18, and WMT19 testset are also used as the validation data.

2.6 Test data

We consider WMT20 and WMT21 "OK" aligned testset as the test data in our system to evaluate the translation quality for the final submission.

3 System Overview

In this section, we describe our system which is based on Transformer architecture (Vaswani et al., 2017). The training details are described in Section 4.1.

3.1 Deep Transformer

Peters et al. (2018) have shown that deeper layers could efficiently extract syntactic and semantic information that could improve the overall performance. Bapna et al. (2018) also have explored deeper encoders for Transformer to improve the translation quality. Several teams that participated in the biomedical shared task last year (Yang et al., 2021; Wang et al., 2021b) have adopted the deep transformer, especially deeper encoders. In this paper, we also adopt the deep transformer architecture which contains 30 encoder layers and 6 decoder layers based on TRANSFORMER-BIG setting (Vaswani et al., 2017).

3.2 Data Augmentation

To augment the in-domain parallel corpus, we adopt back-translation (Sennrich et al., 2016), where the synthetic parallel corpus is generated by translating target-side monolingual data into the source language. Back-translation is one of the effective methods to utilize monolingual data.

In this paper, we first train base models of each direction with the combination of general-domain and in-domain parallel corpus; then, we utilize these trained models to generate source-side sentences from target-side monolingual data.

Moreover, Wang et al. (2021a) present that the overall performance is improved when the in-domain dictionaries are appended to the training corpus. We also consider in-domain terminology dictionaries as the training data.

3.3 Soft-Constrained Terminology Translation

The common approach for the terminology translation is constrained decoding (Hokamp and Liu, 2017), where the translation results are forced to contain pre-specified subsequences, such as the terminology, at decoding time. Since it is the hard-constrained method, it can aggravate the translation quality. Moreover, constrained decoding methods increase the complexity of the decoding process. To address these problems, Dinu et al. (2019) and Molchanov et al. (2021) propose the soft-constrained methods, where pre-specified terminologies are given as input with the source sentence. Although there is no guarantee that translation results always contain these pre-specified terminologies, it can learn a copy behavior at training time without compromising the overall performance.

In this paper, we adopt the soft-constrained strategy of Molchanov et al. (2021) for the terminology translation; that is, we add the desired translation result of the terminology as input with special tokens such as <*term_start*>, <*term_end*>, and <*term_trans*>. Figure 1 presents the example of the revised source sentence including the desired translation result with special tokens. For this, the training corpus should be revised to reflect this input format. First, $N\%$[11] sentence pairs of the training data are randomly extracted and both source and target sentences are tokenized by SpaCy[12] which not only supports tokenization but also provides neural network models for part-of-speech tagging. To obtain the word alignment information between

[9] https://github.com/PlanTL-GOB-ES/MeSpEn_Glossaries

[10] https://zenodo.org/record/6497373#.YxHGtXZBz-j

[11] This is a heuristic value. In this paper, we set it to 15.

[12] https://github.com/explosion/spaCy

Source sentence: Patient had a MI or CVA in last year, or has unstable cardiovascular disease.
Terminology in the source sentence: MI
Desired translation result: IM

New source sentence: Patient had a <*term_start*> *MI* <*term_end*> *IM* <*term_trans*> or CVA in last year, or has unstable cardiovascular disease.

Figure 1: Example of the revised source sentence for the soft-constrained terminology translation

source and target sentences, the word-aligner[13] is applied. Among aligned words, we only consider *Nouns* as candidates of pre-specified terminologies. In each sentence pair, up to three[14] candidates are randomly selected to provide the desired translation result. Finally, the source sentence is revised by adding a subsequence of the target sentence that is aligned to the selected candidate of the source sentence with special tokens.

For the inference of test data, the biomedical terminology dictionaries described in Section 2.4 are utilized to provide pre-specified terminology information. As we mentioned, terminology dictionaries should be refined. We first remove duplicate terminologies; for instance, if one terminology in the source language is matched with multiple terminologies in the target language, it should be removed since we don't know which of them is the desired translation result. Moreover, if the frequency of the terminology is high in general-domain data, we don't need to consider it. Thus, dictionaries are filtered based on the frequency in general-domain data. For test data, the desired translation results which are from refined dictionaries are added to each source sentence for up to three terminologies, such as the training corpus. If the source sentence in test data doesn't contain any term which is in refined dictionaries, we just input the original source sentence.

3.4 Ensemble

From several NMT studies (Sutskever et al., 2014; Garmash and Monz, 2016; Firat et al., 2016), it has been already shown that ensembling methods can improve the overall performance. In this paper, we conduct the ensemble strategy with the top three models based on our testset for the final submission.

[13] eflomal, https://github.com/robertostling/eflomal

[14] This is a heuristic value. Based on our training data, we decide this value.

3.5 Reranker

The current NMT system utilizes the beam search approach to generate the final translation result. However, since it is the auto-regressive model, it considers only a limited target context to get the probability of a target token. To address this issue, there are several reranking methods that generate several different hypotheses from the NMT model and rerank them. Since reranking models can consider the entire target context, it can improve the overall performance over the beam search (Lee et al., 2021).

In this paper, we adopt two reranking methods: noisy channel modeling (Yee et al., 2019) and discriminative reranking (Lee et al., 2021). Noisy channel modeling is based on Bayes' rule; it generates translation results based on a backward model and a pre-trained target-side language model. We use a translation model in the opposite direction as a backward model and train transformer language models for the target-side language model. The discriminative reranking model is a transformer architecture that takes the source sentence and the n-best list of output hypotheses as input. It also includes position embeddings and language embeddings for representing two different languages' inputs. As in Lee et al. (2021)'s work, we use XLM-R (Conneau et al., 2020) which is a transformer-based multilingual masked language model as the pre-trained model.

4 Experiments

In this section, we present training details and experimental results of our systems.

4.1 Training details

The baseline models are trained based on TRANSFORMER-BIG setting (Vaswani et al., 2017) which contains 6 encoder layers. We first train baseline models with only general-domain corpus and incrementally train them using in-domain parallel corpus to confirm the effectiveness of in-domain

System	Data	En2Es		Es2En	
		WMT20	WMT21	WMT20	WMT21
Best Offiical 20 (Bawden et al., 2020)		0.4672		0.5075	
Best Official 21 (Yeganova et al., 2021)			0.5117		**0.5382**
Baseline	GD	0.4761	0.5134	0.4952	0.5148
	GD+ID	0.4956	0.5305	0.5060	0.5183
Deep Transformer	GD+ID	**0.5174**	0.5485	0.5186	0.5360
+ Data Augmentation	GD+ID+BT+IND	0.5151	0.5523	0.5236	0.5346
+ Ensemble	GD+ID+BT+IND	0.5169	**0.5524**	**0.5255**	0.5332
+ SC Terminology Translation	GD+ID+BT+IND	0.5158	0.5450	0.5216	0.5362
+ Noisy Channel Modeling	GD+ID+BT+IND	0.5143	0.5454	0.5110	0.5255
+ Discriminative Reranking	GD+ID+BT+IND	0.5159	0.5481	-	-

Table 2: BLEU scores on the WMT20 and WMT21 OK aligned test set.

corpus. The deep transformer models which contain 30 encoder layers are trained with the combination of the general-domain and in-domain parallel corpus; based on them, the synthetic data are generated from in-domain monolingual data. Finally, we train the deep transformer models on all corpus: general-domain (GD) and in-domain (IN) parallel corpus, synthetic data (BT), and in-domain dictionary (IND) information. The soft-constrained (SC) terminology translation models are also trained based on deep transformer models with revised training corpus described in Section 3.3. In addition, the ensemble strategy and reranking methods explained in Section 3.5 are applied. For the implementation, we use Fairseq[15], and all models are trained using 8 A100 GPUs. Adam optimizer is used. The batch size is 4K tokens, and the frequency of parameter update is 20. The learning rate, the dropout, and the label smoothing are set to 0.0007, 0.1, and 0.1, respectively. For the inference, the beam size is set to 8. The BLEU scores are calculated using the mt-eval script from Moses (Koehn et al., 2007).

4.2 Experimental results

The experimental results of English to Spanish (En2Es) and Spanish to English (Es2En) directions are shown in Table 2. The baseline models show that the in-domain corpus improves the overall performance in the domain translation. We then apply the deep transformer with the general-domain and in-domain data and it achieves a significant improvement over baseline models. With data augmentation by back-translation of monolingual in-domain data and in-domain dictionaries, there is a slight improvement on average; even though the performance drops slightly in the WMT20 testset of En2Es and WMT21 testset of Es2En, it improves more in other testset of each direction.

The ensemble models show better performance than a single model in general.

In the soft-constrained terminology translation, the performance is slightly improved in one testset while the performance is decreased in the other testset in each direction. Since the soft-constrained terminology translation models are trained with revised corpus, the testset also should be revised by adding desired translation results with special tokens in order to evaluate the performance accurately. Table 3 shows BLEU scores of soft-constrained terminology translation models on plain testsets and revised testsets which contain desired translation results. We observe that soft-constrained terminology translation models are more effective when the desired translation results of some terminologies are given such as training corpus.

As we mentioned in Section 3.5, two reranking methods are adopted, but as a result, the overall

En2Es	WMT20	WMT21
Plain Testset	0.5158	0.5450
Revised Testset	**0.5325**	**0.5505**
Es2En	**WMT20**	**WMT21**
Plain Testset	0.5216	0.5362
Revised Testset	**0.5294**	**0.5472**

Table 3: BLEU scores of soft-constrained terminology translation models on plain testsets and revised testsets with soft-constrained terminologies.

[15] https://github.com/facebookresearch/fairseq

System	En2Es	Es2En
Best Official	0.5235	0.6045
SRT run1	0.5214	0.5954
SRT run2	0.5196	0.5943
SRT run3	**0.5235**	**0.6045**

Table 4: Official BLEU scores of our submissions for WMT22 biomedical task.

performance is not improved. (The discriminative reranking is experimented only on En2Es.)

Since there is no improvement with two reranking methods, we exclude their results in our final submissions. Our final submissions are results of data augmentation, ensembling models, and soft-constrained terminology translation.

5 Official Evaluation Results

The official evaluation results of our submissions (SRT) for WMT 2022 biomedical translation task are shown in Table 4. All our submissions show the best BLEU scores.

6 Conclusion

This paper presents the Samsung Research's Translation system (SRT) for the WMT22 biomedical translation shared task in two language directions: English to Spanish and Spanish to English. We perform experiments with several strategies such as deep transformer, data augmentation, soft-constrained terminology translation, ensembling models, and reranking methods. Our experiments show the effectiveness of each strategy. The deep transformer, data augmentation, and ensemble strategies improve effectively the overall performance in the domain translation. Moreover, we present that the soft-constrained terminology translation is a reasonable method to achieve good performance in the domain translation. Our systems show the best BLEU scores in the official evaluation results.

References

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033, Brussels, Belgium. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264, Online. Association for Computational Linguistics.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. PROMT systems for WMT21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *LREC MultilingualBIO: Multilingual Biomedical Text Processing. ELRA*.

Weixuan Wang, Wei Peng, Xupeng Meng, and Qun Liu. 2021a. Huawei AARC's submissions to the WMT21 biomedical translation task: Domain adaption from a practical perspective. In *Proceedings of the Sixth Conference on Machine Translation*, pages 868–873, Online. Association for Computational Linguistics.

Xing Wang, Zhaopeng Tu, and Shuming Shi. 2021b. Tencent ai lab machine translation systems for the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 874–878, Online. Association for Computational Linguistics.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, chuanfei xu, Min Zhang, and Ying Qin. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884, Online. Association for Computational Linguistics.

Kyra Yee, Yann N. Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5695–5700. Association for Computational Linguistics.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.