Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation

Xabier Soto, Olatz Perez-de-Viñaspre, Gorka Labaka, Maite Oronoz
HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country UPV/EHU
{xabier.soto, olatz.perezdevinaspre, gorka.labaka, maite.oronoz }@ehu.eus

Abstract

In this paper we describe the systems developed at Ixa for our participation in WMT20 Biomedical shared task in three language pairs, en-eu, en-es and es-en. When defining our approach, we have put the focus on making an efficient use of corpora recently compiled for training Machine Translation (MT) systems to translate Covid-19 related text, as well as reusing previously compiled corpora and developed systems for biomedical or clinical domain. Regarding the techniques used, we base on the findings from our previous works for translating clinical texts into Basque, making use of clinical terminology for adapting the MT systems to the clinical domain. However, after manually inspecting some of the outputs generated by our systems, for most of the submissions we end up using the system trained only with the basic corpus, since the systems including the clinical terminologies generated outputs shorter in length than the corresponding references. Thus, we present simple baselines for translating abstracts between English and Spanish (en/es); while for translating abstracts and terms from English into Basque (en-eu), we concatenate the best en-es system for each kind of text with our es-eu system. We present automatic evaluation results in terms of BLEU scores, and analyse the effect of including clinical terminology on the average sentence length of the generated outputs. Following the recent recommendations for a responsible use of GPUs for NLP research, we include an estimation of the generated CO_2 emissions, based on the power consumed for training the MT systems.

1 Introduction

The WMT20 Biomedical shared task calls for developing systems for translating biomedical abstracts and terminologies between several languages. In our case, we participate in the task of translating biomedical terms and abstracts from English into Basque (en-eu), as well as translating biomedical abstracts between English and Spanish (en-es and es-en). For translating the test data from English into Basque, we concatenate our best en-es system with our es-eu system, both for translating abstracts and terminologies.

2 Related work

For translating biomedical texts from English into Catalan, Costa-jussá et al. (2018) use a pivoting or cascade approach, translating the texts first from English into Spanish (en-es), and then from Spanish into Catalan (es-ca). This technique is useful when there are more bilingual in-domain sentences for each of the language pairs (en/es and es/ca) than for the desired source and target languages (en/ca). Since there are low resources for en/eu biomedical domain, but we have access to many resources for en/es and es/eu in the biomedical or clinical domain, we follow the same approach for translating the test sets from English into Basque (en-eu).

Since most of the available in-domain corpus is monolingual, we also make use of traditional back-translation and forward translation techniques (Sennrich et al., 2016).

In our previous work for translating clinical texts between Basque and Spanish, we showed that including clinical terminologies directly into the training corpus was useful for domain adaptation when no bilingual in-domain sentences were available (Soto et al., 2019a). As clinical terminologies, we refer to the automatic translation into Basque of SNOMED CT (IHTSDO, 2014), which is considered the most comprehensive, multilingual clinical health care terminology collection in the world. In this work, we extend the number of clinical terminologies as part of the ongoing translation of SNOMED CT into Basque (Perez-de-Viñaspre,

Proceedings of the 5th Conference on Machine Translation (WMT), pages 875–880

2017), and include the provided ICD-10 resources plus other smaller terminology collections recently created for translating Covid-19 related texts.

3 Resources

For training our baseline en/es systems, we make use of the Medline corpus provided by the organisers of the WMT20 Biomedical shared task, as well as the recently compiled TAUS Corona Crisis Corpus.[1]

For backtranslation (es-en) and forward translation (en-es), we use the English corpus prepared by Sketch Engine[2], based on the Covid-19 related corpus compiled for a recent Kaggle competition (Wang et al., 2020).

As a final step, we include several clinical terminologies: 1) the ICD-10 (en-eu) corpus provided by the organisers of the WMT20 Biomedical shared task, adding the corresponding Spanish counterparts; 2) terms obtained from the automatic translation into Basque of SNOMED CT (Perez-de-Viñaspre, 2017), including terms up to 11 tokens; 3) a recent SNOMED CT interim release of Covid-19 related terms[3], manually translated into Basque by a translator of the Basque public health service (Osakidetza); and 4) a collection of Covid-19 related terms recently compiled by Elhuyar[4], including all the terms published until June 18[5].

For training our es-eu system, we use the aforementioned terminologies together with an out-of-domain corpus formed mainly by news (Etchegoyhen et al., 2016), previously applying a language identification tool[6] to exclude sentences where most of the terms are named entities like locations or person names. Doing this, a bigger part of the vocabulary can be used to translate biomedical or clinical terms. Furthermore, as in-domain corpus we use clinical notes in Spanish coming from the hospital of Galdakao-Usansolo for forward translation and copying (Currey et al., 2017). This corpus was compiled between 2008 and 2012.[7]

For the evaluation of en/es systems, we use Khresmoi;[8] while for es-eu we use templates of clinical notes in Basque written in the Donostia hospital (Joanes Etxeberri Saria V. Edizioa, 2014), together with their manual translations into Spanish made by a bilingual doctor.

Table 1 presents the description and statistics of our corpora.

	Description	Sentences
en/es	Medline (WMT Biomedical)	388,068
	TAUS Corona Crisis Corpus	902,133
	Sketch Engine Covid-19 (en)	4,671,609
	ICD-10 (WMT Biomedical)	27,696
	SNOMED CT corpus	385,800
	SNOMED CT Covid-19 corpus	84
	Elhuyar Covid-19 corpus	113
	Khresmoi (dev set)	500
	Khresmoi (test set)	1,000
es-eu	out-of-domain	3,703,757
	in-domain (es)	2,023,811
	ICD-10 (WMT Biomedical)	27,696
	SNOMED CT corpus	896,898
	SNOMED CT Covid-19 corpus	84
	Elhuyar Covid-19 corpus	126
	Donostia hospital (dev set)	1,038
	Donostia hospital (test set)	1,038

Table 1: Description and statistics of the used corpora.

4 Systems

For en/es we develop 3 systems: 1) using only the bilingual in-domain corpus (Medline + TAUS Corona Crisis Corpus), 2) including the Sketch Engine Covid-19 (en) corpus for backtranslation (es-en) or forward translation (en-es), and 3) adding all the clinical terminologies from ICD-10, SNOMED CT and Elhuyar.

For es-eu we train a unique system using the out-of-domain corpus and the clinical terminologies, as well as the in-domain (es) corpus for forward translation and copying.

For training the backtranslation (en-es) and forward translation (es-en) systems, we used the bilingual in-domain corpus (Medline + TAUS Corona Crisis Corpus); while for es-eu we used the out-of-domain corpus and a reduced set of SNOMED CT terminologies, as used in Soto et al. (2019b).

[1] https://md.taus.net/corona

[2] https://www.sketchengine.eu/covid19/

[3] http://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release%2DCOVID-19

[4] https://www.elhuyar.eus/eu/site/prentsa-aretoa/368/covid-19-gaitzaren-inguruko-terminologia%2Dgure-hiztegietako-azkenaldaketak

[5] when the English term was missing, if there was no doubt about how to translate it, the first author manually translated it; while if there wasn't a clear translation into English or the term was more related to socioeconomics than biomedical domain, it wasn't included in the en/es corpus.

[6] https://github.com/saffsd/langid.py

[7] Due to privacy requirements, this corpus is not publicly available. Prior to use, it was de-identified by reordering sentences, and only authors who had previously signed a non-disclosure commitment had access to it.

[8] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122

All the systems are Transformer (Vaswani et al., 2017) models trained with OpenNMT (Klein et al., 2017), using the recommended hyperparameters.[9] When necessary, we halved the batch-size so that it could fit in 2 GPUs, and accordingly doubled the value for gradient accumulation.

We applied joint BPE-dropout (Provilkov et al., 2020), with 32,000 merge operations for en/es and 90,000 for es-eu.

5 Results

Table 2 shows the BLEU scores of our systems on the validation (dev) and test sets presented in Table 1, together with previously published (es-eu) results for comparison.

Lang.	System	dev	test
es-en	Baseline (Medline + TAUS)	56.57	52.55
	Baseline + backtranslation (bt)	**61.60**	**57.25**
	Baseline + bt + terminologies	60.95	56.89
en-es	Baseline (Medline + TAUS)	48.02	46.30
	Baseline + forward translation (ft)	**50.20**	**47.19**
	Baseline + ft + terminologies	49.92	47.15
es-eu	Soto et al. (2019a)	11.30	**12.04**
	Soto et al. (2019b)	**11.85**	11.24
	This work	6.21	5.15

Table 2: BLEU scores for systems developed for es-en, en-es and es-eu translation directions (Lang.).

As expected, backtranslation significantly improves the es-en results (around 5 BLEU points); while the gains obtained with forward translation (en-es) are smaller (around 2 BLEU points in the dev set and around 1 BLEU point in the test set). However, we observe a slight decrease on BLEU values when including the clinical terminologies on the training corpus for both es-en and en-es systems. For further analysing this, we calculate the average sentence length of the different evaluation corpora as translated by the different systems. Table 3 shows the average sentence length of the validation (dev) and test sets after being translated by each of the es-en and en-es systems. As a reference, the average sentence length of the original dev and test sets are 22.70 (es) / 21.06 (en) and 24.03 (es) / 21.91 (en).

We observe that, except for the dev set translated by the en-es systems, the lower sentence length is always obtained when using the system including the clinical terminologies. This is confirmed by a fast check of the outputs generated when translating

Lang.	System	dev	test
es-en	Baseline (Medline + TAUS)	20.54	22.02
	Baseline + backtranslation (bt)	20.56	21.73
	Baseline + bt + terminologies	20.40	21.56
en-es	Baseline (Medline + TAUS)	22.75	23.87
	Baseline + forward translation (ft)	22.93	23.84
	Baseline + ft + terminologies	22.99	23.76

Table 3: Average sentence length of the different evaluation corpora as translated by the systems developed for es-en and en-es translation directions (Lang.).

the official test sets provided by the organisers, where we see that the sentences translated by these systems usually end before having translated all of the terms that appear in the input. Overall, the sentence lengths of the generated translations are closer to the original sentence lengths when using the baseline systems; therefore, for en-es and es-en we submit as best systems the translations produced by the baseline systems, using only Medline and TAUS corpora.

Regarding es-eu, in Table 2 we can see a severe decrease on BLEU scores comparing to our previous works. For training the system in Soto et al. (2019a) we used the same out-of-domain corpus (without applying langid.py) and a reduced set of SNOMED CT terminologies (151,111 entries), both directly and inserted into artificial sentences; while in Soto et al. (2019b) we used this same corpus without the artificial sentences, which didn't prove to be useful. Nevertheless, after manually checking the outputs generated by these 3 systems, we observe that the system developed for this work performs generally better, so we submit the translations produced by this system.[10] As we use a cascade approach for en-eu, we use the en-es system including the terminologies for translating abstracts; and the baseline system for translating terminologies, as these were the best performing systems on a fast human evaluation. [11]

Once we have selected the best performing systems for each of the language pairs, since we are allowed to submit 3 runs, in the case of en/es, for each of the developed systems we submit an ensemble of the 3 models which obtained higher BLEU

[9] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model (Accessed on July 18, 2020.)

[10] It has to be noted that the evaluation corpus used for es-eu has strong limitations, since the original sentences are written for encouraging medicine students to write correctly; while the translations into Basque made by a doctor are overall shorter, use simplified grammar, often omit verbs and punctuation, and use many acronyms.

[11] Both for en/es and en-eu systems, the translations of the first 10 sentences of the official test sets were checked; and in case of tie, the next 10 sentences were also observed.

scores in the dev set during training; while for en-eu we alternate between single and ensemble systems for each of the en-es and es-eu systems. Specifically, we submit as best system an ensemble of the baseline en-es system and a single es-eu system for translating terminologies; while we use a single en-es system including the terminologies and an ensemble es-eu system for translating abstracts.

Table 4 shows the BLEU scores obtained on the official test sets for each of the language pairs and submitted runs for translating abstracts, as provided by the organisers. We present in italics the result of the expected best system for each language pair, and in bold the highest BLEU score, as in previous tables.

Lang.	System	BLEU
es-en	Baseline (Medline + TAUS)	*40.65*
	Baseline + backtranslation (bt)	**40.71**
	Baseline + bt + terminologies	39.96
en-es	Baseline (Medline + TAUS)	***41.71***
	Baseline + forward translation (ft)	38.36
	Baseline + ft + terminologies	38.58
en-eu	single (en-es) + ensemble (es-eu)	*8.15*
	ensemble (en-es) + single (es-eu)	7.82
	ensemble (en-es) + ensemble (es-eu)	**8.84**

Table 4: BLEU scores on the official test sets for translating abstracts in es-en, en-es and en-eu translation directions (Lang.).

Comparing to the submissions made by other teams, our systems submitted for en/es obtain the lowest BLEU scores among all the participants; while for en-eu our best run is the second among the best runs of each participant, only surpassed by the three runs submitted by Elhuyar.

Finally, Table 5 presents the accuracy and BLEU scores obtained by our systems when used for translating terminologies (en-eu), as provided by the organisers.

Lang.	System	Acc.	BLEU
en-eu	single (en-es) + ensemble (es-eu)	0.12	13.14
	ensemble (en-es) + single (es-eu)	*0.08*	*7.21*
	ensemble (en-es) + ensemble (es-eu)	**0.13**	**14.81**

Table 5: Accuracy (Acc.) and BLEU scores on the official test set for translating terminologies in en-eu translation direction (Lang.).

Surprisingly, the obtained automatic scores are much lower than the ones obtained by the rest of the participants (between 0.73 and 0.78 for accuracy, and approximately 71 to 74 BLEU scores). However, the generated translations look quite sensible, so we expect the human evaluation will shed some light about the performance of our systems.

6 Measured power consumption and estimated CO_2 emissions

Following the recommendations by Strubell et al. (2019), we report the power consumed by our GPUs when training the systems developed for this work, along with the estimated CO_2 emissions. For calculating the training time, we use the time shown in the first and last lines of the log file generated while training the systems, including also the initial time for preparing the data, so the presented values constitute an upper bound of the actually consumed power. Nonetheless, we have to point out that OpenNMT makes an efficient use of the power capabilities of the GPUs, so we can say that the numbers shown here are an accurate estimation. Table 6 shows the number of GPUs, training time, power consumption and estimated CO_2 emissions for each of the developed systems. All the GPUs used for this work are Nvidia Titan Xp models with 250W power. We present the values of the different systems in the same order as in Table 2, and estimate the CO_2 emissions by applying equations (1) and (2) in Strubell et al. (2019), considering only the power consumed by our GPUs. Overall, the CO_2 emissions generated by our GPUs are approximately 329.44 lbs.

Lang.	GPUs	Time (hh:mm)	Power (kWh)	CO_2e (lbs)
es-en	4	43:19	43.33	65.31
	2	46:30	23.26	35.06
	2	45:37	22.82	34.39
en-es	4	45:09	45.16	68.07
	2	47:24	23.70	35.73
	2	47:21	23.68	35.69
es-eu	2	73:14	36.62	55.20
TOTAL				**329.44**

Table 6: Number of GPUs, training time, power consumption and estimated CO_2 emissions for each of the developed systems (same order as in Table 2).

7 Conclusion and future work

In this work, we have presented a simple proposal using previously compiled corpora from the biomedical or clinical domain, as well as clinical terminology included directly to the training corpora. Apart from calculating BLEU scores, we have also calculated the average sentence length of the generated translations for en/es systems, and observed that the systems including terminologies

performed generally worse than the baseline systems.

As future work, we plan to incorporate these clinical terminologies in a more efficient way (Dinu et al., 2019; Wang et al., 2019). For improving both training and evaluation, we'll also use bilingual clinical domain corpora being compiled now in collaboration with the Basque public health service (Osakidetza). Furthermore, since we have observed that some of the translations generated by the es-eu systems remain in Spanish, we'll study techniques to leverage in-domain monolingual data in Basque like the one provided by the organisers from Wikipedia.

Finally, we plan to keep reporting the consumed power and consequently generated CO_2 emissions, probably making use of recently developed automatic tools (Henderson et al., 2020)[12].

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045, and projects BigKnowledge (BBVA foundation grant 2018), DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE) and DOTT-HEALTH (PID2019-106942RB-C31, MCIU/AEI/FEDER, UE).

References

Marta R. Costa-jussá, Noé Casas, and Maite Melero. 2018. English-catalan neural machine translation in the biomedical domain through the cascade approach. *Computing Research Repository*, arXiv:1803.07139. Version 2.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

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.

Thierry Etchegoyhen, Andoni Azpeitia, and Naiara Pérez. 2016. Exploiting a large strongly comparable corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3523–3529, Portoroz, Slovenia.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Computing Research Repository*, arXiv:2002.05651.

International Health Terminology Standards Development Organisation IHTSDO. 2014. *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation.

Joanes Etxeberri Saria V. Edizioa. 2014. Donostia unibertsitate ospitaleko alta-txostenak. *Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Olatz Perez-de-Viñaspre. 2017. *Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque*. Ph.D. thesis, University of the Basque Country, Donostia, Spain.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. 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.

Xabier Soto, Olatz Perez-De-Viñaspre, Gorka Labaka, and Maite Oronoz. 2019a. Neural machine translation of clinical texts between long distance languages. *Journal of the American Medical Informatics Association*, 26(12):1478–1487.

Xabier Soto, Olatz Perez-De-Viñaspre, Maite Oronoz, and Gorka Labaka. 2019b. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 8–18, Dublin, Ireland.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *Computing Research Repository*, arXiv:1906.02243.

[12] https://github.com/Breakend/experiment-impact-tracker

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, Long Beach, CA.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *Computing Research Repository*, arXiv:2004.10706. Version 4.

Tao Wang, Shaohui Kuang, Deyi Xiong, and António Branco. 2019. Merging external bilingual pairs into neural machine translation. *Computing Research Repository*, arXiv:1912.00567.