This document is part of the Research and Innovation Action "European Live Translator (ELITR)".
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D4.1

Initial Report on Multi-Lingual MT

Ondřej Bojar (CUNI), Bohdan Ihnatchenko (CUNI),
Philip Williams (UEDIN), Dominik Macháček (CUNI),
Rico Sennrich (UEDIN)

Dissemination Level: Public

Final (Version 1.0), 31st December, 2019

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	Doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 36 months
Dissemination level	Public
Contractual date of delivery	31st December, 2019
Actual date of delivery	31st December, 2019
Deliverable number	D4.1
Deliverable title	Initial Report on Multi-Lingual MT
Type	Report
Status and version	Final (Version 1.0)
Number of pages	14
Contributing partners	CUNI, UEDIN
WP leader	CUNI
Author(s)	Ondřej Bojar (CUNI), Bohdan Ihnatchenko (CUNI), Philip Williams (UEDIN), Dominik Macháček (CUNI), Rico Sennrich (UEDIN)
EC project officer	Alexandru Ceausu
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic
bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary

This deliverable summarizes the progress in WP4 Multi-Lingual MT during the first year of the project. The work package consists of 5 tasks, three of which are running during the first year.

T4.1 Baseline MT Models was planned and carried out during the first 6 months of the project. It provided MT systems to the rest of the main processing pipeline, so that intergration and technical testing could start soon. More details are in Section 2.

T4.2 Document-Level Translation is a research goal somewhat more independent of the remaining tasks. The aim is to substantially improve the practice of handling document-level context across MT processing stages: training, translation and evaluation. In Section 3, we report on our progress in all three aspects: several separate means of evaluation with more or less conclusive results as well as a post-processing strategy to improve document-level coherence.

T4.3 Multi-Target Translation explores the setup most needed for ELITR central event, the EUROSAI congress where a single speech needs to be translated into up to 43 target languages. We report on our baseline massively-multilingual system and on our exploration of the trade-off between the number of languages covered in a single system vs. the loss in translation quality. These experiments proved to be more time-consuming than expected and we will continue with them also in year 2 of the project.

T4.4 Multi-Source Translation aims to improve translation quality by considering other language versions of the same content. The task is scheduled to start in year 2 and can consider both written or spoken multi-source. As preparatory steps ahead of time, we have begun gathering data from training lessons of interpreters to assess if multi-source could be applied in the ELITR setup of live conference interpretation. More details are in Section 5.

T4.5 Flexible Multi-Lingual MT is planned for year 3 of the project.

2 Task T4.1 Baseline MT Models (CUNI, UEDIN, KIT)

Using the ELITR OPUS Corpus described in Deliverable 1.1, UEDIN has trained baselines for all EU translation directions and a majority of EUROSAI translation directions in the form of a massively multilingual MT model (Aharoni et al., 2019). The dataset is 'English-centric,' meaning that all sentence pairs include English on either the source or the target side. Translation for pairs not including English is therefore zero-shot or must be pivoted through English.

We used scripts from the Moses toolkit (Koehn et al., 2007) to normalize, tokenize, and truecase the data. We used subword-nmt[1] to segment the text into subword units using the byte pair encoding (BPE) algorithm (Sennrich et al., 2016) with 40,000 merge operations. Following Johnson et al. (2017), we prepended a tag to each source sentence to indicate the target language. For instance, in an English-Czech sentence pair, the first token of the source sentence is `<2cs>`.

For training and inference we used the Marian toolkit (Junczys-Dowmunt et al., 2018). Our model is a Transformer, configured using the 'base' hyperparameters (Vaswani et al., 2017). We used a multilingual validation set containing 200 sentence pairs (100 into English and 100 out of English) for each language pair covered by the dataset. This amounts to $39 \cdot 200 = 7800$ test sentence pairs. Training was stopped when the validation set cross entropy had failed to improve for 10 consecutive validation points.

	Source						
Target	cs	de	en	es	fr	it	ru
ar	5.2	4.4	**9.2**	5.4	4.9	5.7	4.3
cs	-	11.4	**22.4**	6.9	9.6	10.4	12.0
de	12.2	-	**22.0**	10.6	9.8	9.6	9.7
en	33.5	29.0	-	**33.8**	28.2	33.4	27.2
es	11.2	15.6	**33.5**	-	20.5	20.0	14.5
fr	14.2	12.8	**25.5**	17.2	-	15.5	12.7
it	14.7	13.1	**30.2**	18.7	16.3	-	11.9
nl	14.3	15.2	**25.3**	15.0	13.1	13.3	12.3
pt	17.7	17.0	**38.5**	25.8	19.6	20.7	14.8
ru	13.4	10.7	**18.3**	11.9	12.0	11.0	-
Average	15.2	14.4	25.0	16.1	14.9	15.5	13.3

Table 1: BLEU scores for translation into the nine target languages in the News Commentary v14 test set. The seven source languages are the languages supported by ASR and intended for deployment. The results are comparable within a given target language and we highlight in bold the best result in each row. It is not a big surprise that English is generally the best source, although BLEU cannot reliably assess subtle properties such as the preservation of the gender.

We evaluated the system using the News Commentary v14 test set (described in Deliverable 1.2). Table 1 gives average BLEU scores for the 10 target languages covered by the test set when translating out of the seven ASR-supported source languages. These results serve as baselines for systems developed for deployment as part of the ELITR pipeline.

For selected language pairs, we also trained dedicated bilingual models to measure the quality difference between massively multilingual and bilingual models. Table 2 gives multilingual BLEU scores for out-of-English translation for all language pairs covered by the News Commentary v14 test set as well as bilingual scores for the dedicated models. Our baseline results show that our multilingual systems, while enabling the coverage of many translation directions, still trail behind dedicated bilingual models in terms of quality, and work on other tasks (T4.3/4.4/4.5) has begun to close this gap.

[1] `https://github.com/rsennrich/subword-nmt`

System	ar	cs	de	es	fr	it	nl	pt	ru
	Target								
Multilingual	9.2	22.4	22.0	33.5	25.5	30.2	25.3	38.5	18.3
Bilingual	16.6	29.1	27.5	-	-	-	-	-	20.7

Table 2: Bleu scores for individual target languages when translating out of English on the News Commentary v14 test set.

3 Task T4.2 Document-Level Machine Translation (CUNI, UEDIN)

Task 4.2 aims at improved handling of document-level phenomena in MT practice for evaluation, training and translation.

In Section 3.1, we describe the test suites we prepared and used in WMT19. Section 3.2 outlines the test suite for 2020 which is in preparation and finally, Section 3.3 presents UEDIN and CUNI techniques for improved translation.

3.1 WMT19 Document Level Test Suites

CUNI in cooperation with the Supreme Audit Office of the Czech Republic (SAO) processsed multiple audit reports which were published on SAO's website and other supreme audit institutions. These selected audit reports always had always several language variants. Additionally, we included one document sample from the domain of agreements in the test suite. The data was cleaned and organised in a suitable form for a WMT19 test suite. The source texts were then distributed to all WMT19 News Task participants and thus translated by MT systems participating in the shared task.

Since this test suite had two different parts, we used two different approaches for the evaluation. As for the part of audit reports, we have complemented an automatic evaluation with a manual evaluation carried out by audit experts. In the case of the sample agreement, the evaluation was fully manual.

In the extensive manual annotation of the MT outputs participating in the shared task, our annotators identified types of translation errors related to document-level translation. The results document that recent NMT systems achieved such a high level of translation quality that it becomes difficult or impossible to evaluate them on the basis of a simple comparison with a single reference translation.

On the other hand, at least one type of documents is mishandled catastrophically by current MT system, namely documents defining their own fixed terminology. A prime example are formal agreements. In the coming months, we plan to focus on this domain, constructing a new test suite for 2020 focused on this type of inputs, see below.

The details about the 2019 test suite were published in the respective WMT19 paper by Vojtěchová et al. (2019).

The test suite itself has been made available as one of public ELITR repositories at github:

`https://github.com/ELITR/wmt19-elitr-testsuite`

Another extensive annotation of document-level phenomena was carried out in Rysová et al. (2019). Here the focus was on news-style sentences from the WSJ section of the Penn Treebank (where explicit discourse annotation exists), specifically on discourse connectives and their alternative lexicalizations. Similarly to the audit domain above, the results indicated again that the current quality of MT is in general high enough so that the comparison with a single reference translation becomes non-discerning. In fact, in some cases the reference translation was scored lower because it was not adhering to wording of the source as the MT systems did. Furthermore, this evaluation did not show any benefit from the few MT systems that were trained with some cross-sentential context taken into account.

3.2 A Testsuite on Agreements in Preparation

In the new test suite, we will focus on several types of agreements, namely lease and sublease agreements and purchase contracts of cars and real estate. We have collected 30 Czech examples in those four categories. We will try to get a comparable number of English agreements, too, but so far we were unsuccessful.

Given the experience with the WMT19 SAO test suite, where a single reference translation prepared beforehand proved insufficient because it did not account for multiple correct translations of domain-specific terminology, we did not plan to provide reference translations for the agreements. Instead, we wanted to mimic our strategy used for the sublease agreement in the 2019 test suite, i.e. to identify "markables" in the source documents and, once the candidate translations are collected, check if the translations of these "markables" are correct and consistent within the given candidate.

However, when we tried this approach on the 30 new agreements, we realised that it will not be possible due to an overhelming number of specific terms, which are furthermore different for each agreement category. In other words, it is not possible to identify "markables" before knowing at least partially the set of candidate translations.

Therefore, our current plan is to use the strategy briefly described in Section 4.3 of Popel et al. (2019), i.e. to automatically list source terms that have multiple target-side counterparts across and within individual candidate translations and manually validate which of these translations are (a) acceptable on their own, and (b) in accordance with other lexical choices within the given candidate translation.

This approach starts with collecting a number of candidate translation so we will definitely submit our new test suite to WMT20. Note that up until now, we have source documents in Czech, so we need MT systems participating in translation from Czech to English. We double checked with WMT organizers that this direction will not be omitted as it was in WMT19.

3.3 Better Document-Level Evaluation and Translation

UEDIN is investigating better models for document-level MT, and their automatic evaluation. Voita et al. (2019b), published at ACL, makes contributions in both aspects. In terms of modeling, we propose a two-pass translation process where a first-pass model, trained on sentence-level parallel data, produces a baseline translation, which a context-aware second-pass model then refines. This two-pass strategy has the advantage of allowing training with a mix of sentence-level and document-level training data. For evaluation, we have produced test sets for contrastive evaluation, similar to those by Bawden et al. (2018), to target specific translation phenomena that require context. These novel test sets are larger-scale, and cover more translation phenomena (namely deixis, ellipsis, and lexical consistency) than that by Bawden et al. (2018). We find that targeted test sets are very useful for development, allowing to measure the impact of design decisions that may have little impact on generic metrics such as BLEU, but affect how effectively the model learns to take context into account.

Our most recent work, Voita et al. (2019a), focuses on the challenging case where there is no document-level parallel data, and all the available document-level data is monolingual. We show that consistency in translation can be improved with a monolingual repair model, essentially a model that performs automatic post-editing purely on the basis of the primary system's translation output. Such a setup is attractive because the monolingual repair system can be trained without document-level parallel data, but it also has advantages from a deployment perspective, since it allows for some independence between the development of the (sentence-level) main translation system, which may be multilingual, and language-specific monolingual repair modules to improve document-level consistency.

CUNI has experimented with improving document-level coherence by translating a windows of subsequent techniques. The system was submitted to the WMT19 news translation task, see Popel et al. (2019).[2] Based on overall scores, no clear benefit of this style of training is apparent

[2]Note that this particular publication received support from other grants, not ELITR.

but a more targeted evaluation is yet to be performed.

4 Task T4.3 Multi-Target MT (CUNI, UEDIN, KIT)

Task 4.3 was proposed to reduce primarily the computational costs of training MT models for the highly multilingual setting needed to support EUROSAI Congress, translating from 7 source languages up to 43 target languages.

With multi-lingual models, described in this section, we also benefit from the GPU parallelism and translate the given input sentence to many targets at once, in one GPU batch. We talk about "rainbow translation models" and adjust the integration pipeline to handle them well. The details of this integration are not the focus of this deliverable and we thus omit them.

While Task T4.3 was originally planned only for year 1 of the project, the experiments, esp. those described in Section 4.2 proved more time (and resource) consuming than expected. We will thus continue the work on finding the best balance of languages in multi-target models also during year 2.

4.1 Massively Multi-Lingual Model

UEDIN has trained baseline multi-target machine translation models to cover all EU and EUROSAI translation directions (see T4.1). These massively multilingual baseline models exhibit a quality drop in translation quality compared to dedicated bilingual machine translation models – on a selection of 4 language pairs (EN→{DE,ZH,BR,TE}), the average drop is 2.9 BLEU (20.9→18.0). We have identified model capacity as a limiting factor in massively multilingual models, and we have investigated methods to increase model capacity without incurring too much cost in efficiency. Unfortunately, just increasing the number of layers in a typical Transformer model leads to vanishing gradient and unstable training. Thus, we first developed methods to train deep and efficient models in a bilingual setting (Zhang and Sennrich, 2019; Zhang et al., 2019). Our contributions consist of a novel depth-scaled initialization for Transformers that allows training of deep models (up to 30 encoder and decoder layers) without gradient vanishing, a more efficient variant of layer normalization, and a merged attention mechanism for the decoder that increases efficiency.

Specifically for multilingual models, we also investigate methods to keep some parameters in the encoder specific to the target language. Specifically, we consider having language-aware bias terms in the model's layer normalization (LALN), and a language-aware linear transformation on top of the encoder (LALT).

Results of UEDIN experiments in multi-target MT are shown in Table 3. We can see that both the language-aware components and deep models benefit multi-target MT models. Compared to our baseline, we see an average improvement of 3 BLEU for high-resource languages, 7.5 BLEU for medium-resource languages, and 9.6 BLEU for low-resource languages. On the selected language pairs with bilingual results, we see an average improvement of 3.2 BLEU. While we outperform the 6-layer bilingual baseline, performance is still below that of deeper bilingual systems, but the gap has become smaller.

4.2 Exploring Mid-Sized Multi-Lingual Models

CUNI has performed language clustering experiments for multi-target MT, with the aim of exploring the effects of language relatedness and determining the optimal number of target languages in a single multi-lingual model. No modifications were applied to the Transformer model in the experiments described below. The model size and other hyper-parameters are constant for all setups inside a particular experiment.

In our first experiments, we used the 'en-to-36' dataset, which is the English-sourced half of the dataset described in Section 2. With a relatively high number of languages in the dataset, it is possible to train a sufficient number of models which include target languages related in many different ways (e.g. related by script, by language group, or by some of WALS features;

	High	Med	Low	Avg	Avg (DE/ZH/BR/TE)
bilingual (6 layers)	-	-	-	-	20.9
bilingual (12 layers)	-	-	-	-	22.8
one-to-many (6 layers)	21.8	26.5	24.3	24.2	18.0
one-to-many (6 layers + LA*)	22.8	30.5	34.5	28.6	20.1
one-to-many (12 layers + LA*)	23.8	31.6	32.5	29.3	19.9
one-to-many (24 layers + LA*)	24.8	34.0	33.9	30.9	21.2

Table 3: One-to-many translation performance for deep models, and models with language-aware components (LA*). Scores are grouped based on amount of training data into high-resource ($\geq 0.9M$; 45), low-resource ($\leq 0.1M$; 18) and medium-resource (others; 31) languages. Bilingual systems are trained and evaluated on typologically different languages DE, ZH, BR, and TE.

Dryer and Haspelmath, 2013), as well as randomly selected languages for comparison. Here we expect to see the effects of shared subword vocabularies, sentence structures, etc. The 'en-to-36' dataset has the benefit of language diversity but it suffers from the differences in the underlying data sources: Aside from the subject of our study, the varying set of languages, the observed differences in performance could be also attributed to the differences in the datasets the models are being trained on.

Here and below '1-to-N model' refers to a model that translates from English to N target languages. For instance, 'en→[de, nl]' is a 1-to-2 model that translates from English to German and Dutch. Its training data is the (shuffled) concatenation of 'en→de' and 'en→nl' training sets. The target language tag is prepended to each source sentence as described in Section 2.

For the 'en-to-36' dataset, there are multiple setups being considered. First of all, in a 'random' setup the models are randomly grouped into sets by 9. For each set, a number of one-to-many experiments are generated, so that every language from the set occurs in 1-to-2 up to 1-to-5 setting three or more times. As of now, only for one of 9 languages sets the models were trained. This way, we expect to observe an average performance decreasing with the number of target languages in the model. Also, we expect to observe a more pronounced decrease when target languages with different scripts are mixed in a model.

Next, languages can be grouped by particular linguistic characteristics. So far, two sets were considered: Slavic languages with Cyrillic script and Germanic languages. To this end, we ran experiments from 1-to-2 to 1-to-5 for Germanic languages and to 1-to-4 for Cyrillic-written languages, organizing the experiments in the same manner as for random sets.

Table 4 and Table 5 show results for some of target languages. In total, we trained 83 models for these experiments. For presentation purposes, we focus on targetting Bulgarian (bg) and Ukrainian (uk) in the Slavic experiment and Danish (da) and Swedish (sv) in the Germanic experiment. In other words, one can see this as a study of how multi-target models cater for Bulgarian, Ukrainian, Danish and Swedish.

In both tables, we vary the number of target languages in the model (see the column "#TG") and in the Slavic experiment, we consider two different test sets for Bulgarian. In all cases, we report the average BLEU score when translating into the given language using a 1-to-#TG model. The "surrounding" target languages in the model affect the performance, but there are too many possible sets of these languages so we have to only sample from from. The column "#" indicates how many different model trainings (with different target language sets) are included in the average BLEU.

Comparing the average BLEU scores in the column "Random" with BLEU scores in the column "Cyrillic" (or "Germanic", resp.), we see a gain of 1.0 to 1.5 BLEU when model is trained on closer languages, i.e. when the surrounding target languages are all from the Cyrillic or Germanic group.

In both tables, we observe a clear decrease in BLEU when more target languages are included in the model and the technique described in Section 4.1 should clearly be used in our future

Target language	Dataset	version	Biling. BLEU	#TG	random		Cyrillic	
					BLEU	#	BLEU	#
bg	Europarl	v7	41.70	2	39.13	3	**40.75**	2
				3	37.88	4	**39.25**	2
				4	37.04	5	**38.00**	1
				5	36.10	3	-	-
	OpenSubtitles	v2018	22.80	2	21.17	3	**23.20**	2
				3	20.43	4	**22.20**	2
				4	19.70	5	**21.30**	1
				5	19.87	3	-	-
uk	OpenSubtitles	v2018	14.00	2	**12.75**	2	12.15	2
				3	11.00	3	**12.20**	2
				4	10.03	4	**11.30**	1
				5	9.88	4	-	-

Table 4: Sample BLEU scores for experiments with Slavic languages with Cyrillic script (ru, uk, mk, bg) and with random set of target languages. "#TG" is the main parameter of interest, the number of target languages per one model. The column "BLEU" is the average BLEU score and the column "#" reports the number of models in this group across which the average is reported. "Biling. BLEU" is the benchmark, BLEU of the simple pairwise model (i.e. #TG of 1).

Target language	Dataset	version	Biling. BLEU	#TG	random		Germanic	
					BLEU	#	BLEU	#
da	Europarl	v7	33.70	2	32.70	1	**33.05**	2
				3	31.57	3	**32.50**	2
				4	31.00	3	**31.95**	4
				5	30.28	6	**31.77**	3
sv	Europarl	v7	33.60	2	32.35	2	**32.60**	2
				3	31.15	4	**31.85**	2
				4	29.90	2	**31.40**	2
				5	**31.20**	1	31.15	2

Table 5: BLEU scores for Germanic languages (da, de, nl, no, sv, is) and random set of target languages. Columns as in Table 4. Underscored values indicate unreliable result: 'en → [bg, da, ka, sv, uk]' training data contains ∼50% of sv test set from Europarl v7, while 'en → sv' training data contains only ∼3% of this test set.

experiments.

Unfortunately, there are also anomalous values observed, see the underlined BLEU scores in Table 5. The reason for this may be in the sampling issue described in the following.

For our '1-to-36' experiments, we relied on the datased prepared and described in Deliverable 1.1. The main focus when the dataset was prepared was the coverage of the target languages. Because some of the languages were known to have very limited data sources, no strict filtering was applied to the training vs. test sets. This is particularly problematic in our multi-target experiments, because (source) sentences in the test set for one language can occur among the (source) sentences in the training data for another target language. The full detail of these overlaps is reported in Appendix A.[3] Some of our experiments are affected too much by this overlap. Tables 4 and 5 contain only the results where the overlap was not too big and the obtained scores are generally trustworthy. In these tables, 'Europarl v7' and 'OpenSubtitles v2018' refer to the parts of test set sampled from Europarl v7 (Tiedemann, 2012) and OpenSubtitles v2018

[3] In Section 2, the problem of test set overlap was avoided by testing on News Commentary v14, a distinct test set. However, this test set does not provide a sufficient number of different target languages, so we couldn't have used it here.

(Lison and Tiedemann, 2016).

Since the desired target language in multi-lingual models is indicated only as one of the input symbols, it is technically possible that the model starts producing a wrong language. This behaviour is rather rare, but we still observed it in the 'en → [ru, uk]' setup. The Russian training data contains the data from common domains as well as news domain and the political domain. The Ukrainian training data consists almost completely of common domain sentences. When various sentences from newspapers were passed to the 'en → [ru, uk]' model with the desired language tag <2uk> prepended (which means the requested target language is Ukrainian), sometimes the translation was produced partially in Russian, partially in Ukrainian. A possible reason for that may be that with lack of domain data in one language the model may prefer switching into another language which has more training data in this domain instead of attempting to translate into the requested target language. To check this observation, more setups with target languages that have a big portion of the sub-word vocabulary shared among them will be tested.

WALS features setup is currently a work in progress. In this setup, models will be grouped with the nearest models in the selected WALS features embedding space. Selection of particular features and suitable embedding type (PCA, UMAP or t-SNE) is to be decided.

Additionally, we are starting to experiment with the UN parallel corpus (Ziemski et al., 2016) instead of the '1-to-36' corpus. The full multi-parallelism in the UN corpus allows us to exclude the effect of the text content and shared linguistic structures or features and to concentrate solely on measuring the negative effect of adding one more target language to the model of the same size. The downside is that the set of languages needed by ELITR, the EUROSAI languages, is quite considerably different from UN languages.

5 Task T4.4 Multi-Source MT (CUNI, UEDIN, KIT)

This task is planned for year 2 and it can focus on translating in the written or spoken domain.

Since the spoken domain is generally harder to obtain, we already started gathering data from the seminars and mock interpreted conferences of students of interpreting from Institute of Translation Studies, Faculty of Arts, CUNI. We have recordings from three mock interpreted conferences, around 213 minutes of speech in Czech, French, German, English and Spanish. The Czech source is interpreted into all the mentioned languages. Non-Czech source is interpreted into Czech and from Czech into other languages. Depending on the availability of the students during the conference and their need for breaks, some directions are missing or are provided multiple times in parallel. There are at most 8 parallel channels. Similarly, we have recordings from seminars of simultaneous interpreteting between Czech and German (69 minutes, 7 channels), Russian (32 minutes, 3 channels) and French (32 minutes, 7 channels). There is one source interpreted into the other language in independent parallel channels.

The data from the Institute of Translation Studies are unique as a source of Czech interpreting. There are some publicly available corpora of interpreted speech (e.g. Iranzo-Sánchez et al., Di Gangi et al., 2019), but none of them contains Czech. Although the data contain students' interpreting, and therefore may contain imperfections in the interpretation, it is unique because of the parallelism. It may be used for analysis of interpreting (together with other sources) or for evaluation of multi-source MT.

6 Task T4.5 Flexible Multi-Lingual MT (CUNI, UEDIN, KIT)

This task is planned for year 3.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. *CoRR*, abs/1903.00089, 2019. URL `http://arxiv.org/abs/1903.00089`.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating Discourse Phenomena in Neural Machine Translation. 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 1304–1313, New Orleans, Louisiana, 2018. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N18-1118`.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June 2019.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/`.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. Submitted to ICASSP2020.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018. URL `https://arxiv.org/abs/1804.00344`.

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. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. English-czech systems in wmt19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy, August 2019. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W19-5337`.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. A test suite and manual evaluation of document-level nmt at wmt19. In *Proceedings of the Fourth*

Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 455–463, Florence, Italy, August 2019. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W19-5352.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair 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)*, pages 876–885, Hong Kong, China, November 2019a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1081.

Elena Voita, Rico Sennrich, and Ivan Titov. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019b. Association for Computational Linguistics. URL https://arxiv.org/pdf/1905.05979.pdf.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. Sao wmt19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy, August 2019. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W19-5355.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019.

Biao Zhang, Ivan Titov, and Rico Sennrich. Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1561.

A Test/Train Data Overlap

The following table reports the number of source (en) sentences in test sets (rows) that are also present in training sets (columns):

test	total	train																																	
		ar	az	bg	bs	cs	da	de	el	es	et	fi	fr	ga	he	hr	hu	is	it	ka	lt	lv	mk	mt	nl	no	pl	pt	ro	ru	sk	sl	sq	sr	sv	tr	uk
ar	26k	**1674**	1519	675	791	1299	32	40	57	49	485	44	42	1635	801	762	510	713	43	936	452	457	749	226	40	727	445	43	588	3266	496	481	761	763	40	766	699
az	4k	16	**3666**	16	17	15	1	3	2	3	18	1	869	20	4	28	8	2	515	18	23	8	137	3	18	24	19	19	12	16	22	20	1	23	17		
bg	26k	1474	2023	**2831**	1703	3453	1117	991	1419	1048	3841	1042	1057	2864	1593	1649	3679	1655	1026	1259	3907	3777	1216	2499	1035	1544	3952	1088	4068	1559	3900	3714	1593	1622	1162	2027	1489
bs	10k	593	1873	587	**960**	50	34	29	38	28	433	38	38	855	692	708	446	702	38	880	440	421	734	150	28	663	389	33	540	532	453	440	706	717	24	691	630
cs	38k	4541	1912	2749	1881	**3480**	1063	939	1444	1036	3706	1079	1035	3334	1467	1533	3586	1588	1006	1475	3775	3763	1272	2203	1002	1649	3691	1023	3083	5987	4105	3724	2072	1571	1111	1800	1396
da	22k	1255	181	1481	1456	1121	**1573**	1853	1652	1876	1594	1835	1900	2006	1398	1403	1569	1400	1804	1056	1613	1649	1128	1530	1894	1400	1536	1891	1482	1251	1652	1635	1478	1410	1819	1621	1371
de	56k	4803	2934	1663	1599	3539	1925	**1302**	1631	1822	1630	1849	1895	3147	1495	1526	1683	1557	1711	1940	1686	1691	1421	1770	1897	1605	1616	1967	1626	5327	1782	1644	1881	1544	1942	1765	1416
el	32k	1506	826	1768	1678	1430	1928	1850	**2081**	1843	1952	1822	1798	3139	1607	1628	1926	1658	1722	1676	1969	1937	1166	1681	1838	1545	1864	1897	1728	1552	2025	1966	1651	1694	1887	1914	1525
es	54k	5007	2242	1759	1808	3445	1976	1836	1645	**1625**	1754	1866	1898	2756	1736	1712	1767	1706	1825	1574	1724	1728	1417	1702	1914	1849	1660	1940	1687	5416	1820	1754	2165	1826	1855	2059	1577
et	24k	1382	795	2682	1621	3227	1039	994	1441	1021	**2594**	1062	1039	2968	1555	1534	3611	1625	1002	1694	3839	3817	1292	2202	985	1480	3689	1046	3054	1363	3791	3708	1598	1555	1098	1818	1409
fi	28k	1410	1056	1624	1664	1280	1925	1855	1620	1921	1770	**1782**	1890	3896	1561	1626	1736	1650	1808	1818	1805	1862	1148	1749	1941	1770	1710	1901	1667	1487	1833	1777	2151	1630	1960	1894	1499
fr	68k	5151	2462	1696	1778	3351	1892	1897	1692	1880	1655	1927	**1689**	3359	1697	1722	1702	1645	1761	1790	1673	1652	1342	1809	1887	1800	1641	1930	1631	5483	1708	1690	2052	1723	1888	1924	1601
ga	8k	29	629	820	71	1072	13	13	16	21	1029	7	10	**4504**	29	40	973	102	15	1111	1101	983	43	2411	19	34	1125	11	1040	275	923	1131	44	53	14	412	18
he	12k	815	374	780	986	59	44	44	45	37	554	48	58	2053	**1184**	956	592	927	46	1303	548	583	965	196	43	1142	534	45	738	759	596	571	1413	981	40	945	863
hr	18k	675	1282	670	812	81	61	55	57	61	503	58	40	2629	761	**983**	520	724	44	1483	484	464	784	305	43	748	447	51	596	587	497	489	789	792	54	767	711
hu	26k	1206	388	2723	1500	3159	1061	981	1441	1007	3675	1058	1068	2839	1375	1445	**2498**	1439	1008	1508	3773	3766	1270	2158	1012	1483	3718	1036	3076	1301	3751	3706	1874	1442	1125	1747	1268
is	8k	711	665	670	854	66	37	44	45	46	516	38	37	2770	809	793	508	**1791**	39	1787	555	569	909	298	29	866	457	37	616	599	541	532	989	804	32	791	752
it	50k	5853	2490	1683	1788	2957	1813	1726	1538	1785	1666	1825	1846	3354	1654	1743	1666	1732	**1655**	1812	1641	1699	1399	1737	1780	1875	1601	1841	1613	4280	1661	1673	2230	1788	1796	2011	1592
ka	6k	603	105	575	747	39	40	36	30	30	404	28	29	1567	692	690	434	821	27	**3370**	453	450	941	138	32	724	345	31	529	526	437	423	968	713	31	679	745
lt	20k	1174	722	2585	1359	3174	1043	964	1425	1047	3573	1072	1058	3301	1268	1315	3499	1559	1011	1485	**2458**	3732	1107	2238	1025	1340	3617	1062	2840	1257	3636	3505	1580	1326	1146	1586	1281
lv	20k	1181	166	2446	1341	3095	1078	931	1442	1065	3541	1090	1059	2648	1295	1248	3472	1522	1021	1114	3695	**2515**	1069	2047	1068	1289	3505	1040	2757	1144	3571	3515	1557	1295	1180	1502	1338
mk	14k	717	125	654	844	49	41	38	35	38	481	50	44	1642	834	831	500	851	35	1059	521	505	**1378**	154	36	803	431	39	587	651	512	506	1012	799	38	754	786
mt	10k	512	31	705	508	1324	55	12	57	60	1159	45	54	1417	515	516	1183	558	56	98	1300	1178	63	**2078**	56	506	1269	55	1060	682	1116	1281	513	516	51	830	503
nl	46k	2284	2481	1597	1684	2125	1878	1906	1661	1961	1645	1905	1890	3661	1632	1662	1707	1655	1793	1968	1704	1664	1309	1705	**1681**	1820	1604	1877	1603	2770	1721	1681	2162	1708	1869	1860	1487
no	8k	839	1715	802	1039	54	45	44	41	37	604	46	40	76	996	1017	595	1107	34	686	607	619	1076	44	41	**1399**	540	42	768	760	656	615	1047	1040	39	955	954
pl	36k	1362	2101	2802	1648	3110	1131	977	1538	1077	3668	1125	1078	2839	1533	1655	3599	1602	1029	1418	3720	3703	1253	2081	1077	1639	**2476**	1111	3135	1425	3768	3583	2011	1691	1238	1836	1459
pt	40k	2757	2617	1667	1746	2005	1886	1812	1773	1879	1728	1899	1911	2565	1702	1705	1765	1608	1813	1690	1749	1731	1442	1568	1835	1573	1698	**1587**	1728	2445	1760	1734	1704	1759	1883	1912	1570
ro	34k	1402	2938	3580	1663	3201	1078	1012	1448	1034	3673	1116	1075	3818	1556	1645	3660	1576	1017	1979	3728	3724	1253	2495	1025	1757	3771	1034	**3142**	1439	3781	3710	2095	1630	1102	1904	1409
ru	36k	4624	2043	1124	1452	2810	100	96	99	92	871	99	82	1896	1411	1408	884	1415	90	1466	899	876	1420	319	84	1570	824	93	1089	**4937**	920	871	1925	1449	91	1426	1298
sk	20k	1214	378	2386	1345	3222	1046	918	1416	989	3484	1049	1021	3119	1318	1360	3413	1326	1014	1255	3599	3573	850	2052	1048	1490	3546	979	2769	1252	**2278**	3464	1807	1372	1133	1626	1221
sl	22k	1439	383	2680	1647	3213	1084	949	1454	1061	3623	1091	1057	3261	1571	1667	3545	1706	992	1446	3743	3710	1274	2203	1043	1768	3639	1036	2978	1471	3677	**2523**	2133	1700	1143	1917	1524
sq	12k	722	2117	671	886	44	23	29	25	27	485	23	30	943	820	826	512	969	23	1062	529	532	1013	108	25	827	437	31	620	621	555	472	**1658**	838	21	758	835
sr	14k	643	114	605	766	54	36	27	36	38	430	37	41	1342	716	719	449	654	38	872	435	416	687	138	41	652	424	39	547	533	465	463	707	**918**	31	723	630
sv	28k	1358	2100	1561	1628	1239	1927	1894	1776	1907	1814	1984	1941	2751	1513	1611	1734	1627	1812	1420	1789	1792	1211	1711	1849	1700	1709	1882	1644	1395	1838	1816	2092	1624	**1796**	1769	1412
tr	22k	1046	2035	971	1290	82	53	59	59	59	771	56	52	1852	1237	1306	742	1220	54	1374	716	762	1275	177	62	1324	693	59	907	934	767	725	1679	1268	47	**1942**	1118
uk	8k	803	108	743	935	55	38	39	46	44	556	44	53	1031	940	911	555	902	37	889	567	576	986	119	37	935	518	43	713	754	576	577	973	953	35	909	**2655**