M5.1 – Intrinsic Evaluation of the Machine Translation Engines

Cristina España-Bonet and Ádám Varga
Universität des Saarlandes

– v1.2 –
July 2018

Abstract

This document describes the statistical and neural machine translation systems used as baselines within the CLuBS project and their evaluation with automatic metrics. We also present several experiments to improve the baselines and we justify the choice of the system for the first translation prototype. Those systems are trained with large out-of-domain and/or small in-domain corpora, which is described in deliverable M1.2. Modifications and improvements for the final prototype are also sketched.

Contents

1 Introduction

In the last years, statistical machine translation systems (SMT) have shown to offer a competitive translation quality when *enough data* is available. When compared to rule-based translation systems (RBMT), SMT systems are cheaper and easier to build and they only rely on the existence of parallel corpus –preferably on the specific domain to translate– in order to be trained. These advantages are common to all data-based systems. Recently, a new paradigm has been applied within data-based machine translation: neural machine translation (NMT). NMT systems have shown to outperform SMT systems for language pairs where huge amounts of parallel corpora are available [3, 2].

As always, there are pros and cons for using one family of systems or the other. SMT systems need less data to get a similar performance and can be trained in a desktop computer. On the other hand, the most modest NMT systems need at least one week of GPU time to get competitive results. A more detailed discussion is reported in deliverable M1.4.

In the CLuBS project we face two main issues. First, we want to translate among all four languages *en–de–*fr*–es*. Second, we want translations on a very specific domain: psychology. The combination of the two aspects makes gathering the appropriate data difficult: we either cannot get data for all language pairs or, in case we can, the data does not match our domain. For translating between languages with few data, SMT engines offer the possibility of pivot translation. Translation models for translating L1-to-L2 and L2-to-L3 can be joined to build a translator L1-to-L3 [31]. Usually the pivot language L2 is a rich language such as English. Then, only a small amount of parallel data in the pair L1-L3 is needed for tuning the system. On the other hand, NMT systems offer the possibility of a joint training and zero-shot translation [17, 16, 11]. Training a single model for L1-to-L2 and L3-to-L4 allows a zero-shot translation L1-to-L4. This is probably due to the fact that context vectors (or attention vectors) for different languages seem to share a space when trained together.

In the CLuBS setting, one needs either 12 domain adapted SMT or bilingual NMT systems to cover all translation directions, or a single multilingual NMT system. This document presents the first results with the three different families. Systems are trained with the generic corpora described in deliverable M1.2 and evaluated on both out-of-domain data and in-domain whenever it is possible. Section 2 introduces the automatic metrics we use for this evaluation. Next, Sections 3 and 4 describe the SMT and NMT engines respectively, and show their performance. Section 5 introduces several experiments conducted to improve specific aspects in NMT. Finally, we summarise and depict the next steps in Section 6.

2 Automatic Evaluation Metrics

Manual evaluation is the most reliable way to quantify the quality of a translation but it is also a very costly way (both in time and money). For a fast and objective evaluation during the system development one needs to make use of automatic metrics. Automatic metrics are costless, objective and reusable but they cannot capture all the aspects that a human evaluator takes into account. In order to somehow overcome this limitation, we do not use a single metric such as the standard BLEU but an heterogeneous set.

The `Asiya` evaluation package [13, 15] includes more than 500 metrics and their variants at lexical, syntactic and semantic levels. Syntactic and semantic metrics need linguistic processors to annotate the translations and are not available to all the languages, but a large set of lexical metrics can always be used. In the following, we select a subset of lexical metrics to be applied to English, French, German and Spanish:

Lexical metrics

- PER [29], TER [27], WER [22]: Based on edit distances
- BLEU [25], NIST [6], ROUGE [20]: Based on n-gram matching (lexical precision: BLEU, NIST; and lexical recall: ROUGE)
- GTM [21], METEOR [1]: Based on the F-measure
- ULC [14]: *U*niform *L*inear *C*ombination. When applied to lexical metrics it includes WER, PER, TER, BLEU, NIST, RG-S* (ROUGE-S*), GTM-2, MTRpa (METEOR with paraphrasing).

3 Statistical Machine Translation Systems

In our experiments, we build a state-of-the-art phrase-based SMT system based on `Moses` [19] trained both on a large general domain corpus and on a smaller in-domain one. Domain adaptation of the general domain system via the development set is applied. Translators are trained for the *es–en*, *de–en* and fr*–en* language pairs and pivot translation is not used at this point. The corpora for training, development and testing and their pre-processing are described in deliverable M1.2 and summarised in Tables 1 and 2. These tables include the complete corpora which is treated differently depending on the engine. SMT engines apply an additional cleaning step, where sentences with more than 100 tokens in any of the languages and sentences 9 times longer in one language than in the other one are discarded. These sentences are not likely to be parallel and, besides, the alignment software is not able to perform well on this data. On the other hand, neural systems discard sentences that are longer than 50 words, since the translation quality for these sentences drops and would damage the training process.

The development of the system has been done using standard freely available software. Word alignment is done with `GIZA++` [24] and both phrase extraction and decoding are done with the `Moses` package. The optimisation of the weights of the model is trained with Minimum Error Rate Training (MERT) [23] against the BLEU evaluation metric. Our model considers the standard features: language model, direct and inverse phrase probabilities, direct and inverse lexical probabilities, phrase and word penalties, and lexicalised reordering.

A 5-gram language model is estimated using interpolated Kneser-Ney discounting with `SRILM` [28]. Since we have monolingual corpora of four main topics (Wikipedia, Politics, Medicine and Psychology), we build the corresponding language models and compile a general one by interpolation of the previous four and tuning the weights on the PubPsych development set.

3.1 Analysis of the Results

We evaluate the translation engines involving English, that is, those language pairs for which we have parallel corpora, with a set of lexical metrics. We use `news-test2013` as out-of-domain test set and the `pubPsych test` of titles and abstracts as in-domain test sets. We run our experiments with three different systems:

Table 1. Summary of the size of the General, EMEA and Scielo parallel corpora used to train the translation engines.

	en–de			*en–es*			*en–fr*		
	Sentences	*en* tok.	*de* tok.	Sentences	*en* tok.	*es* tok.	Sentences	*en* tok.	*fr* tok.
UN	162,981	6,098,083	5,617,876	11,196,913	320,064,682	366,072,923	12,886,831	361,877,676	421,687,471
EP	1,920,209	53,091,548	50,548,739	1,965,734	54,505,707	57,047,216	2,007,723	55,730,752	61,888,789
ComCrawl	2,399,123	58,864,439	54,570,779	1,845,286	46,855,705	49,557,537	3,244,152	81,084,856	91,281,890
Subtotal	**4,482,313**	**118,054,070**	**110,737,394**	**15,007,933**	**421,426,094**	**472,677,676**	**18,138,706**	**498,693,284**	**574,858,150**
EMEA	1,108,752	14,477,119	13,197,725	1,098,333	14,334,648	15,975,506	1,092,568	14,317,365	17,046,979
ScieloBio	–	–	–	117,862	3,252,183	3,382,511	–	–	–
ScieloHealth	–	–	–	558,714	14,382,853	15,031,533	9,129	244,486	308,055
Subtotal	**1,108,752**	**14,477,119**	**13,197,725**	**1,774,909**	**31,969,684**	**34,389,550**	**1,101,697**	**14,561,851**	**17,355,034**
pubPsychAbstracts	241,749	6,584,364	6,135,612	88,848	2,640,441	2,909,559	–	–	–
pubPsychTitles	306.640	3.480.727	3.059.048	25.105	293.164	340.203	45.137	463.610	567.618
Total	**5,832,814**	**139,115,553**	**130,070,731**	**16,871,690**	**456,036,219**	**509,976,785**	**19,240,403**	**513,255,135**	**592,213,184**

Table 2. Size of the development and test sets used for the evaluation of the translation engines.

	en–de			*en–es*			*en–fr*		
	snt.	*en* tok.	*de* tok.	snt.	*en* tok.	*es* tok.	snt.	*en* tok-	*fr* tok.
news-test2012	3,003	72,988	72,603	3,003	72,988	78,887	3,003	72,988	81,797
news-test2013	3,000	64,809	63,411	3,000	64,809	70,540	3,000	64,809	73,658
EMEA dev	2,000	38,658	37,945	2,000	36,676	39,959	2,000	34,554	41,026
EMEA test	2,000	36,864	35,773	2,000	34,359	38,615	2,000	33,316	39,674
pubPsych devAbs	1,500	39,968	37,557	1,500	45,611	50,831	–	–	–
pubPsych testAbs	2,162	60,219	55,610	2,486	74,382	81,575	823	25,884	29,226
pubPsych testTit	737	9.691	8.202	575	6.935	8.002	187	2.589	3.012

- **SMTgen.** Statistical system trained exclusively with the general domain data and using `news-test2012` as development set
- **SMTgenPP.** Statistical system trained exclusively with the general domain data and using `pubPsych dev` as development set for domain adaptation. We do not use this system for testing out-of-domain data as it has been domain-adapted
- **SMTpp.** Statistical system trained exclusively with the in-domain PubPsych data and using `pubPsych dev` as development set. Notice that we do not have in-domain data for French–English, so this system cannot be trained for the language pair

Tables 3 and 4 report the automatic evaluation for German-to-English and English-to-German. As expected, for translating out-of-domain data the general system SMTgen performs much better than SMTpp, since the general domain corpus is 24 times larger. However, the specialised system performs better for translating the in-domain tests because the vocabulary is more adequate. Notice that domain-adapting the general system via the development set (SMTgenPP) the translation quality on PubPsych is only slightly improved, but the increment is statistical significant when using SMTpp. One can see for instance in the German-to-English translation, that the BLEU scores for abstracts increase with these three systems from 12.82→13.16→15.71 for abstracts and 23.79→24.96→33.18 for titles. The same trend is observed for English-to-German but with lower scores. The translation quality is however too low especially for abstracts (BLEU is 15.71 for *de2en* and 11.05 for *en2de*) and this is a first motivation to try to use neural systems to make the most of the data.

Tables 5 and 6 report the automatic evaluation for Spanish-to-English and English-to-Spanish respectively. The conclusions and trends are the same as for German–English, only with a better translation quality as measured by the automatic metrics. In general, the inflexions and compound nature of German, make it a difficult language to translate from and especially into. This together to the fact that the general domain corpus is relatively small for German ($\sim 5M$ parallel sentences compared to the $\sim 16M$ in Spanish) justify the difference in the automatic metrics scores.

Finally, for French–English we do not have data to train the specialised system, but the large amount of general corpora allows a translation quality for this pair similar to that obtained for Spanish–English with the in-domain corpus for titles which are easier to translate, and around 6 BLEU points worse for abstracts. The concrete figures can be read in Tables 7 and 8 for French-to-English and English-to-French respectively. The

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	61.58	39.99	56.55	24.67	7.14	27.28	30.82	33.65	68.30
SMTpp	70.99	48.78	66.66	15.25	5.52	21.98	24.07	22.80	45.69

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	85.87	59.36	82.24	12.82	4.33	17.65	18.55	17.11	57.45
SMTgenPP	87.48	58.36	83.78	13.16	4.30	17.54	18.90	17.34	57.71
SMTpp	84.93	55.56	80.98	15.71	4.80	19.12	20.54	20.61	66.34

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	62.31	47.26	58.90	23.79	5.97	33.49	26.29	33.99	78.14
SMTgenPP	60.88	45.94	57.75	24.96	6.04	33.99	26.96	34.80	79.56
SMTpp	54.87	38.23	51.23	33.18	7.13	39.50	32.24	44.86	92.25

Table 3. Automatic evaluation of the baseline MT systems for German-to-English on three test sets: News (news-test2013), Abstracts (PubPsych testAbst) and Titles (PubPsych testTit). See Section 3 for system's description.

capability of neural systems to deal with zero-shot language pairs is a second motivation to explore their performance in our setting.

4 Neural Machine Translation Systems

In order to train our neural systems we use the state-of-the-art toolkit Nematus [26]. The data are the same as described in the previous section for the SMT systems and summarised in Tables 1 and 2. Here, though, only sentences shorter than 50 tokens are used for training and validation. Besides, the in-domain data from PubPsych is not enough to train an NMT system, so we use this subset for domain adaptation as explained in Section 4.1.

We build our multilingual many-to-many NMT engine similarly to [17], that is, we train our system on parallel corpora for several language pairs Li–Lj simultaneously, adding a tag in the source sentence to account for the target language "<2Lj >" (e.g., <2de> if the target language is German). Systems are trained on 4 language pairs or, at least, using data from the pairs: *de–en*, fr–*en*, *es–en* and *es*–fr. Although some corpora exist for the remaining two (*es–de* and fr–*de*), we exclude them in these experiments to study these pairs as instances of zero-shot translation[1]. With the additional cleaning for NMT, we obtain ∼15 M parallel sentences per language pair —for *de–en* in the multilingual system, we oversampled to reach that amount by tripling the original sentences.

We conducted a full analysis of the multilingual system to chose the best configuration. This analysis also included a new methodology to select parallel sentences from comparable corpora using the context vectors of the multilingual NMT system. This way, we are able to obtain additional in-domain parallel sentences from Wikipedia in the psychological domain and use them for domain adaptation purposes. All the details can be found in the

[1] All the data will be compiled for the final CLuBS translation engine and trained on the best configuration.

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	69.39	48.05	65.67	17.99	6.00	23.40	38.35	24.16	67.26
SMTpp	75.21	56.05	72.37	10.73	4.70	19.18	29.45	15.11	45.86

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	97.57	72.01	94.96	8.40	3.31	14.73	23.91	10.49	58.66
SMTgenPP	93.65	66.56	91.00	8.82	3.45	14.99	23.50	10.63	61.43
SMTpp	91.15	63.20	88.53	11.05	3.88	16.35	25.72	12.79	70.59

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	79.55	68.95	77.57	16.45	4.41	28.36	37.55	23.23	75.87
SMTgenPP	74.97	63.42	72.92	17.85	4.64	29.29	37.71	24.00	78.38
SMTpp	66.11	54.71	63.92	24.52	5.60	34.84	44.96	32.94	92.41

Table 4. As Table 3 for English-to-German.

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	55.07	36.54	51.34	30.04	7.79	30.04	34.25	37.92	68.92
SMTpp	65.08	45.18	61.73	19.73	6.12	24.10	27.65	26.20	46.79

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	58.78	37.87	55.19	30.03	7.64	26.87	32.58	40.46	65.97
SMTgenPP	58.63	37.36	54.94	30.54	7.70	27.01	33.06	40.87	66.87
SMTpp	58.43	36.94	54.70	31.16	7.77	27.43	32.75	41.24	67.62

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	52.36	35.24	46.16	33.88	7.07	41.25	34.45	51.95	84.42
SMTgenPP	51.55	35.10	45.16	34.86	7.16	42.14	35.01	52.85	85.78
SMTpp	51.32	34.35	44.50	35.88	7.23	42.78	34.88	52.76	86.54

Table 5. As Table 3 for Spanish-to-English.

associated journal paper [10] and the effect of the extracted parallel sentences is reported in the next section.

The analysis in [10] determined that the most appropriate configuration for the neural system uses the following parameters:

Word embedding dimension:	512
Hidden layer dimension:	2,048
Vocabulary size:	80,000 + 2,000 BPE units
Optimizer:	AdaDelta
Learning rate:	0.0001
Batch size:	80

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	57.02	38.87	52.93	28.54	7.53	28.26	52.07	34.51	68.36
SMTpp	66.52	46.76	62.79	19.47	6.04	23.20	41.74	24.50	47.71

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	59.85	38.23	56.11	32.16	7.61	26.99	52.90	39.17	63.65
SMTgenPP	60.14	38.76	56.41	31.95	7.59	26.85	53.03	38.91	63.14
SMTpp	59.56	37.98	55.53	32.88	7.76	27.27	53.85	40.30	65.11

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	54.20	37.24	49.58	35.46	7.03	40.47	57.61	47.56	72.60
SMTgenPP	54.79	37.90	50.04	35.11	6.97	40.05	57.06	47.10	71.70
SMTpp	54.69	38.20	49.18	36.80	7.13	41.40	58.50	48.59	73.75

Table 6. As Table 3 for English-to-Spanish.

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	55.58	37.08	51.72	30.05	7.72	30.13	33.98	37.61	67.08
SMTpp	–	–	–	–	–	–	–	–	–

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	67.37	43.54	64.10	23.96	6.12	22.82	28.94	33.91	63.39
SMTgenPP	66.57	42.77	63.46	24.83	6.24	23.21	29.33	34.81	65.24
SMTpp	–	–	–	–	–	–	–	–	–

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	51.64	35.65	46.81	36.19	6.24	40.00	36.47	49.73	72.84
SMTgenPP	50.68	35.65	46.16	37.32	6.33	41.02	37.27	50.00	74.33
SMTpp	–	–	–	–	–	–	–	–	–

Table 7. As Table 3 for French-to-English.

4.1 Domain Adaptation via Transfer Learning

There are various possibilities available to do domain adaptation in neural systems: transfer learning using a general out-of-domain system with in-domain data, ensembling adapted and unadapted systems [12], or performing target-forcing within the NMT system with respect to the training sentences' domains for instance [5]. In our project, we follow the pure transfer learning approach. That implies training a complete general domain system as explain in the previous section and, afterwards, run a few number of additional iterations on in-domain corpora. This way, the parameters of the model are optimised for the adequate data. While during the training of the general system no dropout is applied, for adaptation purposes we utilise a dropout rate of 0.2 to avoid overfitting. A master thesis

News

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	58.17	40.31	54.51	29.27	7.50	28.21	49.96	34.13	68.14
SMTpp	–	–	–	–	–	–	–	–	–

Abstracts

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	69.51	45.52	66.33	24.62	6.16	21.98	43.05	32.67	63.33
SMTgenPP	68.81	45.33	65.65	25.25	6.23	22.33	43.52	33.13	64.59
SMTpp	–	–	–	–	–	–	–	–	–

Titles

	WER	PER	TER	BLEU	NIST	GTM-2	MTRpa	RG-S*	ULC
SMTgen	54.95	37.92	51.93	34.62	6.21	37.40	55.30	43.92	82.95
SMTgenPP	55.01	37.18	51.66	35.46	6.31	38.07	55.31	43.26	83.58
SMTpp	–	–	–	–	–	–	–	–	–

Table 8. As Table 3 for English-to-French.

in the framework of the project [30] explored various scenarios regarding the available adaptation data, namely:

- using only parallel, clean data (*pubPsych*, **NMTpp**)
- using only the extracted data from comparable corpora (*Wikipedia*, **NMTwp**) with three different sizes of additional data (5k, 10k and 30k parallel sentences per language pair)
- using the combination of the two data sets (**NMTmrg**) with three different sizes of additional data (5k, 10k and 30k parallel sentences per language pair)

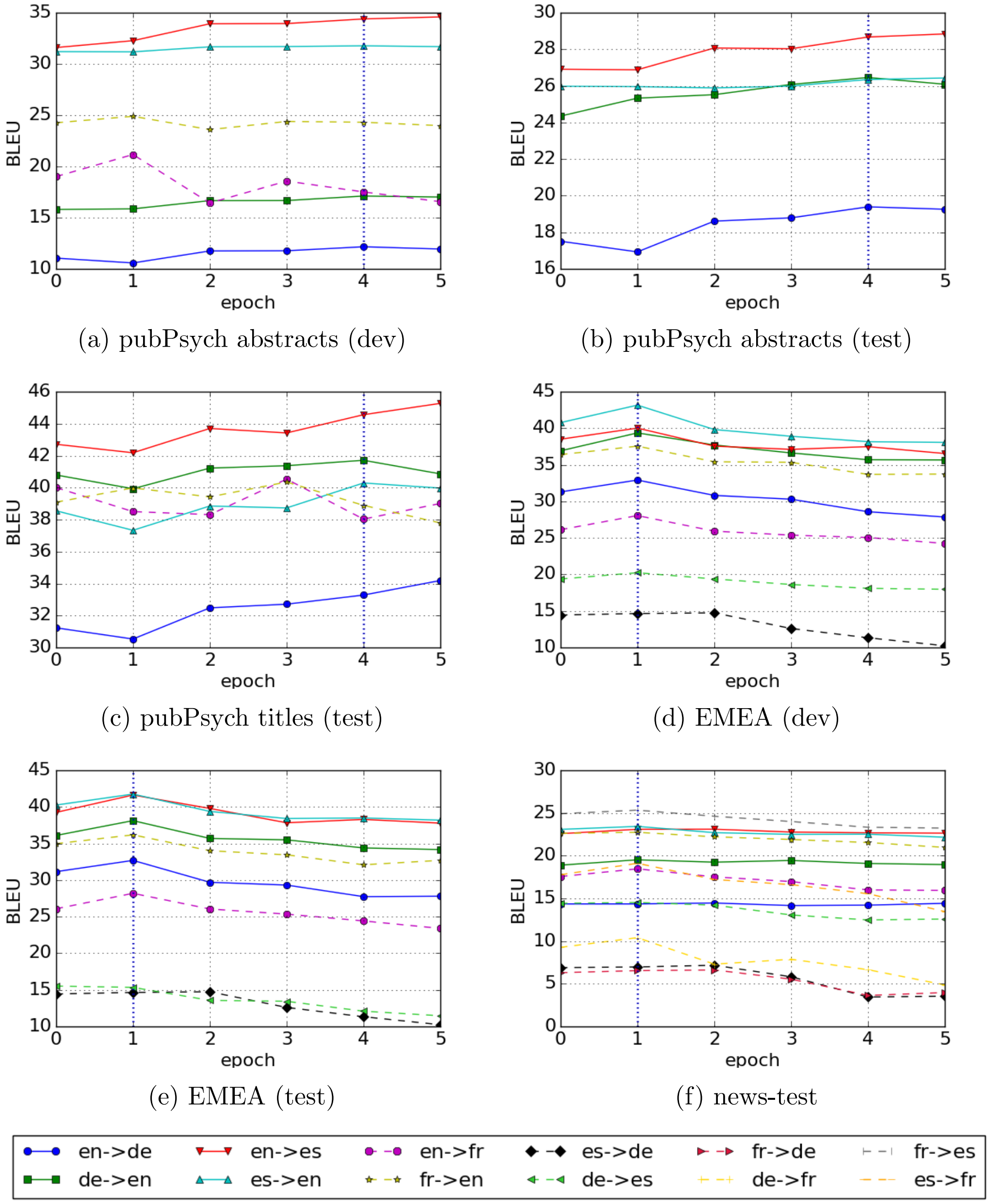
Notice that the additional data partitions are significantly smaller than the PubPsych adaptation data (1,414,958 sentence pairs in total with respect to 60,000, 120,000 and 360,000 sentence pairs added), and their covered domain is less homogeneous as they are extracted automatically from articles about health and psychology. On the other hand, they are available for all language pairs. We perform the partitioning in order to investigate the tradeoff between the amount of data and its quality. Here we include the basic results with the three approaches and refer the reader to Ádám Varga's thesis [30] for the full analysis and details.

In all cases, besides the language target forcing tags that are included in the general-domain multilingual NMT system, we additionally introduce a category tag that specifies the origin of the sentence ("title" or "abstract") for the transfer learning phase with PubPsych data. This improves translation quality on the test data where the same information is available (*pubPsych titles* and *abstracts* test sets).

For the adaptation, we run the training process for an additional 5 epochs and we examine the evolution of the translation performance in terms of BLEU score after each of these epochs.

4.2 Analysis of the Results

The adapted systems are tested on the in-domain *pubPsych* test sets, on a close-domain test sets (EMEA) and an out-of-domain one (*news-test2013*). Apart from the advantage



(a) pubPsych abstracts (dev)

(b) pubPsych abstracts (test)

(c) pubPsych titles (test)

(d) EMEA (dev)

(e) EMEA (test)

(f) news-test

Figure 1. Evolution of domain adaptation through five epochs using the pubPsych adaptation corpus. Dashed lines indicate missing language pairs from adaptation data.

of being able to see the performance evolution for the missing language pairs in the SMT systems where no data is available, this is also beneficial for observing the effect domain adaptation has on out-of-domain datasets.

Figure 1 displays the evolution of the NMT system throughout the 5 adaptation epochs. The figure is broken down by dataset, and different lines correspond to source-language pair combinations. Regarding the change in translation quality, several observations can be made. First, there are two different trends to be observed that depend on the type of the test data. It can be said in general that on in-domain datasets, the system's performance keeps improving until the fourth or fifth epoch. In most cases, the performance starts decreasing at the fourth epoch (a phenomenon that can be explained by overfitting). In the case of the three close- and out-of-domain datasets, however, the results saturate after

	en2de				*de2en*			
	Abstracts	**Titles**	**EMEA**	**News**	**Abstracts**	**Titles**	**EMEA**	**News**
NMT	11.05	31.23	31.15	14.35	15.80	40.79	36.10	18.89
NMTpp	↑**12.16**	↑33.28	32.69	14.35	↑**17.11**	41.70	↑**38.12**	19.52
NMTwp5	↓10.04	30.44	↓26.70	↓13.17	↓15.01	40.59	↓34.28	19.26
NMTwp10	↓9.80	↓28.59	↓26.67	↓13.25	15.12	↓39.24	↓33.66	↑19.37
NMTwp30	↓7.76	↓23.12	↓25.95	↓13.12	15.34	↓37.86	↓33.18	18.39
NMTmrg5	↑11.74↓	↑33.60	↓28.32↓	14.34	↑16.49↓	41.34	↓34.84↓	19.26
NMTmrg10	↑11.69↓	↑33.89	↓29.54↓	14.62	↑16.66↓	↑**42.02**	↓34.81	↑**19.46**
NMTmrg30	↑11.96	↑**33.97**	↓27.33↓	13.46↓	↑16.95	41.44	↓33.03↓	19.41

Table 9. BLEU scores for the different systems used to domain-adapt the general NMT system for the phsycological domain German–English. Arrows mark statistical significance (see text) and best results on each test set are shown in bold (in case of significant improvements).

one epoch of adaptation, and worsen after this point. The explanation for this behavior is that the overfitting takes effect much earlier due to the mismatch between the domain of the adaptation and test sets. Furthermore, this trend is repeated in the case of in-domain test sets for the language pair that is underrepresented in the adaptation set (*en–fr*, only titles are available). This is caused by the fact that even though the new data is beneficial, in this case its quality is inferior to the original training data, as that set contains sufficient parallel data for this language pair. Because of this, while one epoch of adaptation has positive effects, further training leads to decreasing quality.

In the light of the previous analysis, we choose to test our systems using the model after fourth epoch in the adaptation process. Domain adaptation can even be beneficial for translating data that is not strictly in-domain or out-of-domain. In this case, however, we use systems that have only been adapted for one additional epoch of training. Similarly, this seems to be the best point in training for translating in-domain texts between language pairs that do not have any or sufficient parallel data in the adaptation set.

Tables 9–13 show the BLEU scores for the different systems and language pairs. Arrows before BLEU scores represent significant changes with respect to the baseline performance (NMT) as calculated by the bootstrap resampling method implemented in Moses [18]. Statistical significance is indicated at $p = 0.005$. Arrows after the BLEU scores indicate significant differences when compared to the adapted results achieved by NMTpp. The evaluations are performed after 1 and 4 epochs of adaptation respectively, depending on the test set as described above. With the NMTpp system on the in-domain data, we obtain significant improvement in all cases, that generally lies between 1 and 2 BLEU scores. In certain cases the improvements are not significant though, most notably on the *en–fr* language pair where there is no adaptation data available (other cases are $es \rightarrow en$ on the *pubPsych abstracts* sets, and $de \rightarrow en$ on *pubPsych titles*).

When enriching our data with parallel sentences extracted from Wikipedia in the health domain, we see that results are 1–2 BLEU points worse compared to the baseline system's performance, when the training corpus already has a large amount of data in the given language pair (e.g. *en–de*) or we evaluate on the in-domain PubPsych tests. The quality of the data seems to be not high enough to significantly improve the in-domain performance. It has to be noted that in addition to the lower quality, the actual domain coverage of the data does not necessarily align completely with that of the test sets. This is backed up by the fact that adapting with the Wikipedia data leads to significant improvements

	en2es				*es2en*			
	Abstracts	**Titles**	**EMEA**	**News**	**Abstracts**	**Titles**	**EMEA**	**News**
NMT	31.60	42.71	39.25	22.58	31.20	38.55	40.25	23.08
NMTpp	↑34.39	↑44.56	↑**41.62**	↑23.09	31.77	↑40.29	41.75	23.43
NMTwp5	31.25	41.05	↓36.49	↓21.74	↓30.12	38.16	↓36.91	↑23.85
NMTwp10	32.09	40.91	↓36.17	↓21.92	↓30.35	37.20	↓37.56	↑23.73
NMTwp30	31.73	↓40.26	↓36.11	22.49	↓30.10	↓35.47	↓36.45	↑**23.94**
NMTmrg5	↑34.02	↑**45.31**	↓36.79↓	↑23.49	32.22	↑40.25	↓38.05↓	23.49↑
NMTmrg10	↑33.44	↑44.72	↓38.58↓	↑**23.68**↑	32.21	↑**40.37**	↓38.67↓	23.36↓
NMTmrg30	↑**34.60**	↑44.98	↓36.42↓	↑23.43	31.91	40.19	↓37.52↓	↑23.78↑

Table 10. As Table 9 for the Spanish–English language pair.

	en2fr				*fr2en*			
	Abstracts	**Titles**	**EMEA**	**News**	**Abstracts**	**Titles**	**EMEA**	**News**
NMT	19.01	40.00	26.08	17.56	24.60	39.09	34.97	22.62
NMTpp	↑21.16	↓38.35	↑**28.19**	18.46	24.89	39.98	36.18	↑22.74
NMTwp5	↑20.96	↓36.41	25.33	↑18.94	23.39	↑**40.84**	33.67	↑23.12
NMTwp10	↑22.41	↓35.39	25.36	↑19.20	22.41	39.22	↓32.85	22.85
NMTwp30	↑**22.87**	↓35.03	25.83	↑**19.28**	23.51	38.44	32.03	↑**23.38**
NMTmrg5	19.41↓	39.36	25.15↓	↑18.38↓	24.60	39.33	↓32.91↓	22.78↑
NMTmrg10	18.71↓	38.87	↑26.22↓	↑18.89↑	24.92	38.16↓	33.05	22.51
NMTmrg30	↑21.40↑	38.89	↓25.60↓	↑18.94↑	24.25	40.99	↓32.87	22.19↓

Table 11. As Table 9 for the French–English language pair. Notice that we have not used direct parallel data for this language pair during the transfer learning phase.

on the out-of-domain *news-test2013* dataset (except for *en* → *es* and *en* → *de*), while on the close-domain EMEA test sets we observe a behavior similar to the strictly in-domain *pubPsych* sets. As we aimed for covering a large number of articles when extracting from the health and psychology domains, the automatic adaptation set contains many parallel sentences that are only slightly related to these topics and they have proven not to be useful. Since in this setting we have available data for language pairs that are under-

	es2de		*de2es*	
	EMEA	**News**	**EMEA**	**News**
NMT	15.52	6.89	20.57	14.42
NMTpp	↓15.34	6.97	20.90	14.52
NMTwp5	16.58	↑9.78	↓20.55	↑**15.03**
NMTwp10	16.84	↑9.73	↑20.65	14.62
NMTwp30	16.39	↑9.94	↓19.45	↓13.98
NMTmrg5	**17.46**↑	↑10.71↑	↓20.43↓	14.75↑
NMTmrg10	17.71	↑10.68↑	↓20.05	14.90↑
NMTmrg30	17.12	↑**10.76**↑	↓20.21↓	14.96↑

Table 12. As Table 9 for Spanish–German. Notice that we have not used direct parallel data for this language pair when training the NMT system or during the transfer learning phase.

	$es2fr$	$fr2es$	$de2fr$	$fr2de$
NMT	17.77	24.92	9.27	6.30
NMTpp	↑19.11	25.33	↑10.40	6.54
NMTwp5	↑**21.30**	↑25.77	↑11.29	↑10.14
NMTwp10	↑**21.30**	↑25.92	↑**10.98**	↑10.01
NMTwp30	↑21.28	↑26.15	8.23	↑9.01
NMTmrg5	↑20.21↑	↑25.48↑	↑10.53↑	↑**10.41**↑
NMTmrg10	↑20.88↑	↑25.82↑	↑10.97↑	↑10.03↑
NMTmrg30	↑20.80↑	↑**26.18**↑	9.78↓	↑9.44↑

Table 13. BLEU scores for the different systems used to domain-adapt the general NMT system for the phsycological domain in the *es*–*fr* and *de*–*fr* language pairs. Notice that at this point of the project we only have available the out-of-domain news test set.

or zero-resourced in either the PubPsych data set, or even the general NMT system, the Wikipedia extractions have a clear positive effect for these language pairs and the adapted systems often outperform the ones that have been adapted on PubPsych data.

Finally, we combine (merge) the parallel sentences extracted from Wikipedia with those available in PubPsych in the NMTmrg system and and use them also to adapt the NMT baseline. The conclusions obtained with the NMTwp systems are also valid here as we can see in the third block of rows of Tables 9–13.

5 Additional Experiments

The previous section presents a summary of the main work done to develop the CLuBS neural machine translation system. All the experiments related to the methodology for extracting parallel sentences and a deep study of different configurations for domain adaptation can be read in the following references:

[7] Cristina España-Bonet and Alberto Barrón-Cedeño. Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 144–149, August 2017

[10] Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350, December 2017

[30] Ádám Csaba Varga. Domain adaptation for multilingual neural machine translation. Master's thesis, Universität des Saarlandes, 2017

Notice, that the fact that the project itself is the responsible for generating the evaluation data implies that we cannot evaluate our systems in the real scenario till the end of the project. Because of this, several experiments have been conducted in other settings always focusing on multilinguality and domain adaptation, that is the case of references [7] and [10].

The same happens with experiments conducted to exploit multilinguality by including factors that increase the number of common elements among languages. In these experiments, we use the TED corpus provided for the IWSLT 2017 multilingual task [4]. The corpus includes 9161 talks in five languages, 4,380,258 parallel sentences when all the language pairs are considered. The intersection of talks among languages is high, 7945 documents are common to all of them, and therefore the same sentence is available in multiple languages. Notice that the size of the corpus is small as compared to the collections of bilingual corpora that we have used for training our general domain system. However, its multilingual nature makes it adequate for our study.

In a first work [8], we focus on including several additional sources of information in a NMT system. All this information is added as factors to every word, implying that the final vector representation of a word is the concatenation of the vectors of all its features. We use high level factors such as phonetic coarse encodings and synsets, besides more common information such as shallow part-of-speech tags, stems and lemmas. Document level information is also considered by including the topic of every document.

The most promising feature turned to be BabelNet synsets, especially when combined with other factors, and in a subsequent work we study how semantic networks can help improve NMT. We show that they improve a state-of-the-art baseline and that they facilitate the translation from languages that have not been seen at all in training (beyond zero-shot translation). A small enhancement for regular language pairs comes about because cross-language synsets help to cluster words by semantics irrespective of their language and to map the unknown words of a new language into the multilingual clusters. However, improvements are at the current state moderate, and cannot be tested on the PubPsych setting because of the lack of a test set. Since the new architecture involves changing the current decoder Marian to Nematus, and that implies a significant reduction in the translation (and training) speed, we do not use this architecture in our final prototype. The detailed experiments can be read in:

[8] Cristina España-Bonet and Josef van Genabith. Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI System at IWSLT 2017. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, pages 15–22, Tokyo, Japan, December 2017

[9] Cristina España-Bonet and Josef van Genabith. Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems. In *Proceedings of the LREC 2018 MLP-MomenT Workshop*, pages 8–13, Miyazaki, Japan, May 2018

6 Conclusions

This document presents an exploration of the translation engines developed during the course of the project. We have created statistical and neural translation systems for the most representative languages in PubPsych, *en–de–fr–es*, with the aim to translate into the four languages the titles and the abstracts available as metadata in our database and improve the cross-lingual article retrieval.

After a first comparison between statistical and neural systems, we decided to focus on the latter and adapt them to better translate our domain and specific characteristics. For this, we developed methods to obtain more corpora and strategies to improve the translation of the in-domain vocabulary. The experiments reported here led to chose the system named NMTpp as the engine to translate the database of our first prototype at month 20 of the project. This system has been trained with general corpora in the *en–de*, *en–es*, *en–fr* and *fr–es* language pairs and domain-adapted with the parallel sentences extracted from the current PubPsych titles and abstracts. We use four additional iterations with our data to adapt the system to psychology via transfer learning.

In their current status, systems that include parallel sentences extracted from Wikipedia or use factors to better translate common vocabulary among languages, do not provide consistent enhancements among language pairs and imply more complex architectures to be maintained, so they are not implemented in the prototype.

For the final CLuBS translation engine to be submitted at month 36 and after analysing current translations, we are including some improvements in the NMTpp architecture such as:

- A better pre-processing of the data as reported in M1.2
- Compilation of general domain data in the missing language pairs, *de–fr*, *de–es* and *es–fr*, and in the less resourced ones adding new corpora, also reported in M1.2
- A better selection of the vocabulary
- A better data balance among the different language pairs, specially for domain adaptation where we observed a decrement in quality for under-represented pairs

Regarding research aspects, we will focus on:

- Experiment with new methods to obtain additional parallel sentences in the psychological domain
- Include in a more efficient way information available in BabelNet to resolve sense ambiguities and better adapt the system to our domain and multilingual setting
- Compare the current sequence to sequence architecture with the, also neural, transformer architecture

References

[1] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

[2] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation. In *Proceedings of the Second Conference on Machine Translations (WMT 2017)*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation (WMT 2016)*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, December 2017.

[5] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of simple domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada, 2017.

[6] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145, 2002.

[7] Cristina España-Bonet and Alberto Barrón-Cedeño. Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 144–149, August 2017.

[8] Cristina España-Bonet and Josef van Genabith. Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI System at IWSLT 2017. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, pages 15–22, Tokyo, Japan, December 2017.

[9] Cristina España-Bonet and Josef van Genabith. Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems. In *Proceedings of the LREC 2018 MLP-MomenT Workshop*, pages 8–13, Miyazaki, Japan, May 2018.

[10] Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350, December 2017.

[11] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. *CoRR*, abs/1606.04164, June 2016.

[12] Markus Freitag and Yaser Al-Onaizan. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*, 2016.

[13] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.

[14] Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic mt evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June 2008. The Association for Computational Linguistics.

[15] M. González, J. Giménez, and L. Màrquez. A graphical interface for MT evaluation and error analysis. In *Proc. of the 50th ACL Conference, System Demonstrations*, pages 139–144, 2012.

[16] Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the International Workshop on Spoken Language Translation*, Seattle, WA, November 2016.

[17] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. 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 Linguist*, 5:339–351, October 2017.

[18] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain, 2004.

[19] 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, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[20] Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[21] I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.

[22] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

[23] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003.

[24] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318, 2002.

[26] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: A Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 65–68, Valencia, Spain, April 2017. Association for Computational Linguistics.

[27] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, , and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231, 2006.

[28] A. Stolcke. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, 2002.

[29] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, 1997.

[30] Ádám Csaba Varga. Domain adaptation for multilingual neural machine translation. Master's thesis, Universität des Saarlandes, 2017.

[31] Zhongjun He Hua Wu Conghui Zhu Haifeng Wang Zhu, Xiaoning and Tiejun Zhao. Improving pivot-based statistical machine translation by pivoting the co-occurrence count of phrase pairs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics.