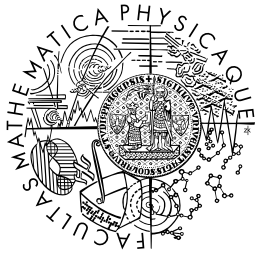

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bohdan Ihnatchenko

Multi-Target Machine Translation

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: doc. RNDr. Bojar Ondřej, Ph.D.
Study programme: Computer Science
Study branch: Artificial Intelligence

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

Dedication.

Title: Multi-Target Machine Translation

Author: Bohdan Ihnatchenko

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Bojar Ondřej, Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: Machine translation words

Contents

Introduction

With increasing availability of computational resources and enormous amount of publicly available corpora it is now possible to obtain a MT system which produces translations of acceptable quality. But in the use cases similar to conferences, where one speech is translated into multiple target languages, the same amount of models needs to be deployed. Another option is to use multilingual MT system for all needed languages together, which may lead to a decreased quality of translations.

1. Background

1.1 History of machine translation

1.2 Transformer model

1.3 Translation evaluation

2. Experiment setup

In this chapter we describe the data used for experiments, training setup and In this chapter we describe the data, training setup and experiments that were run to answer the questions asked in this thesis.

2.1 Questions and constraints

XXX TODO: Limited resources, reasonable quality is still needed
Constraints:

Translation quality for multi-lingual system is insignificantly worse than for mono-lingual one-to-one tranlsation system.

Maximum possible target languages are combined in one model.

Questions:

How *in average* adding one more randomly selected target language to the multitarget model affects its En→De performance?

How is it different if we add a linguistically similar, not randomly selected language?

How is adding one more language from the same language family or group *in average* affects tranlsation performance for selected language pair (e.g. En→De)?

2.2 Experiments

2.2.1 Starting point

In Johnson et al. [2017] autors proposed a way to build a multi-lingual machine tranlsation model without any changes to the *Transformer* architecture. In fact, the only change should be performed on the input data. To make the *Transformer* model process multi-lingual data the language tag is added to the source sentence. For example, the following En→Cz sentence pair:

Hello world! → Ahoj světe!

is modified to:

<2cs> Hello world! → Ahoj světe!

With given method it is possible to produce translations in multiple languages using the same model just by altering the prepended target language tag. It was also demonstrated that this method slightly improves translation quality for low resource languages when compared to monolingual translation model.

In this and the following (Arivazhagan et al. [2019], Aharoni et al. [2019]) papers from Google many different cases are tried and described. However, in each setting there is usually only one model of each kind considered. For example, when in [Aharoni et al., 2019] authors compare 5-to-5, 25-to-25, 50-to-50, etc. models, there is only one 5-to-5 model, one 25-to-25, etc.

2.2.2 Proposed experiments

XXX Experiments: monolingual baseline

Target language tags do not affect BLEU: Siddhant et al. [2020]. mNMT models en-to-4 and 4-to-en trained; 1) <2xx> added to the source; 2) target language encoded separately. BLEU scores are comparable using both approaches.

XXX Experiments: n-lingual baselines (random)

Multilingual models with random set of languages. The purpouse is twofold: to show BLEU score decrease with increasing number of target languages and to serve as a baseline for multitarget models with target languages grouped by in non-random way, e.g. by language group or linguistic similarity.

XXX Experiments: group by language group

If all target languages are from one language group we expect to observe better translation quality comparing to multilingual baseline results with randomly selected target languages. This is expected due to shared parts of vocabulary (todo: expand with examples) and linguistic properties (again, expand with examples). Germanic group: da, de, is, no, nl, sv. Slavic with cyrillic script: bg, mk, ru, uk. Slavic: bg, cs, hr, mk, pl, ru, sk, sl, sr, uk

XXX Experiments: group by linguistic similarity

From Siddhant et al. [2020] follows that languages' script and similarly the amount of shared vocabulary is not so important for XX→En translation direction. Example with Serbian and Croatian, with the same vocabulary but in different scripts.

2.3 Dataset(s)

2.3.1 English to 36 languages

To observe effects of linguistic similarity of target languages is is important to examine enough possible variations of those. The OPUS dataset (Tiedemann [2012]) is an open and free collection of texts covers more than 90 languages with data from several domains.[1]

For our experiments the source language is English only. Sampling and splitting of the data is the one used for ELITR project.[2] For each of language pairs and each sub-dataset the data was splitted to training, validation and testing sets. For each of the two latter sets 2000 random sentences were selected and the rest of the data remained for the training set. In cases where the sub-dataset contained less than 16000 sentence pairs no data went to the validation set. Later for each language pair there were 1000000 sentence pairs sampled from all training subsets. Firstly, if available, the sentences were taken from Europarl, then EUbooks, OpenSubtitles, and then all remaining sub-datasets. The same procedure was used to sample x000 of validation set sentences per each language pair. The test sets were left separate, so that the result on each domain would be observable.

Later an overlap in the source side of different language pairs was found. Although this would not directly lead to unfair increase of the test score, such

[1] Available at `http://opus.nlpl.eu/`

[2] `https://elitr.eu/wp-content/uploads/2019/07/D11.FINAL_.pdf`

sentence pairs were removed from the training sets. This filtering decreased the amound of sentence pairs to 0.85-0.95 millions per language pair.

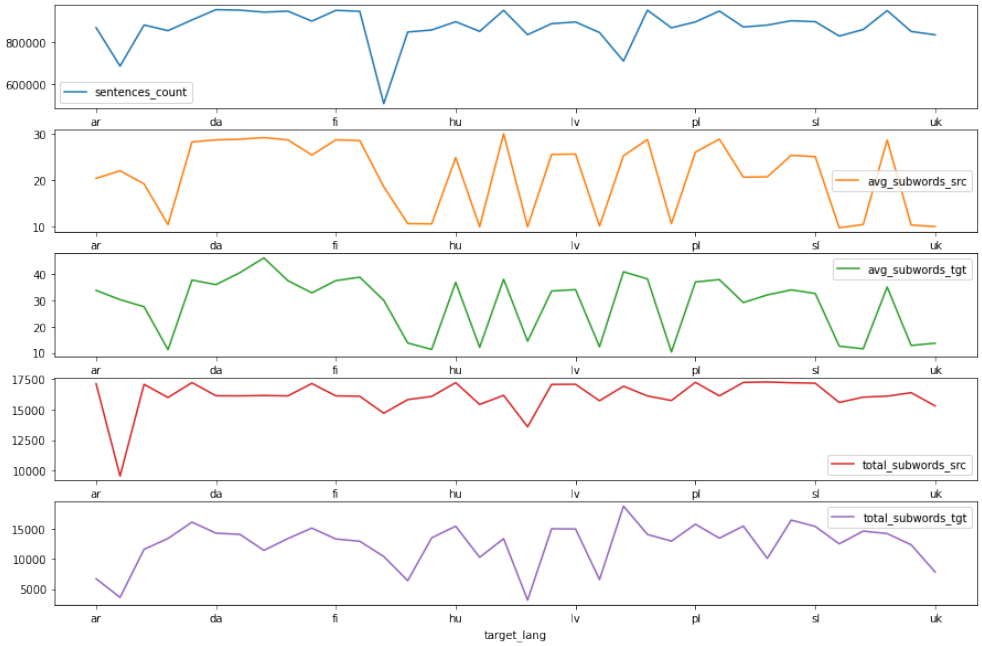

Figure 2.1: **Training data language statistics.** Languages are on the X axis sorted as in appendix. From top to bottom: total number of sentence pairs in training set per language, average amount of subwords per sentence on the source side, the same on the target side, total amount of unique subwords for this target language on the source side, the same on the target side.

To train a model on a specific subset of target languages, only related sentence pairs are subsampled. For example, to prepare data for En→{Fr, De} setup only sentences which source side starts with tags <2fr> or <2de> are selected to the training set. Development set is selected in the same way.

2.4 Training

2.4.1 Tool kits

There exists a number of different tools that can be used for training a NMT model. General purpouse deep learning programming libraries like Tensorflow[3] and PyTorch[4] are most popular for deep learning related research. With their help it is possible to construct any of today's state-of-the-art NMT models; pre-built and pre-trained models are initially present in such frameworks, but it is also possible to describe a model from scratch.

Another option is presented by specialized NMT tool kits. They usually contain efficient and tested implementations of NMT models as well as some of usefull preprocessing tools. For the experiments described in 2.2 there is a need to train significant amount of models with the same architecture and settings but

[3] `https://tensorflow.org/`
[4] `https://pytorch.org/`

different datasets. Due to that fact, in this work the use of specialized NMT tool kit is more suitable. Let us consider the foolowing list of broadly used tool kits as for year 2020, presented in Koehn [2020]: XXX Book is not yet published (expected in June)

- OpenNMT (based on Torch/pyTorch)[5]
- Sockeye (based on MXNet)[6]
- Fairseq (based on pyTorch)[7]
- Marian (stand-alone implementation in C++)[8]
- Google's Transformer (based on Tensorflow)[9]
- Tensor2Tensor (based on Tensorflow) [10]

We chose *MARIAN-NMT* tool kit[11] as a fast solution with stable and efficient *Transformer* Vaswani et al. [2017] implementation, minimum of third-side dependencies, and ability to train models on multiple GPU units in parallel.

2.4.2 Computational cluster

Many computations - cluster used.

Resourses are used by other people, disc quota is limited – parallel launching of experiments, switching to the next each 2 hours, saving only best models and the last one, removing subsampled datasets

2.4.3 Model settings

The initial parameter selection is made with respect to Popel and Bojar [2018]. First of all, the hyperparameters of MT model are tuned on couple of language pairs from one dataset. The parameters leading to the same result in shorter time were preferred. Then the selected parameters were used on all experimends with the dataset.

[5] `https://opennmt.net`
[6] `https://github.com/awslabs/sockeye`
[7] `https://github.com/pytorch/fairseq`
[8] `marian-nmt.github.io`
[9] `https://github.com/tensorflow/models/tree/master/official/transformer`
[10] `https://github.com/tensorflow/tensor2tensor`
[11] Junczys-Dowmunt et al. [2018]

3. Random choise of target languages

3.1 Overview

In this chapter we explore the effect of increasing number of target languages on the model performance in general. Multiple possible outcomes can be expected at this experiment: either performance drop due to the increased amount of languages to be processed by the model of the same size, or the opposite - performance increase due to shared knowledge gained by the model from bigger and varying dataset. Also, either of these options can be true for different target languages in different scale.

First of all, performance drop is expected. Considering that the size of the model is fixed, so is its capacity. At some moment adding more target languages should lead to the decrease in translation quality for each of every target language

3.2 Performance drop on massively multilingual setup

1-to-3, 5, 7, etc. models on en-to-36 dataset (0.9 mil. sentences per target language)

When the size of the model is fixed, adding more translation directions usually causes worsening of its performance. Multiple studies have shown this to be truth for many-to-many setup.

In Aharoni et al. [2019] models with up to 103 languages were tested. English centric in-house dataset was used to train En→Any and Any→En multilingual models. The average number of examples per language pair is 940k: for 13 out of the 102 pairs there were less than one million examples available. All languages from 5-to-5 model are present in 25-to-25, same is true for all languages from 25-to-25 with respect to 50-to-50 and so forth. In all cases they trained large Transformer model with 473.7M parameters. As can be seen on Table 3.1, the quality of translation is significantly worse when model is trained to translate more languages. However, it is worth reminding that this many-to-many experiment may have different reasons due to many-to-one direction present in it.

The decrease of model's performance with adding more target langueges is clearly shown in Aharoni et al. [2019].

3.3 Performance decrease on richer data sets

1 to 3, 4, 5 on UN corpus (much more sentence pairs per target language) Eisele and Chen [2010]

	En-Ar	En-Fr	En-Ru	En-Uk
5-to-5	**12.42**	**37.3**	**24.86**	16.48
25-to-25	11.77	36.79	23.24	**17.17**
50-to-50	11.65	35.83	21.95	15.32
75-to-75	10.69	34.35	20.7	14.59
103-to-103	10.25	34.42	19.9	13.89

Table 3.1: **BLEU scores for translation in one direction (part of Table 7 from [Aharoni et al., 2019]) .** Model trained on 5-to-5 English centric dataset (English to any and any to English) scores 12.42 BLEU for English-Arabic test set. Every language from 5 languages of 5-to-5 data set is included into 25-to-25 set, as well as every language from 25-to-25 data set is included into 50-to-50 and so forth.

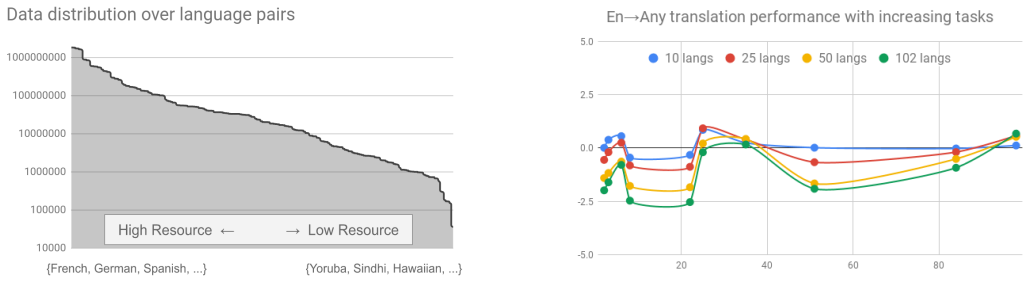

Figure 3.1: **Tranlsation performance for 102 languages from Arivazhagan et al. [2019] .** Axis X is shared between left and right plot. On axis X there are languages sorted by amount of training data. Left: amount of training data (axis Y) for a language. Right (best viewed in color): Effect of increasing the number of languages on the translation quality. On the axis X the languages are sorted the same way as on the left plot. The points visualized are 10 languages that are present in all setups from En $\leftrightarrow$ 10 to En $\leftrightarrow$ 102.

n_targets	mean	std	count
1	41.40	—	1
2	40.60	0.20	3
3	39.39	0.62	8
4	39.40	0.71	2
5	38.45	0.52	6

(a) En→Bg for *Europarl/v7* dataset.

n_targets	mean	count	std
1	19.50	1	–
2	18.88	4	0.39
3	17.45	4	0.52
4	17.80	2	0.42

(b) En→Ru for *OpenSubtitles/v2016* dataset.

Table 3.2: **BLEU score change with adding target languages.** (a) First row: for mono-lingual En→Bg model test BLEU score is 41.40. Second row: for 3 (column *count*) En→Any models with two target languages (column *n_targets*) one of which is Bulgarian the mean BLEU score is 40.60 with standard deviation 0.20. (b): same way as (a)

4. Group by language groups

4.1 Language groups

1 to 2, 3, 4, 5, etc. models on en-to-36 dataset (0.9 mil. sentences per target language) compared with random runs

4.1.1 Germanic group

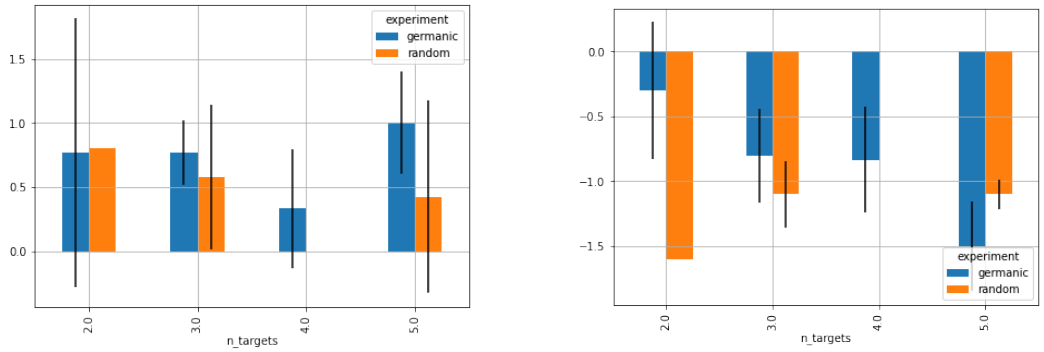

Figure 4.1: **En→De BLEU score difference: Random vs. Germanic.** On X axis - number of target languages. On Y axis - difference score comparing with monolingual BLEU. Left: OpenSubtitles/v2018, monolingual model's result 13.1 BLEU. Right: MultiUN, monolingual BLEU is 25.4.

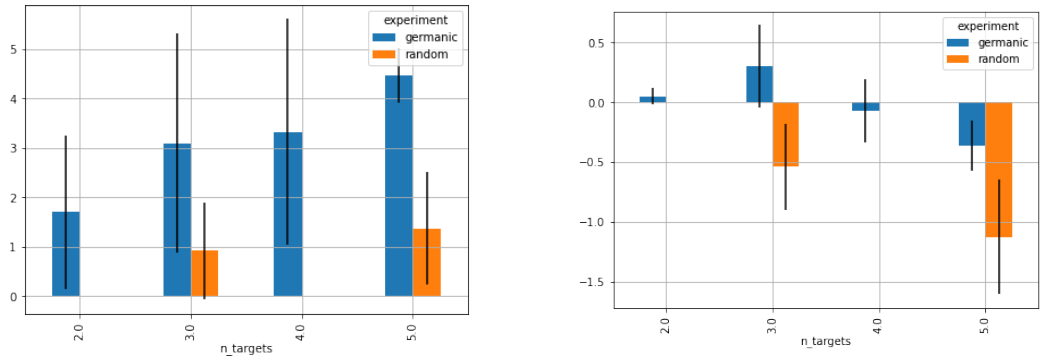

Figure 4.2: **En→Da BLEU score difference: Random vs. Germanic.** Axis are same as above. Left: OpenSubtitles/v2018, monolingual model's result 15.6 BLEU. Right: Europarl/v3, monolingual BLEU is 24.6.

4.1.2 Slavic with cyrillic script

4.2 Selecting target languages by linguistic similarity

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	14.950000	15.2625	6.0	8.0	1.209545	0.757699
3	15.383333	14.5625	6.0	8.0	0.725029	0.616297
4	15.200000	14.2750	2.0	4.0	0.282843	0.206155

(a) MultiUN/v1

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	22.883333	22.875	6.0	8.0	0.507609	0.514782
3	22.333333	21.050	6.0	8.0	0.273252	0.705084
4	21.050000	20.950	2.0	4.0	0.494975	0.465475

(b) NewsCommentary/v11

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	19.266667	18.7375	6.0	8.0	0.273252	0.396187
3	19.350000	17.7875	6.0	8.0	0.653452	0.418970
4	18.900000	17.5750	2.0	4.0	0.282843	0.613052

(c) OpenSubtitles/v2018

Table 4.1: Mean BLEU score, its standard deviation and number of trained models (count) for Russian at various datasets

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	40.433333	40.683333	6.0	6.0	0.273252	0.183485
3	39.216667	39.506250	6.0	16.0	0.318852	0.611521
4	37.750000	39.450000	2.0	4.0	0.494975	0.525991
5	NaN	38.483333	NaN	12.0	NaN	0.511386

(a) Europarl/v7

n_targets	mean		len		std	
	cyrillic	random	cyrillic	random	cyrillic	random
2	23.683333	23.250000	6.0	6.0	0.598052	0.225832
3	23.216667	22.406250	6.0	16.0	0.470815	0.619106
4	22.300000	22.600000	2.0	4.0	0.424264	0.081650
5	-	21.741667	-	12.0	-	0.494439

(b) OpenSubtitles/v2018

Table 4.2: BLEU score for Bulgarian at various datasets

n_targets	mean		count		std	
	cyrillic	random	cyrillic	random	cyrillic	random
1	41.40		1		-	
2	40.30	40.60	3	3	0.17	0.20
3	38.97	39.39	3	8	0.23	0.62
4	37.40	39.40	1	2	–	0.71
5	–	38.45	–	6	–	0.52

Table 4.3: BLEU score for bg at dataset Europarl/v7

Conclusion

Bibliography

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://www.aclweb.org/anthology/N19-1388.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. jul 2019. URL http://arxiv.org/abs/1907.05019.

Andreas Eisele and Yu Chen. {M}ulti{UN}: A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation ({LREC}'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/686{_}Paper.pdf.

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. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024.

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. *Neural Machine Translation*. Cambridge University Press, 2020.

Martin Popel and Ondřej Bojar. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110, 03 2018. doi: 10.2478/pralin-2018-0002.

Aditya Siddhant, Ankur Bapna, Henry Tsai, Jason Riesa, Karthik Raman, Melvin Johnson, Naveen Ari, and Orhan Firat. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. 2020.

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, Łukasz 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. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment