This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

Health in my Language

HimL

D1.1: Report on Building Translation Systems for Public Health Domain

Author(s): Ondřej Bojar, Barry Haddow, David Mareček, Roman Sudarikov, Aleš Tamchyna, Dušan Variš

Dissemination Level: Public

Date: February 1st 2017

Version 1.0

Grant agreement no.	644402
Project acronym	HimL
Project full title	Health in my Language
Funding Scheme	Innovation Action
Coordinator	Barry Haddow (UEDIN)
Start date, duration	1 February 2015, 36 months
Distribution	Public
Contractual date of delivery	February 1st 2017
Actual date of delivery	February 1st 2017
Deliverable number	D1.1
Deliverable title	Report on Building Translation Systems for Public Health Domain
Type	Report
Status and version	1.0
Number of pages	24
Contributing partners	CUNI, UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Ondřej Bojar, Barry Haddow, David Mareček, Roman Sudarikov, Aleš Tamchyna, Dušan Variš
EC project officer	Tünde Turbucz
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom
	Univerzita Karlova V Praze (CUNI), Czech Republic
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUENCHEN), Germany
	Lingea SRO (LINGEA), Czech Republic
	NHS 24 (Scotland) (NHS24), United Kingdom
	Cochrane (COCHRANE), United Kingdom

For copies or reports, updates on project activities and other HimL-related information, contact:

Barry Haddow bhaddow@staffmail.ed.ac.uk
University of Edinburgh Phone: +44 (0) 131 651 3173

Contents

Overview

The goal of WP1 Data and Adaptation is to collect and maintain domain-specific data and to improve the accuracy of statistical machine translation using domain adaptation techniques.

The work package is organized into three tasks: *Task 1.1 Building and maintaining training data*, *Task 1.2 Creating domain-adapted systems*, and *Task 1.3 Mining terminology from non-parallel data.* This deliverable describes the state of the first two tasks after year 2 of the HimL project. The results of Task 1.3 will be covered by Deliverable 1.2 at the end of year 3.

All tasks proceed as planned.

Details on the progress and experiments in Tasks 1.1 and 1.2 are provided in the respective Sections 1 and 2 below.

1 Building and Maintaining Data for Training MT Systems

This task runs throughout the full duration of the project, to feed other work packages with relevant training data.

We report here on two main corpus gathering exercises. The first is new release of the Czech-English parallel corpus CzEng (version 1.6), described in Section 1.1. Additionally, towards the end of year 2, we assembled a large collection of parallel texts for HimL languages and focusing on the HimL domain. This new dataset, described in Section 1.2, was not ready for year 2 experiments but it will be used in the last year of the project wherever possible to improve the comparability of experiment results across languages and work packages.

1.1 New Release of CzEng

CzEng is a parallel corpus maintained by CUNI since 2006. In the second year of HimL, a new release of CzEng was prepared and the support from the project HimL was used to specifically add texts from the medical domain.

CzEng has been released in two sub-versions: a pre-release without linguistic annotation[1], just in time for the WMT16 translation task, and the final automatically annotated and better filtered release version 1.6.[2] The new release represents a large increase in corpus size, from about 15 million sentence pairs in CzEng 1.0, to over 50 million in CzEng 1.6.

More details, including details on the medical section, are available in the paper by Bojar *et al.* (2016a).

1.2 HimLCorpus

The HimLCorpus (version 1.0) is a collection of parallel data for HimL languages, collected with focus on the medical domain. Collected corpora are sentence aligned and non-tokenized if the tokenization was not already performed by the source. Table 1 summarizes the number of parallel segments in the resulting corpus for each of the sources.

The data was collected mainly from the OPUS[3] (Tiedemann, 2009) website and the Khresmoi project[4], so we refer the reader to these two sources for more detailed data descriptions, and focus here on new data gathered for HimL. In particular, we gathered additional data by performing a new crawl of documents available on the European Medicines Agency (EMA) website[5], previously known as European Agency for the Evaluation of Medicinal Products (EMEA).[6] The process of data extraction is described in the following section.

We gathered following parallel data:

- **Cordis** - The CORDIS news database sentence-aligned with the mAligna aligner using the Church & Gale algorithm. Original texts were crawled from the CORDIS website.
- **DBpedia** - Large multi-domain ontology which has been derived from Wikipedia.
- **ECDC translation memory**[7] - Translation memory produced by the European Centre for Disease Prevention and Control. Collection of health-related documents with professional translations into 25 languages.

[1] http://ufal.mff.cuni.cz/czeng/czeng16pre
[2] http://ufal.mff.cuni.cz/czeng/
[3] http://opus.lingfil.uu.se/
[4] http://khresmoi.eu/resources/data-sets/
[5] http://www.ema.europa.eu/ema
[6] Even though the organization name was changed we will use the old EMEA abbreviation in the span of this report.
[7] http://ipsc.jrc.ec.europa.eu/?id=782

Corpora	**cs-en**	**de-en**	**pl-en**	**ro-en**
Cordis	-	-	175,531	-
ECDC	2,324	2,379	2,202	2,363
EMEA (merged-uniq)	807,395	780,971	795,648	743,741
EMEA (old crawl from OPUS)	*1,051,462*	*1,106,373*	*1,044,864*	*992,790*
EMEA (new crawl)	*1,308,338*	*2,394,544*	*1,209,565*	*1,144,810*
EUbookshop	455,472	9,333,066	539,941	324,553
EUROPARL	645,795	1,954,622	629,549	399,037
JRC-Acquis	1,273,092	719,071	1,610,513	455,168
Medical Web Texts from CzEng 1.6	7,029	-	-	-
MuchMore	-	33,318	-	-
MultiUN	-	168,734	-	-
News Commentary	191,432	242,770	-	-
OpenSubtitles	61,799,474	15,557,228	50,610,379	79,972,303
PatTR Medical	-	1,848,303	-	-
PatTR Other	-	9,320,237	-	-
Rapid	-	-	144,091	-
Subtitles	3,143	85,326	3,044	133,428
Total Parallel Segments	65,185,156	40,012,707	54,510,898	82,030,593
Total Words (target language/en)	373M/450M	894M/857M	316M/397M	500M/533M
Dictionaries	**cs-en**	**de-en**	**pl-en**	**ro-en**
DBpedia	148,181	681,494	549,600	-
MeSH	20,084	24,394	-	-
UMLS Metathesaurus	1,640,448	2,326,035	731,196	-
Total Entries	1,808,713	3,007,529	1,280,796	-

Table 1: Summary of parallel data collected for HimL languages into the new HimlCorpus. We report number of parallel segments per source for each of the HimL languages. Total Parallel Segments and Total Words contain sum over all datasets (excluding the two versions of EMEA before merging, in italics).

- **EMEA corpus (old crawl from OPUS)**[8] - Parallel corpus composed of documents from the European Medicines Agency, as released in the open-source corpus OPUS.
- **EMEA corpus (new crawl)** - Parallel corpus created by crawling the European Medicines Agency[9] document database. Corpus acquisition was done using our set of tools, see Section 1.3 below.
- **EMEA corpus (merged-uniq)** - Concatenation of EMEA corpus and EMEA corpus (new crawl). To gather a union without many duplicates, the resulting corpus was sorted and in case of duplicate sentence pairs, only the first occurrence was kept. The original corpora contain large number of duplicates themselves, but we do not eliminate them in their case. Therefore, the resulting size of the merged corpus is lower than the sizes the original corpora.
- **EUbookshop**[10] - Corpus of documents from the EU bookshop.
- **EUROPARL**[11] - Parallel corpus extracted from the proceedings of the European Parliament.
- **JRC-Acquis**[12] - Parallel corpus extracted from Acquis Communautaire, the total body of European Union law applicable in its member states, by the Language Technology group of the European Commission's Joint Research Centre.
- **Medical Web Texts from CzEng 1.6**[13] - Parallel sentences collected in CzEng 1.6 excluding the EMEA data. Duplicate sentence pairs were dropped during preprocessing.
- **MeSH**[14] - Medical Subject Headings thesaurus provided by U.S. National Library of Medicine.

[8] http://opus.lingfil.uu.se/EMEA.php
[9] http://www.ema.europa.eu/ema
[10] http://opus.lingfil.uu.se/EUbookshop.php
[11] http://www.statmt.org/europarl/
[12] http://www.jrc.it/langtech
[13] http://ufal.mff.cuni.cz/czeng/czeng16pre
[14] https://www.nlm.nih.gov/mesh/

- **MuchMore Springer Bilingual corpus**[15] - Parallel corpus of English-German scientific medical abstracts obtained from the Springer web site. The corpus was aligned on the sentence level.
- **MultiUN**[16] - Parallel corpus that was extracted from the United Nations Website, cleaned and converted to XML.
- **News Commentary**[17] - Parallel corpora made available for the WMT 16 translation task.
- **OpenSubtitles**[18] - A collection of documents from the OpenSubtitles website.
- **PatTR Medical**[19] - A sentence-parallel corpus extracted from the MAREC patent collection. In-domain medical data created by extracting documents with relevant identifiers from the corpus.
- **PatTR Other** - A sentence-parallel corpus extracted from the MAREC patent collection. Out-of-domain data created as a complement of PatTR Medical.
- **Rapid** - The RAPID press releases of the EU sentence-aligned with the mAligna aligner using the Church & Gale algorithm. Original texts were crawled from from the European Commission Press releases database.[20]
- **Subtitles** - Subtitle files downloaded from the major subtitle servers via Subliminal[21] search and download library.
- **UMLS Metathesaurus**[22] - Very large, multi-purpose, and multi-lingual vocabulary database.

1.3 Crawling EMEA

From our experience when building the new release of CzEng (Section 1.1), we knew that the European Medicines Agency has changed the set of multilingual documents they provide. Some older documents are no longer available online and new documents were created.

The new documents were gathered in this process:

1. We use the EMEA document search engine[23] to collect relative paths of documents within the database. We search either using a document reference number known from the previous crawls or using a keyword from a list of most frequent words in the previous version of the corpus.
2. Next, we download the PDF files for all the search results from the EMEA database. We use language codes to ask for specific language version of each of the downloaded document.
3. Every downloaded PDF file is then transformed to plain text using PDFMiner.[24]
4. Plain texts are split into sentences using NLTK[25] sentence splitter from the punkt package.[26]
5. We create sentence-level alignment for each transformed document pair using hunalign.[27] For better alignment quality, we provide hunalign with word-to-word translation dictionary extracted automatically from word-aligned Europarl corpora for each language pair.

To get the largest possible collection of relevant texts, we eventually combine the older OPUS crawl of EMEA with our new crawl and de-duplicate it at the level of sentences.

[15] http://muchmore.dfki.de/resources1.htm
[16] http://www.euromatrixplus.net/multi-un/
[17] http://www.statmt.org/wmt16
[18] http://opus.lingfil.uu.se/OpenSubtitles.php
[19] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/
[20] http://europa.eu/rapid/search.htm
[21] https://github.com/Diaoul/subliminal
[22] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[23] http://www.ema.europa.eu/ema/index.jsp?curl=pages%2Fdocument_library%2Flanding%2Fdocument_library_search.jsp&
[24] http://www.unixuser.org/~euske/python/pdfminer/
[25] http://www.nltk.org/
[26] http://www.nltk.org/_modules/nltk/tokenize/punkt.html
[27] http://mokk.bme.hu/resources/hunalign/

1.4 HimLCorpus vs. CzEng 1.6 for Czech

From now on, HimLCorpus should, for all HimL languages and the medical domain, serve as the only source of parallel data.

This deliverable mentions two new resources for Czech-English data: HimLCorpus and CzEng 1.6. CzEng 1.6 is meant as a general-purpose corpus and medical data are only a part of it. As domain adaptation experiments in Section 2.2 below confirm, domain-only data of this size are sufficient to obtain comparable or better results than by employing full CzEng. Therefore, we specifically included only the medical sections of CzEng in HimLCorpus.

We also included OpenSubtitles corpora in HimLCorpus even though they cover a general-purpose domain. These corpora are available across all HimL languages and should serve as a comparison between these languages.

For Czech, as for other languages, we will thus use HimLCorpus for year 3 experiments.

1.5 Availability of HimLCorpus

Due to copyright concerns, we still do not distribute HimLCorpus release version 1.0 publicly. The corpus is available from the consortium upon request but we are considering to release at least some parts of it.

2 Creating Adapted SMT Systems

In this section we discuss the approaches to domain adaptation tried in HimL and present experiment results. Our work consists of three more or less independent strands: a targeted probe for the hybrid MT system used as Czech year 2 translation system (excluding the final automatic grammar correction component) is described in Section 2.2, adaptation for all HimL languages in a phrase-based setup then carefully examined in Section 2.3 and finally Section 2.4 evaluates adaptation techniques relevant for neural machine translation.

We start with a quick overview of domain adaptation for phrase-based MT in Section 2.1. Section 2.2 makes use of some of these techniques and Section 2.3 then systematically looks at all of them.

2.1 Overview of Domain Adaptation Techniques

In the experiments in Sections 2.2 and 2.3, we focused on adaptation of the two main models used in a phrase-based system – the language model and the translation model. Adaptation of the other models used in a phrase-based system has received some attention in prior work (for example the reordering model (Chen *et al.*, 2013a), the alignment model (Cuong and Sima'an, 2015) and the operation sequence model (Durrani *et al.*, 2015)). However DA results on these other models are much scarcer, and we believe that because the language and translation models have the largest effect on translation quality, they should be the most important to consider for adaptation.

2.1.1 Language Model Adaptation

The baseline approach is to create one large language model using the target side of the parallel data, plus any additional available monolingual data in the target language. We use KenLM (Heafield *et al.*, 2013) to train 5-gram language models, with Kneser-Ney smoothing, and by default we prune all singletons of order 3 and above.

To adapt the language model, the overall approach is to split the data into sections, train separate models on each of these corpora, and then combine these models using some form of *interpolation* to create a new language model. We optimise the weights of the interpolation for our domain, with the aim of improving the translation quality of a phrase-based system built with the interpolated language model. There are two main decision points in building interpolated language models: how to split the data into corpora, and how to interpolate the constituent models.

To split the data into individual corpora, we use the original corpus boundaries of the data, and experiment with two different approaches. Either we train separate language models on each of the consituent corpora, or we group the corpora into "in-domain" and "out-of-domain" corpora, and train just two language models. The idea of these larger groupings is that splitting to constituent corpora gives many small corpora, some of which are too small for good LM estimation, and making training of interpolation weights more difficult.

The two approaches we use to interpolation are linear and log-linear. Linear interpolation uses the Moses wrapper to the SRILM implementation of the perplexity minimisation algorithm, to minimise perplexity on a heldout set (we use the tuning set). There are known issues with this method: (i) the Moses wrapper gives different results depending on the order of interpolation; and (ii) it is not possible to represent a linear interpolation as a single ARPA file, and even though SRILM produces a single ARPA

file, it is not a correct interpolation (Heafield *et al.*, 2016). Despite these problems, this method of linear interpolation has given good results in the past, so we employ it here. For log-linear interpolation, we consider each LM is a separate feature in the MT system, and tune these feature weights along with all the other weights of the system, to maximise BLEU. This is not true log-linear interpolation (as implemented in (Heafield *et al.*, 2016)), since we do not create a normalised probability distribution when combining the LMs, however we have the advantage of that we can optimise directly for a measure of translation quality.

From Edinburgh's work on building systems for WMT (Huck *et al.*, 2016; Williams *et al.*, 2016) we noted two additional adaptation techniques. Firstly, when doing log-linear interpolation, we additionally include a monolithic LM trained on all available data in the interpolation. And secondly, irrespective of the type of interpolation, we experimented with adding an additional, unpruned, LM, trained on only in-domain data. The idea behind using an unpruned in-domain LM is that the rare 3,4 and 5-grams in an in-domain LM can be domain-relevant terms and we don't want the LM to "filter" them out just because they have not been seen before.

2.1.2 Translation Model Adaptation

For the translation model, the baseline approach is again to train a single model (a TM, in this case) using all the available data, applying Good-Turing smoothing. The techniques that we use for adaptation of the TM involve either modifying the scoring of phrase-pairs to better reflect the domain, or removing out-of-domain phrase-pairs.

The first adaptation technique is the simplest. For each phrase-pair in the TM, we add a set of binary "domain indicator" or "provenance" features, one for each corpus in our data set. Each feature is switched on whenever a the phrase-pair is extracted from the corresponding corpus. The feature weights are tuned in the usual way (i.e. using MERT or kbmira), along with all the other features in the translation system.

We also consider two other domain adaptation techniques that had pre-existing implementations in Moses. Linear interpolation of the translation model is similar to linear interpolation of the language model. We build a single aligned corpus on the concatenation of all parallel corpora, then from this we build separate translation models on each of the consitituent corpora. The interpolation weights are optimised by minimising perplexity on a heldout set (the tuning set) where perplexity is calculated using the phrases extracted from the heldout set using the standard procedure. We use the Moses implementation accompanying Sennrich (2012), with the `--normalized` argument.

The other domain adaptation technique we use from Moses is modified Moore-Lewis filtering (Axelrod *et al.*, 2011). The idea of this is to filter the training parallel texts, retaining those sentences which are most representative of the domain. The filtering is based on cross-entropy scores calculated using "in-domain" and "general-domain" language models. For the in-domain models, we created English ones using crawled data from NHS 24 and Cochrane websites, and target-language models using the target side of the in-domain training data. For general domain data we use a random sample from all training data. MML filtering is run after word alignment, retaining a specified percentage of the training set to build the translation model.

The final domain adaptation technique that we applied is the *vector space model* (Chen *et al.*, 2013b). The VSM attaches feature(s) to each phrase-pair in the translation model to indicate how typical of the domain it is. To calculate the value of the VSM feature(s) we first need to calculate a "domain vector" for the phrase-pair, made up of the *tf-idf* weights of the phrase-pair in each of the constituent corpora. We implemented two variants of the VSM model. In the "plain" VSM model, we also calculate a domain vector for the tuning set, by running phrase extraction across that set, and adding up the domain vectors for all phrase-pairs in tuning. The VSM feature for a phrase-pair is then the similarity (as measured by the Bhattacharyya Coefficient – BC) between the phrase-pair's domain vector, and the domain vector of the tuning set. In later work (Chen *et al.*, 2014) , the "distributional" VSM was proposed, where instead of calculating a similarity score, the whole domain vector is included in the feature set attached to the phrase-pair. The distributional VSM model can thus be seen as an extension of the provenance feature (above), where the feature value indicates the domain specificity of the phrase-pair.

2.2 Domain Adaptation of Czech Year 2 System

In contrast to other HimL systems, the year 2 system for Czech is a rather complex hybrid setup: a transfer-based system TectoMT followed by a fine-tuned phrase-based translation and complemented with a final automatic correction of grammatical errors. The phrase-based component serves two purposes: bring in knowledge from large parallel corpora and merge it seamlessly with the translation proposed by the transfer-based system.

While the transfer-based system TectoMT can be in principle adapted to a particular domain, such an attempt would go beyond the scope of the HimL project. (For example, it would involve evaluation of parser performance on the medical domain, demanding manual syntactic analysis of NHS24 and Cochrane sentences.) We thus focus on the adaptation of the phrase-based component, but, in contrast to the focus on pure phrase-based MT in Section 2.3, here we evaluate the adapted system with the transfer-based component in place.

The last component of HimL year 2 English-to-Czech pipeline improves grammatical correctness of the output by specifically fixing known frequent errors. It does not address lexical errors, so we do not expect any domain effects here and exclude this final component from our adaptation experiments.

2.2.1 Adaptation by data selection

Our experiments are focused on translation model adaptation, as outlined in Section 2.1.2 above.

We chose XenC tool (Rousseau, 2013) to extract domain-specific data from out-of-domain corpora for English-to-Czech SMT systems. We used two modes provided by XenC to filter out-of-domain corpus. Both of these modes estimate two language models from an in-domain corpus and out-of-domain corpora, using SRILM toolkit. The first mode is a filtering process based on a simple perplexity computation and the second mode is based on the bilingual cross-entropy difference as described in (Axelrod *et al.*, 2011), the same one as is mentioned in Section 2.1.2.

In the following experiments the training part of CzEng 1.6 (Bojar *et al.*, 2016b) was used as out-of-domain corpus. The XenC tool required prior domain-specific data to be used for scoring out-of-domain sentences. For this we used medical section of CzEng 1.6, which contains 1.5 million sentences, gathered from parallel health-related web sites and also by re-crawling EMEA (European Medicines Agency). We have experimented with five different corpora:

- CzEng16Pre – pre-release version of CzEng1.6 corpus, used for WMT16 (Bojar *et al.*, 2016c)
- CzEngMed – medical section of CzEng 1.6, 1.5 million sentences
- CzEngTop1 – top 1% from CzEng 1.6 scored using XenC with CzEngMed as domain-specific data, 1 million sentences
- CzEngTop5 – top 5% from CzEng 1.6 scored using XenC with CzEngMed as domain-specific data, 4.5 million sentences
- CzEngTop10 – top 10% from CzEng 1.6 scored using XenC with CzEngMed as domain-specific data, 9.8 million sentences

2.2.2 SMT Systems Configuration

All systems presented in Section 2.2.3 are configured the same way as the WMT16 English-Czech submission (Tamchyna *et al.*, 2016). They used the following setup:

- Phrase table source-side factors are true-cased word form + a refined word form which includes POS tag for short and frequent word forms, allowing for a better disambiguation.
- Phrase table target-side factors are true-cased word form + lemma + morphological tag
- Language models are (separately) for true-cased word forms, lemmas and morphological tags
- Operation sequence model with 5 features is based on true-cased word forms (refined by POS tag again for short words) on the source side and maps to true-cased word forms on target side

For setups with more than one phrase table, we used log-linear interpolation with weights set during MERT tuning step with both translation tables.

2.2.3 Using Only Domain-Specific Data to Create SMT System

We used the medical section of CzEng 1.6 (CzEngMed), described in Section 2.2, as a source for extraction of phrase tables for experiments with only domain-specific data.

We experimented with the following setups, summarized in Table 2:

- Baseline – one phrase table from CzEng 1.6Pre, another phrase table from TectoMT, language models use monolingual news corpora made available by WMT16, operation sequence model uses CzEng 1.6Pre, same as baseline in Section 2.2
- Setup 0 – one phrase table from CzEngMed, another phrase table from TectoMT, language models use CzEngMed, operation sequence model uses CzEng 1.6Pre
- Setup 1 – one phrase table from CzEngMed, another phrase table from TectoMT, language models use monolingual news corpora from WMT16, operation sequence model uses CzEng 1.6Pre

Setup	TM	Mononews LM (form+lemma+tag)	CzEngMed LM (form)	CzEngMed LM (lemma+tag)	OSM
Baseline	CzEng1.6Pre	+	-	-	+
Setup 0	CzEngMed	-	+	+	+
Setup 1	CzEngMed	+	-	-	+
Setup 2	CzEngTop1	+	-	-	+
Setup 3	CzEngTop5	+	-	-	+
Setup 4	CzEngTop10	+	-	-	+
Setup 5	CzEngMed+CzEngTop5	+	-	-	+
Setup 6	CzEngTop5	+	+	-	+

Table 2: SMT setups

- Setup 2 – one phrase table from CzEngTop1, another phrase table from TectoMT, language models use monolingual news corpora from WMT16, operation sequence model uses CzEng 1.6Pre
- Setup 3 – one phrase table from CzEngTop5, another phrase table from TectoMT, language models use monolingual news corpora from WMT16, operation sequence model uses CzEng 1.6Pre
- Setup 4 – one phrase table from CzEngTop10, another phrase table from TectoMT, language models use monolingual news corpora from WMT16, operation sequence model uses CzEng 1.6Pre
- Setup 5 – one phrase table from CzEngTop5, another phrase table from CzEngMed, a third phrase table from TectoMT, language models use monolingual news corpora from WMT16, operation sequence model uses CzEng 1.6Pre
- Setup 6 – one phrase table from CzEngTop5, another phrase table from TectoMT, language models use monolingual news corpora from WMT16 and CzEngMed (forms only), operation sequence model uses CzEng 1.6Pre

Across the experiments, we use in-domain development and test corpora (Cochrane + NHS24) for MERT tuning and evaluation. The baseline system with which we compare our setups is identical to the year 2 system reported in D4.2/5 (not accessible to general public), except that we do not use the final post-processing by DepFix (Rosa *et al.*, 2012) here.

The results presented in Table 3 show that systems built from domain-specific data can perform nearly as well as systems using the huge out-of-domain corpus, see Setups 0 to 6.

Setup	BLEU	Avg.BLEU
Baseline	24.83 [23.81, 25.90]	24.5
Setup 0	24.32 [23.34, 25.35]	24.1
Setup 1	23.97 [22.93, 24.99]	-
Setup 2	22.58 [21.51, 23.67]	-
Setup 3	24.52 [23.40, 25.66]	24.5
Setup 4	24.78 [23.71, 25.88]	24.5
Setup 5	24.68 [23.60, 25.72]	24.2
Setup 6	25.57 [24.46, 26.62]	25.4

Table 3: SMT using phrase tables extracted from domain-specific data. The first column reports BLEU for a single run with confidence intervals established by bootstrapping sentences from the test set. The second column reports the average BLEU of 4 different MERT runs.

The best system (Setup 6) uses domain adaptation for both phrase tables and the language model and improves over the baseline. Technically, this is again quite a complex setup: the medical texts from CzEng served not only as the basis to select similar sentences from the whole CzEng (the resulting selection is called CzEngTop5, as described above) but they were also directly included in the setup as a secondary language model.

2.3 Domain Adaptation in Phrase-based MT: Comparison Across All HimL Languages

In this section we explore several options for domain adaptation (DA) of a phrase-based MT baseline, applying the results to all HimL languages, and showing results on HimL test sets. We provide some analysis of how each of the DA techniques affects the translation of in-domain terms and *n*-grams.

2.3.1 Experimental Setup

Our experimental setup is based on Moses (Koehn *et al.*, 2007), and for the baseline systems we draw on experience gained in building systems for the WMT shared tasks (Haddow *et al.*, 2015; Williams *et al.*, 2016). Our MT system scores translations using a linear combination of the following features, where the weights are optimised using *k*-best MIRA (Cherry and Foster, 2012).

Translation model We align the parallel data using `fast_align`, symmetrise and extract and score phrases using the standard Moses heuristics. We use forward and backward phrase probabilities and lexical weights, and also binned frequency features. Phrase probabilities are smoothed with Good-Turing.

Language model The baseline is to use the target side of all parallel data, plus any available monolingual data, to train a 5-gram language model with KenLM, applying Kneser-Ney smoothing, and pruning singleton *n*-grams with order 3 and above.

Reordering model The lexicalised hierarchical reordering model classifies forward and backward reorderings as monotone, swap, discontinuous left and right, giving 8 features.

Operation sequence model We use the OSM with 4 count-based features.

Discrete features We add a word penalty, a phrase penalty and a distance-based reordering score.

The data sets we use for training our experiments are drawn from data released for WMT14 medical and WMT15 shared tasks (Dušek *et al.*, 2014; Bojar *et al.*, 2015), the OPUS corpus collection (Tiedemann, 2009), along with some HimL-specific updates. The updates for HimL are:

- A re-crawl of the MuchMore corpus of Springer abstracts (used in WMT14 medical), to fix missing umlauts.
- A re-extraction of the UMLS (also from WMT14 medical) to include new data, and Polish.
- A small (approx. 10,000 sentence) parallel English-German corpus from Cochrane

The statistics (sentence counts) for the training data are shown in Table 4. We class as "in-domain" any data drawn from a medical source, such as UMLS or EMEA (European drug information leaflets).

Name	cs	de	pl	ro
in-domain parallel	2,273,462	3,871,798	931,012	235,636
out-of-domain parallel	85,346,900	39,734,843	57,003,892	83,778,500
in-domain mono	947,708	4,118,619	941,007	240,283
out-of-domain mono	130,722,475	159,335,262	57,179,032	83,895,886

Table 4: Sentence counts in data sets used for domain adaptation experiments. We show the numbers of parallel sentences (after filtering out long sentences) and the number of sentences available for language modelling Normally the latter includes the target side of the parallel data, except for German and Czech UMLS where a much smaller corpus of sentences was used for the LM, as opposed to using the parallel side of the term dictionary.

For tuning and testing we use the HimL test sets, as described in *D5.1: Test Sets for HimL Languages*. The statistics of these sets are shown in Table 5. For tuning, we used a concatenation of the Cochrane and NHS 24 tuning sets, whereas for testing we report results on both sets separately.

	Cochrane		NHS 24	
Section	Sentences	Words	Sentences	Words
Tune	760	15492	1201	14440
Test	673	14446	1258	15758

Table 5: Sentences and word counts in HimL tune and test sets. The word counts are for untokenised English text.

2.3.2 Experiments

The first set of experiments observes the effect of language model adaptation. The baseline system uses a single LM trained on all the available data, i.e. the target side of the parallel data and any available extra monolingual data. We then split the LM data into "in-domain" and "out-of-domain" segments before training an LM on each and interpolating them, either linearly or log-linearly. Next we split the data into all constituent corpora to train LMs on each (combining some corpora that were too small for LM estimation) and again compared the two different interpolation methods. Finally we add an in-domain unpruned LM to the combinations of all LMs. The results are shown in Table 6. We thus end up with three different ways of splitting the

corpus (none, in/out and each), two different interpolation methods (linear and log-linear) and an optional in-domain unpruned LM.

Name	Model			en-cs		en-de		en-pl		en-ro	
	Split	Interp.	+unpruned LM	Coch	NHS24	Coch	NHS24	Coch	NHS24	Coch	NHS24
lm-all	none	n/a	no	27.4	21.3	37.4	29.2	15.6	21.7	34.4	29.3
lm-inout-lin	in/out	linear	no	27.8	21.4	37.9	30.9	16.7	23.9	36.1	31.6
lm-inout-log	in/out	log-lin	no	26.0	21.2	37.4	31.3	**16.8**	24.0	35.5	31.5
lm-each-lin	each	linear	no	26.9	21.2	39.1	32.0	16.6	23.2	36.7	32.0
lm-each-log	each	log-lin	no	27.8	**22.5**	39.4	32.9	16.5	24.2	36.9	31.9
lm-each-lin-unpr	each	linear	yes	27.5	21.0	39.1	32.6	16.7	**24.8**	**37.0**	**32.5**
lm-each-log-unpr	each	log-lin	yes	**28.2**	22.3	**39.5**	**33.1**	16.5	**24.8**	**37.0**	31.9

Table 6: Comparison of LM adaptation strategies. Case-sensitive BLEU scores.

The results of Table 6 show a mixed picture, but we do find some general trends emerging, and the observed differences in BLEU scores indicate that it is important to consider LM adaptation. Overall, using a single monolithic LM is a bad idea, and the best results are usually achieved by splitting the language model data into individual corpora, then interpolating. This is especially true for en-de. It is unclear whether linear or log-linear interpolation is better, but we note that the former seems better for en-ro and en-pl, whereas the latter is better for en-de and en-cs. Most (6/8) of the best performances are achieved when using an upruned in-domain language model, and in the other two cases performance is not far behind the best, so we conclude that this type of LM is a good idea.

We now turn our attention to TM adaptation. For the baseline, we choose the best-performing of the two systems which incorporate an unpruned in-domain LM, i.e. we choose a system from the last two lines of Table 6. We then consider the adaptation techniques proposed in Section 2.1.2. In particular, we consider provenance features, interpolated translation model, two variants of VSM (vector space model), and MML (modified Moore-Lewis) selection, retaining 20% of training data. We consider each technique individually, partly for clarity, and partly because some feature combinations do not have clear interpretations. The results of the comparison are shown in Table 7 (note that TM interpolation is missing for en-de due to the interpolation requiring too much RAM).

Name	Description	en-cs		en-de		en-pl		en-ro	
		Coch	NHS24	Coch	NHS24	Coch	NHS24	Coch	NHS24
base	Baseline (Best from Table 6)	28.2	22.3	39.5	33.1	16.7	24.8	37.0	**32.5**
prov	+ Provenance features	28.0	22.4	39.4	33.2	16.9	24.9	36.9	31.9
interp-n	+ Linear TM interpolation	**28.7**	22.3	–	–	**17.0**	**25.4**	36.8	31.9
vsm	+VSM	27.5	22.5	**39.6**	33.3	16.7	24.8	36.9	31.9
vsm-dist	+Distributional VSM	27.7	22.5	39.4	**33.5**	16.7	24.7	36.8	31.8
mml20	+Modified Moore-Lewis (20%)	27.9	**22.7**	39.4	33.0	16.6	24.8	**37.1**	32.0

Table 7: Comparison of TM adaptation strategies. Case-sensitive BLEU scores.

Looking at the results in Table 7, we see that TM adaptation may contribute little or nothing over LM adaptation. This is in line with the general picture for the literature on the topic, where papers on individual techniques often demonstrate gains (on specific data sets) but no techniques have gained significant traction (unlike LM interpolation). Our conclusion is that it is worthwhile experimenting with TM adaptation, especially as open-source implementations of the important techniques mostly exist, but that results are dependent on language pair and data set.

2.3.3 Analysis

In order to see beyond the BLEU scores, we tried to investigate how well each system translated domain-specific, as compared to domain-general terms. In order to do this, we extracted "terms" from the test sets, devised a measure of the "domain-ness" (treating NHS 24 and Cochrane as separate "domains") and correlated it with accuracy of translation. In the following paragraphs, we explain in more detail how this analysis works.

Extraction of Terms We wish to break down the source sentences into units that can be assessed and scored independently. Such a breakdown will always be an approximation, since it is not really possible to translate sub-parts of the sentences inde-

pendently, and the segmentation is not necessarily preserved by translation. However we hope that by averaging over many segments we will be able to pick out general patterns in translation performance.

After experimenting with *n*-gram-based term extraction, and term extraction from a parse, we decided to use the output of a chunker to pick out terms. Specifically we use TreeTagger for chunking, and extract terms using the following heuristics:

- Consider noun (NC), verb (VC) and adjectival chunks (AD).
- Using part-of-speech tags, remove leading determiners, prepositions, numbers and wh-words.
- Removing leading auxilliary verbs, brackets and negation.

Using these heuristics results in a list of around 2500 terms for each of the domains (NHS24 and Cochrane).

Domain-ness of Terms In order to measure the domain-ness of the source-sentence units, we used the log-likelihood ratio of in-domain and general-domain language models. For the in-domain language models we trained LMs separately on crawls of the Cochrane and NHS 24 websites, and to train the general domain LM we used a random sample from the English side of the large mixed training set from Section 2.3.1, with the vocabulary from the in-domain LM. We trained unpruned LMs for all orders up to 4, and normalised the log-likelihood ratios by length to enable comparison of terms of different length.

Measurement of Translation Precision To measure how well a given source segment is translated, we project it to the translation hypothesis and compare with the reference. The projection uses the word alignment output by the decoder during translation. This alignment is a combination of the phrase-phrase alignment from the hypothesis's derivation, coupled with the phrase-internal (word-word) alignment assigned to each phrase-pair during extraction (the most frequent alignment observed in the training corpus).

The source segment is projected to the hypothesis using this alignment, to obtain a corresponding hypothesis segment. If the resulting hypothesis segment is discontinuous, we treat it as a continuous segment and do not include intervening tokens, so clearly introducing noise but this is unavoidable. We count the number of times that this projected segment occurs in the hypothesis, and the number of times it occurs in the reference and record these counts, clipping the reference count if it exceeds the hypothesis count. To measure the precision of a given set of source segments, we just divide the total of the clipped reference counts, by the total hypothesis counts for the whole set.

Most Common Errors In the tables that follow we list the most commonly mis-translated terms, using the measures of domain-ness and precision detailed above. To select these terms, we consider each domain and target language separately, choosing terms which occur at least 5 times in the test set, and that have a translation precision of less than 0.1. We then rank these terms by domain-ness and show the 10 most in-domain terms in the tables below.

Domain: cochrane Language: Czech

Term	Count	Score	Precision	Translations
rotational thromboelastometry	5	2.590	0.000	rotační thromboelastometry (5)
Cochrane Central Register	6	2.329	0.000	Cochrane centrálního registru (6)
RCTs	28	2.295	0.071	randomizovaných klinických (13), RCTs (8), studie (7)
adult trauma patients	7	1.859	0.000	dospělých pacientů trauma (3), trauma dospělých pacientů (2), dospělých pacientů traumaty (2)
outcome measures	5	1.558	0.000	výsledek opatření (2), opatření výsledek (1), výsledku opatření (1)
Chinese herbal medicines	5	1.440	0.000	čínské rostlinné léčivé přípravky (2), čínských bylinných přípravků (1), čínské bylinné přípravky (1)
alpha blocker treatment	5	1.414	0.000	alfablokátory léčba (1), alfa-blokátory léčbě (1), léčby alfa-blokátorů (1)
abdominal drainage	5	1.381	0.000	břišní drenážní (5)
mortality	5	1.331	0.000	úmrtnosti (4), úmrtnost (1)
controlled trials	11	1.317	0.000	kontrolovaných klinických studiích (7), kontrolovaných klinických studií (3), kontrolovaných studiích (1)

Domain: cochrane Language: German

Term	Count	Score	Precision	Translations
review authors	13	2.870	0.000	Review Autoren (13)
rotational thromboelastometry	5	2.590	0.000	thromboelastometry (3), rotatorische thromboelastometry (2)
unclear risk	5	2.490	0.000	unklares Risiko für (3), unklares Risiko (2)
developing lymphoedema	5	1.915	0.000	Entwicklung Lymphödem (3), Entwicklung Lymphödeme (1), Lymphödem entwickeln (1)
adult trauma patients	7	1.859	0.000	erwachsenen Patienten (7)
outcome measures	5	1.558	0.000	Endpunkte (3), Ergebnis Maßnahmen (1), Outputmaße (1)
Chinese herbal medicines	5	1.440	0.000	chinesische pflanzliche Arzneimittel (3), chinesischen pflanzliche Arzneimittel (2)
alpha blocker treatment	5	1.414	0.000	Alpha-Blocker Behandlung (3), Alpha-Blocker zur Behandlung (1), alpha-Blocker Behandlung von (1)
controlled trials	11	1.317	0.091	randomisierte kontrollierte Studien (6), kontrollierten Studien (4), randomisierten kontrollierten Studien (1)
drain use	5	1.316	0.000	Drain Verwendung (2), Drainage verwenden (1), Drain benutzen (1)

Domain: cochrane Language: Polish

Term	Count	Score	Precision	Translations
review authors	13	2.870	0.077	przeglądu autorów (10), autorzy przeglądu (2), przegląd autorów (1)
rotational thromboelastometry	5	2.590	0.000	thromboelastometry obrotowa (3), rotacyjnej thromboelastometry (2)
Cochrane Central Register	6	2.329	0.000	Cochrane Centralnego Rejestru (6)
RCTs	28	2.295	0.000	RCTs (28)
developing lymphoedema	5	1.915	0.000	rozwoju obrzęk limfatyczny (4), wystąpienia obrzęku limfatycznego (1)
CENTRAL	7	1.889	0.000	AZJA (7)
adult trauma patients	7	1.859	0.000	dorosłych pacjentów urazami (6), urazu dorosłych pacjentów (1)
mean difference	7	1.594	0.000	średnia różnica (6), średniej różnicy (1)
outcome measures	5	1.558	0.000	zastosowaniu środków (4), środki wynikami (1)
Chinese herbal medicines	5	1.440	0.000	leki ziołowe chińskie (2), chińskiego lekach ziołowych (2), chińskie lekach ziołowych (1)

Domain: cochrane Language: Romanian

Term	Count	Score	Precision	Translations
review authors	13	2.870	0.000	autori revizuire (5), revizuirea autori (4), evaluare autori (3)
rotational thromboelastometry	5	2.590	0.000	de rotație thromboelastometry (4), thromboelastometry de rotație (1)
unclear risk	5	2.490	0.091	clar riscul (3), risc neclar (2)
Cochrane Central Register	6	2.329	0.000	Cochrane registrul central (6)
RCTs	28	2.295	0.000	RCTs (28)
CENTRAL	7	1.889	0.000	ASIA (5), CENTRALE (2)
adult trauma patients	7	1.859	0.000	adulți la pacienții traume (1), traume pentru adulți pacienții (1), la pacienții adulți un traumatism (1)
MD	6	1.656	0.000	MD (6)
outcome measures	5	1.558	0.000	măsuri rezultat (2), rezultatul măsurile (1), măsuri rezultatele (1)
Chinese herbal medicines	5	1.440	0.000	chinezesc bază de plante medicinale (2), chinezesc pe bază de plante medicamente (2), chinezesc medicamente pe bază de plante (1)

Domain: nhs24 Language: Czech

Term	Count	Score	Precision	Translations
NHS	28	3.330	0.000	NHS (27), zdravotnictví (1)
inform	27	1.872	0.000	informovat (27)
feet hip-width	6	1.747	0.000	nohou hip-šířku (4), nohou hip-rozkročte (1), nohou kyčle (1)
Health Library	14	1.413	0.000	Knihovny zdraví (5), zdraví Knihovny (4), Health Library (2)
out more	5	1.245	0.000	více informací (5)
blood supply	8	0.935	0.100	krevní zásobení (3), přívod krve (2), krevního zásobení (2)
underlying cause	5	0.829	0.000	základní příčina (4), základní příčinou (1)
affect	7	0.823	0.000	mají vliv (3), mít vliv na (2), mít vliv (1)
next section	12	0.783	0.000	dalším bodě (11), další sekce (1)
support	15	0.587	0.000	podporu (13), podpory (1), podporovat (1)

Domain: nhs24 Language: German

Term	Count	Score	Precision	Translations
inform	27	1.872	0.000	informieren (27)
feet hip-width	6	1.747	0.000	Füße Hip-Breite (6)
falls	11	1.648	0.083	Stürze (7), fällt (2), Sturzrisiko (1)
Health Library	14	1.413	0.000	Health Library (14)
out more	5	1.245	0.000	erfahren mehr (4), informieren über (1)
strength and balance	8	1.189	0.000	Kraft und Gleichgewicht (3), und Gleichgewicht . (2), Stärke und (2)
physical activity	5	0.976	0.000	körperliche Aktivität (2), Aktivität um (1), Aktivität was Sie (1)
positive things	6	0.864	0.000	positive Dinge (6)
next section	12	0.783	0.000	nächster Abschnitt (11), nächsten Abschnitt (1)
cardiovascular disease	8	0.596	0.000	Herz-Kreislauf-Erkrankung (7), Herz-Kreislauf-Erkrankungen (1)

Domain: nhs24 Language: Polish

Term	Count	Score	Precision	Translations
inform	27	1.872	0.000	poinformować o tym (25), poinformować (1), poinformować o (1)
Heart Helpline	5	1.813	0.000	serce Linia (5)
feet hip-width	6	1.747	0.000	stopy hip-szerokość (6)
good posture	6	1.500	0.000	dobrą posturę (3), dobrą postawa (2), wyprostowana i (1)
Health Library	14	1.413	0.000	Health Library (14)
serious condition	7	1.142	0.000	poważny stan (6), ciężki stan (1)
external link	24	1.099	0.000	połączenie zewnętrznych (21), połączenie) (2), związek zewnętrznych (1)
positive things	6	0.864	0.000	pozytywnych rzeczy (6)
heel	5	0.824	0.000	pięta (3), obcas (1), pięty (1)
try	19	0.646	0.000	nan (7), starać (3), próba (2)

Domain: nhs24 Language: Romanian

Term	Count	Score	Precision	Translations
NHS	28	3.330	0.033	NHS (27), neobișnuită (1)
inform	27	1.872	0.000	informa (26), informează (1)
Heart Helpline	5	1.813	0.000	Heart Linia (4), " Linia (1)
feet hip-width	6	1.747	0.000	picioarele depărtate (6)
Health Library	14	1.413	0.000	Health Biblioteca (14)
regular exercise	10	1.366	0.000	exerciții regulate (9), exercițiu , (1)
out more	5	1.245	0.000	afla mai multe (5)
strength and balance	8	1.189	0.000	și echilibru . (2), puterea și echilibru (2), echilibrului putere și (1)
external link	24	1.099	0.000	link externă (20), link externe (2), link) (2)
positive things	6	0.864	0.000	lucruri pozitive (6)

We note that similar terms seem to cause problems across all the languages. We also note that specific abbreviations can be quite problematic, and that there are definite examples of medical terms (e.g. thromboelastometry) which are not covered in training.

Plotting Precision Against Domain-ness To get a view of how translation precision varies between in-domain and out-domain terms, for different systems, we plot graphs of the rolling mean of precision, against the domain-ness score, in Figures 1 and 2. We created these graphs by ranking all term occurences according to domain-ness score, then using a triangular window of size 500, we computed the sums of clipped reference occurrences, and hypothesis occurrences to obtain the rolling mean precision. We did this separately for each (language, domain, system) combination, plotting the LM variants from Table 6 in Figure 1, and the TM variants in Table 7 in Figure 2.

So how do we interpret these graphs? Overall, we do not see a significant trend in the relation between translation precision and domain-ness. In other words, the precision is similar for in-domain terms, as it is for out-of-domain terms, although the NHS24 data does exhibit a dip in performance at the right-hand side of the graph, showing that performance is not as good on highly out-of-domain terms. Possibly this is because there are certain terms used in the NHS 24 text (see tables on previous pages) that have very specific translations. There is also a similarity in the curves across languages, for the same domain, although the closest similarity is observed between Czech and Polish.

In terms of the difference between models, again there is more to be seen in the LM comparison than in the TM comparison. For instance for Cochrane-Czech, there is a distinct separation of the models at around 1.5, showing especially bad performance for `lm-inout-log`, and much better performance for the linearly interpolated models. In NHS24-Romanian, we see that the poor performing `lm-all` model does comparatively worse on the most in-domain terms, suggesting that it is poorly adapted. The TM curves are harder to separate, although there is the same divergence at around 1.5 for Czech-Cochrane, and there is a general tendency for the interpolated model (green) to diverge more from the others (for example in NHS24-Polish).

2.4 Domain Adaptation for Neural MT

Since HimL started, there has been a dramatic shift in the field of machine translation. Neural machine translation (Bahdanau *et al.*, 2014; Sutskever *et al.*, 2014) has achieved state-of-the-art results in shared tasks (Cettolo *et al.*, 2015; Bojar *et al.*, 2016c), and been deployed in commercial systems (Wu *et al.*, 2016; Crego *et al.*, 2016). The speed of the movement from the earliest under-performing lab-based systems, to state-of-the-art deployed systems has been more rapid than many anticipated, and has resulted in a major re-alignment of the research community's efforts. Interest in "traditional" SMT models such as phrase-based, hierarchical and syntax-based has reduced as interest in NMT has increased.

Due to the excellent results shown by NMT on many tasks, we decided to build NMT systems for the HimL languages, and see how they would perform in the public health domain. NMT systems are data-driven, in the same way as earlier SMT systems, so are still potentially sensitive to differences between the training and test. In other words, domain adaptation can still be a problem for NMT. However NMT opens up new possibilities for addressing domain adaptation, due to the simplified training pipeline, and the possibility to include extra information (such as context) in the NMT model.

In the following subsections we describe how we built baseline NMT systems for HimL, and our attempts at improving domain adaptation for NMT.

2.4.1 Baseline NMT Systems

The data sets we use for the baseline NMT systems are similar to those described in Table 4, with the following differences:

- For English-Czech, we only use the permissible data for WMT16, so we do not include any extra data from OPUS.
- For English-German, we set aside 20000 sentences each from JRC-Acquis and OpenSubtitles2016, to use in the future as dev/test for possible future domain adaptation experiments.
- For English-Polish, we set aside 20000 sentences from JRC-Acquis, EMEA and Europarl, again for domain adaptation experiments. We also added a small ($\approx$ 900 sentence) corpus of Cochrane data.
- For English-Romanian, we set aside 20000 sentences from JRC-Acquis.

We also performed some baseline experiments without the subtitles corpora, to see the effect of this large, out-of-domain, and sometimes noisy corpus. The sizes of the parallel training data used in the baseline NMT systems are shown in Table 8. Note that these show the data set sizes with sentence length limited to 80 – we further removed sentences of length greater than 50 when loading data into the NMT system for training.

	en-cs	en-de	en-pl	en-ro
All	52,002,221	43,578,522	57,876,920	83,994,787
w/o Subtitles	–	30,172,195	7,274,455	3,783,779

Table 8: Parallel training set sizes for NMT baseline experiments, showing the size of the whole data set, and the size with subtitles removed.

Our baseline systems are created using the same process described in Sennrich *et al.* (2016a). We use the NMT implementation provided by Nematus[28], based on code released for an NMT tutorial[29]. This implements an updated version of the model described in Bahdanau *et al.* (2014). We use a word-embedding dimension of 500 and a hidden layer dimension of 1024.

The pre-processing involves normalisation, tokenisation and truecasing with Moses scripts, as for the phrase-based MT systems. We then learn a joint byte-pair encoding (BPE) model (Sennrich *et al.*, 2016c) on the training data, with 89500 merges, and apply this model to segment the training data into subwords. We limit the vocabulary for both target and source to 85000, mapping other tokens to UNK.

Training involves optimising cross-entropy with adadelta, using a minibatch size of 80, a learning rate of 0.001, and clipping gradients to 1. We run validation (cross-entropy on the full HimL tuning set) every 10000 steps, and stop training when convergence is detected. This takes between 0.8M and 1.3M steps, depending on language pair.

For testing, we use an ensemble of the last 4 save-points (saving every 30,000 iterations). Decoding is with beam-search with a beam size of 12, and we normalise sentence scores by length to prevent the NMT system from producing overly short sentences.

In Table 9 we show the BLEU scores on the HimL test sets, for training baseline NMT systems with and without subtitles.

Comparing the results of Table 9 with those in Table 7 shows us that the NMT systems still lag behind the phrase-based systems, on BLEU scores at least. It is interesting to note that the only pair where NMT beats PBMT is English-Czech, and in this case the data is drawn from the CzEng corpus, as opposed to drawing extensively from OPUS. Possibly the English-Czech data is

[28] https://github.com/rsennrich/nematus
[29] https://github.com/nyu-dl/dl4mt-tutorial

	en-cs		en-de		en-pl		en-ro	
Model	Coch	NHS24	Coch	NHS24	Coch	NHS24	Coch	NHS24
Baseline	30.2	23.1	37.6	31.6	15.5	19.5	31.5	28.6
w/o Subtitles	–	–	38.2	30.8	17.1	20.7	31.9	27.8

Table 9: Baseline NMT results, using data sets listed in Table 8

.

cleaner than for the other language pairs, and the NMT systems are being harmed by noisy data. It is also notable that removing subtitles has a positive effect on English-Polish, and only has a small adverse effect on English-Romanian, despite reducing the training set size by a factor of about 40. This again could indicate that NMT is being affected by noisy data, or by the severely out-of-domain nature of the Subtitles corpus.

2.4.2 Fine-tuning for Domain Adaptation

In previous work on NMT, it has been observed that "fine-tuning" of NMT systems, with parallel data (including synthetic data) can be very effective, particularly if the data is from the same domain as the test data (Luong and Manning, 2015; Sennrich *et al.*, 2016b,a). We therefore tried to improve the systems of Section 2.4.1 by fine-tuning with appropriate data. In previous work the best fine-tuning results were obtained using large in-domain data sets, however these are not available in the HimL scenario, so we had to try alternative approaches.

The available in-domain parallel data mainly consists of translation memories from Cochrane (around 10,000 sentences for German-English, about 1000 for Polish-English) and the EMEA corpus (crawled from drug information leaflets). The EMEA corpus is related in content to the HimL test sets, but different in style, tends to be repetitive, and suffers from extraction/alignment errors. Initial experiments in fine-tuning with EMEA and Cochrane, suggested that the former did not help, whilst the latter gave small gains but required careful regularisation

Due to the lack of in-domain parallel data we decided to use synthetic data, as in the Edinburgh WMT16 submissions. The problem with the HimL scenario is that there is no clear source of in-domain target data, as compared to WMT where there are large monolingual news corpora in all languages. Instead, we adopted the following procedure for creating synthetic parallel data for HimL systems:

1. Crawl both Cochrane and NHS24 websites, creating two in-domain English corpora of about 170,000 and 60,000 unique sentences, respectively.
2. Translate these corpora to each of the 4 HimL target languages using baseline NMT systems.
3. Use the Moore-Lewis method (Moore and Lewis, 2010) to select a quantity of data from language-specific sections of CommonCrawl, as extracted by Buck *et al.* (2014). We also ignore lines longer than 80 or shorter than 10 words.
4. Back-translate this selection to English (either using Edinburgh WMT16 systems, where available, or another NMT system created from the baseline data) to create a synthetic parallel corpus, with the target side drawn from CommonCrawl.

These synthetic corpora were used for fine-tuning of the baseline NMT systems, with training starting with the final save-point of the baseline training run. For the fine-tuning set, the synthetic data was mixed with a roughly equal sized random selection of the original training data, plus the EMEA corpus, and the Cochrane TMX, where available.

For English-Romanian, we noticed an immediate problem when fine-tuning with the synthetic data mix. Scores went down on the heldout data and did not recover. An investigation showed that the fine-tuned system was creating Romanian without diacritics, and this was due to the lack of diacritics on the target side of the synthetic corpus, drawn from CommonCrawl. In order to address this, we trained a Romanian "diacritiser" to translate from Romanian text without diacritics, into Romanian with correct diacritics. This diacritiser was an NMT system trained on a selection of good-quality Romanian corpora, with the diacritics artificially removed to create the source side. Using the diacritiser we further processed the Romanian selection from CommonCrawl to attempt to reinstate its diacritics, and used this processed version as the target side of the synthetic data.

Statistics on the composition of the fine-tunining corpora are shown in Table 10. The results of the fine-tuning experiments are shown in Table 11. We started with the baselines from Table 9, then continued training from the last save-point using a mix of synthetic and parallel data. The synthetic data was either selected using the Cochrane or the NHS 24 web crawl, using the procedure described above. Again, the BLEU scores are for ensembles of the final 4 save-points. We can see from Table 11 that this type of fine-tuning is always beneficial, with gains of up to 4.7 BLEU points possible. Mostly the biggest improvements are obtained when the corpus used for selection matches the training set, but there are exceptions to this rule (for en-de Cochrane and en-pl NHS 24). Overall, the BLEU scores of the best NMT systems are comparable to those of the best PBMT systems in all language pairs, except for en-ro. For this pair, the NMT systems still lags about 3 BLEU points behind the best PBMT system, perhaps because it is more adversely affected by the noise cause by inconsistent diacritisation.

	en-cs	en-de	en-pl	en-ro
Sampled from baseline	4,000,000	10,000,000	9,000,000	5,000,000
Selected from CommonCrawl (using Cochrane)	3,392,661	10,509,551	8,414,483	4,955,355
Selected from CommonCrawl (using NHS24)	3,392,661	8,018,461	9,832,401	5,345,227
EMEA	259,653	283,520	239,743	240,283
Cochrane TM	–	1,159,900	–	–

Table 10: Data set sizes (sentences) for fine-tuning data, before cleaning and removal of sentences longer than 50 tokens. Note that two different selections from CommonCrawl were used in two separate experiments, each time combined with the other data. Note also that the Cochrane data is duplicated 100 times to upweight it.

	en-cs		en-de		en-pl		en-ro	
Model	Coch	NHS24	Coch	NHS24	Coch	NHS24	Coch	NHS24
Baseline (as Table 9)	30.2	23.1	37.6	31.6	15.5	19.5	31.5	28.6
Fine-tune with synthetic mix, selected with Cochrane	**33.4**	25.6	38.5	31.7	**19.1**	**24.9**	**34.4**	29.0
Fine-tune with synthetic mix, selected with NHS24	33.2	**26.7**	**39.2**	**32.9**	18.9	24.2	34.1	**29.7**

Table 11: Performance of NMT systems after fine-tuning baselines with mixture of parallel and synthetic data.

2.4.3 Domain-sensitive Models for NMT

In the previous section we showed how data improvements could increase the performance (as measured by BLEU) of NMT systems on the HimL data sets. In this section we look at model-based attempts to improve the performance of NMT on specific domains. The work in this section is much more exploratory than in previous sections, with sometimes incomplete experimental results, and a mixture of baselines and test sets.

The idea behind the proposed new models is to find ways for the translation to take into wider context, to sentence-level and beyond. We assume that a badly adapted MT system is one which produces translations which are inappropriate for the context, and seek to include the context in the models.

In principle, neural MT makes it easier to include extra information in the model. There is a single training step to estimate $p(e|f)$, and adding extra information is simply a matter of designing the correct network architecture to include it in the model. We experiment with two types of extra information for translation of a given sentence: the corpus in which it is found, and the surrounding context of the sentence.

Experiment 1: Domain indicator as source factor The domain indicator for any sentence is the name of the corpus in which it was found, as used in the provenance features described in Section 2.1.2. The first method we tried for including the domain indicator in the model is just to add it as an extra feature on each word in the corpus, using the mechanism developed in Sennrich and Haddow (2016). An alternative (and similarly straightforward) way of adding a domain indicator to the source would be to add an extra "pseudo-word" to the source sentence.

For the experiments with domain indicator as source factor, we used a sub-sampled version of the Czech-English WMT16 data consisting of 4 million sentence pairs (from a total corpus size of about 52 million). The sub-sampling was implemented by taking first 1/13th of each of the constituent corpora of CzEng, plus news-commentary and europarl. The test sets used were the HimL test sets (concatenated), the nc-test2008 set from WMT08, plus 2000 sentence corpora extracted from the end of the WMT16 europarl and CzEng subtitles corpora. A dev set was created by taking separate 750 sentence samples from HimL tuning, nc-devtest2008, europarl and subtitles. The training, dev and test sets are all annotated with domain indicators as source factors, although note that whilst the news-commentary, europarl and subtitles domains are represented in training as well as dev/test, the HimL domain only appears in dev/test. The training procedure was the same as in Section 2.4.1, again with testing using ensembles from the last 4 save points. In Table 12 we show the performance of the systems on the 4 different test sets.

System	europarl	news-comm	subtitles	HimL
Baseline	26.7	17.4	14.5	21.6
with domain indicator	27.3	17.6	16.2	19.7

Table 12: Comparison of baseline, with system trained on data with domain indicators as source factors. Trained on a sub-sampled version of CzEng1.6pre. Case sensitive BLEU scores calculated with `multi-bleu.perl`.

Looking at the results in Table 12 we see that there is a small gain for the europarl test set, and an even smaller one for the news-commentary test set. For subtitles, the situation is a little strange since there appears to be a large gain on the ensemble, but decoding with the single best system (measured on the dev set) shows baseline vs. domain indicator as 16.5 vs. 16.9. The baseline ensemble outputs many nonsense sentences, giving a length ratio of 1.2. On the HimL test set, the fact that the HimL domain indicator is not present in training causes performance to suffer when it is used in test, showing that we need to find a different model for scenarios where the test set is not drawn from one of the training domains.

Experiment 2: Domain indicator as target pseudo-word The problem with using the domain indicator in the model, is that the test sentence may not come from any of the training domains. A possible solution to this is to have the model first predict the domain based on the source sentence, then generate the target sentence taking into account that prediction. We can do this easily by making the domain indicator an initial "pseudo-word" in the target sentence during training. At test time, this word is predicted, but stripped from the output, then the rest of the sentence is predicted, conditioning on the pseudo-word.

For the target pseudo-word experiments, we used the HimL setup from Table 9, specifically en-ro, without subtitles. We show results in Table 13 on both HimL test sets, as well as the WMT16 en-ro test set. From Table 13, we see that the domain

System	Cochrane	NHS24	newstest2016
Baseline	31.9	27.8	23.8
with domain indicator	35.6	29.0	23.9

Table 13: Comparison of baseline, with system trained on data with domain indicators as target pseudo-words. Trained on corpus of en-ro, as in Table 8, without subtitles.

indicator model offers improvements in BLEU on both the HimL data sets (substantially for Cochrane) but essentially no change in performance on the WMT16 data. In Table 14 we show the counts of the different target pseudo-words predicted in the test runs above. The results of Table 14 at least confirm that the domain indicator is behaving as expected, with the HimL sets most

Corpus	Cochrane	NHS24	newstest2016
EMEA	365	561	13
DGT	176	324	100
Europarl	57	63	445
SETIMES2	37	37	846
SETIMES	6	3	235
EUBookshop	19	102	66
TED2013	5	72	227
Others	8	96	67

Table 14: Counts of domain indicator predictions in test sets, with domain indicator as 0th pseudo-word

often predicted as medical (EMEA) and the WMT data predicted mostly as news (SETIMES and SETIMES2).

Experiment 3: Soft prediction of domain indicator Predicting the domain indicator as the 0th word of the target sentence seems slightly unnatural, and has the added problem that it enforces a hard decision. If the domain indicator is predicted as "EMEA", for example, then the rest of the sentence is decoded based on this prediction only, without allowing for any doubt. In practice, it may make more sense to predict a distribution across domain labels, and base the word prediction on this distribution.

To accomplish this soft prediction, we change the architecture of the NMT system. We introduce a "sentence factor", which in training is an extra output, corresponding to the domain indicator. This sentence factor is predicted by the network, using a feed-forward gate taking the encoder context (i.e. concatenated forward and backward hidden states) as input. In the "predict-sf-only" version, we just predict the sentence factor using this feed-forward network, and include the cross-entropy of this prediction in the objective, giving a multi-task model. In the "sf-decoder" model, we insert the embedding of the predicted sentence factor into the decoder, by appending it to the attention-weighted encoder context at each output time step. If the domain indicator were known at test time, then we could use the actual domain indicator in the decoder, but we have not yet experimented with this variant.

The experiments for soft prediction of domain indicator, use a different setup again, since we were re-using a setup created for Experiment 4 (below) with unshuffled training data. The language pair is English-German, and the training data is drawn from news-commentary, europarl, books, Acquis communautaire, subtitles and medical patents, totalling approximately 4 million

sentence pairs. Training was as before, except that we used a minibatch size of 60 instead of 80, again following Experiment 4. The results are shown in Table 15, giving scores on HimL test sets, as well as WMT newstest2015.

	en-de			de-en		
System	Coch	NHS24	n2015	Coch	NHS24	n2015
baseline	26.7	22.1	22.4	32.6	30.1	25.9
predict-sf-only	26.9	22.5	22.2	33.0	30.1	26.0
sf-decoder	–	–	–	33.3	29.6	26.1

Table 15: Comparison of baseline system, with models that make predictions of the domain indicator using a feed-forward network on top of the encoder. Test sets are from HimL and WMT (newstest2015). The "predict-sf-only" model is a multi-task model doing domain indicator prediction as well as translation, and "sf-decoder" additionally feeds this domain indicator prediction back to the decoder portion of the model.

In Table 15 we observe that the multi-task model (predict-sf-only) offers small gains on the HimL data, but not on the WMT data. It appears that feeding back the domain prediction to the decoder is not helpful, but we think more experimentation with this technique could be useful.

Experiment 4: Document attention The final domain adaptation for NMT model that we consider here attempts to condition translations on the wider context, and does not consider domain indicators. The idea here is that the model could learn what type of translated language is appropriate for the context. To do this we use the attention mechanism of the NMT model. We introduce a second form of attention where, as well as attending to the source context vectors for a particular sentence, we also take a weighted sum of the average context vectors of all source sentences in the minibatch. The source context and the document context vectors are concatenated, and used in the decoder gates in the same way as the baseline. At test time, the data is also processed in minibatches of consecutive sentences.

The minibatch is thus treated as a pseudo-document, with all sentences in the minibatch able to affect the translation of any sentence in the minibatch. This changes the training setup slightly, in that we must ensure that the minibatch consists of sentences which are adjacent in the training data, and that the training data is not sentence-shuffled. Normally in NMT training, we select "maxibatches" randomly from training, consisting of (say) 20-times the number of sentences that we have in a minibatch. The maxibatch is then sorted by length, and the minibatches are drawn from this maxibatch. This setup means that each minibatch consists of sentences of roughly the same length, optimising GPU utilisation. In order to get coherent minibatches, we do not apply this sorting, so minibatches take (on average) longer to process, increasing training time. We do not have experiments with matching hardware and drivers, but rough estimates suggest that training time increases by 50-100%.

The experiments with document attention use the same training setup as in Experiment 3, including limiting the batch size to 60 since the models occupy more GPU memory. We use a baseline which also has unsorted minibatches, which is why its performance is slightly different. The results of the document attention experiments are shown in Table 16.

	en-de			de-en		
System	Coch	NHS24	n2015	Coch	NHS24	n2015
baseline	27.5	23.5	22.8	33.3	29.5	26.1
with doc attention	27.3	23.6	23.0	33.8	31.1	26.5

Table 16: Comparison of baseline model, with one which adds document attention.

The results in Table 16 show that the document attention model helps in the de-en direction across all data sets, but does not appear effective in en-de (although does not reduce performance).

Summary In this section we have shown several different models for including domain information in NMT. At this stage, whilst there are some encouraging results, we do not have a sufficiently consistent set of experiments to draw conclusions from. The results here should be treated as a preliminary set of experiments into what could be done, which need further investigation.

The domain indicator experiments show how to integrate simple pieces of information (the corpus containing a training sentence) into the model, however this is only one limited piece of domain data. The power of neural network models is better shown by experiment 4, where the document context can be incorporated into the model using a reasonably straightforward extension of the standard model. Whilst we believe that this is a step in the right direction, there are still questions to be resolved about the

best way to efficiently incorporate document context into the model, and of course the practical problem that much MT training data is shuffled before being released.

Conclusion

In this deliverable we describe the efforts to collect data for the HimL-specific domains, and to build translation systems with this data. We showed how domain adaptation, especially of language models, and to a lesser extent translation models, improve the translation performance of a phrase-based MT system built on large diverse data sets. We also showed early results with Neural MT, and how it could also be improved using domain adaptation techniques. Indications are that the NMT systems are more affected by noise than phrase-based systems.

References

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. "Domain Adaptation via Pseudo In-Domain Data Selection." *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. "Neural machine translation by jointly learning to align and translate." *CoRR*, abs/1409.0473.

Bojar, Ondřej, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016a. "CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered." *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, no. 9924 in Lecture Notes in Computer Science, 231–238. Cham / Heidelberg / New York / Dordrecht / London.

Bojar, Ondřej, Ondřej Dušek, Tom Kocmi, Jindřich Libovickỳ, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. "Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered." *International Conference on Text, Speech, and Dialogue*, 231–238.

Bojar, Ondřej, 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. 2016c. "Findings of the 2016 conference on machine translation." *Proceedings of the First Conference on Machine Translation*, 131–198. Berlin, Germany.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. "Findings of the 2015 workshop on statistical machine translation." *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Buck, Christian, Kenneth Heafield, and Bas van Ooyen. 2014. "N-gram counts and language models from the common crawl." *Proceedings of the Language Resources and Evaluation Conference*.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. "The iwslt 2015 evaluation campaign." *Proceedings of IWSLT*.

Chen, Boxing, George Foster, and Roland Kuhn. 2013a. "Adaptation of reordering models for statistical machine translation." *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chen, Boxing, Roland Kuhn, and George Foster. 2013b. "Vector space model for adaptation in statistical machine translation." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chen, Boxing, Roland Kuhn, and George Foster. 2014. "A comparison of mixture and vector space techniques for translation model adaptation." *Proceedings of AMTA*.

Cherry, Colin and George Foster. 2012. "Batch tuning strategies for statistical machine translation." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 427–436. Montréal, Canada.

Crego, J., J. Kim, G. Klein, A. Rebollo, K. Yang, J. Senellart, E. Akhanov, P. Brunelle, A. Coquard, Y. Deng, S. Enoue, C. Geiss, J. Johanson, A. Khalsa, R. Khiari, B. Ko, C. Kobus, J. Lorieux, L. Martins, D.-C. Nguyen, A. Priori, T. Riccardi, N. Segal, C. Servan, C. Tiquet, B. Wang, J. Yang, D. Zhang, J. Zhou, and P. Zoldan. 2016. "SYSTRAN's Pure Neural Machine Translation Systems." *ArXiv e-prints*.

Cuong, Hoang and Khalil Sima'an. 2015. "Latent domain word alignment for heterogeneous corpora." *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Durrani, Nadir, Hassan Sajjad, Shafiq Joty, Ahmed Abdelali, and Stephan Vogel. 2015. "Using joint models for domain adaptation in statistical machine translation." *Proceedings of AMTA*.

Dušek, Ondřej, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, and Daniel Zeman. 2014. "Machine translation of medical texts in the khresmoi project." *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Haddow, Barry, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. "The edinburgh/jhu phrase-based machine translation systems for wmt 2015." *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 126–133. Lisbon, Portugal.

Heafield, Kenneth, Chase Geigle, Sean Massung, and Lane Schwartz. 2016. "Normalized log-linear interpolation of backoff language models is efficient." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. "Scalable modified kneser-ney language model estimation." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Huck, Matthias, Alexander Fraser, and Barry Haddow. 2016. "The edinburgh/lmu hierarchical machine translation system for wmt 2016." *Proceedings of the First Conference on Machine Translation*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, *et al.* 2007. "Moses: Open source toolkit for statistical machine translation." *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180.

Luong, Minh-Thang and Christopher D. Manning. 2015. "Stanford neural machine translation systems for spoken language domain." *International Workshop on Spoken Language Translation*.

Moore, Robert C. and William Lewis. 2010. "Intelligent selection of language model training data." *Proceedings of the ACL 2010 Conference Short Papers*.

Rosa, Rudolf, David Mareček, and Ondřej Dušek. 2012. "Depfix: A system for automatic correction of czech mt outputs." *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 362–368.

Rousseau, Anthony. 2013. "Xenc: An open-source tool for data selection in natural language processing." *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Sennrich, Rico. 2012. "Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation." *Proceedings of EACL*.

Sennrich, Rico and Barry Haddow. 2016. "Linguistic input features improve neural machine translation." *Proceedings of the First Conference on Machine Translation*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. "Edinburgh neural machine translation systems for wmt 16." *Proceedings of the First Conference on Machine Translation*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. "Improving neural machine translation models with monolingual data." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016c. "Neural machine translation of rare words with subword units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." *ArXiv e-prints*.

Tamchyna, Aleš, Roman Sudarikov, Ondrej Bojar, and Alexander Fraser. 2016. "Cuni-lmu submissions in wmt2016: Chimera constrained and beaten." *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Tiedemann, Jörg. 2009. "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces." *Recent Advances in Natural Language Processing (vol V)*.

Williams, Philip, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. "Edinburgh's statistical machine translation systems for wmt16." *Proceedings of the First Conference on Machine Translation*.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *ArXiv e-prints*.

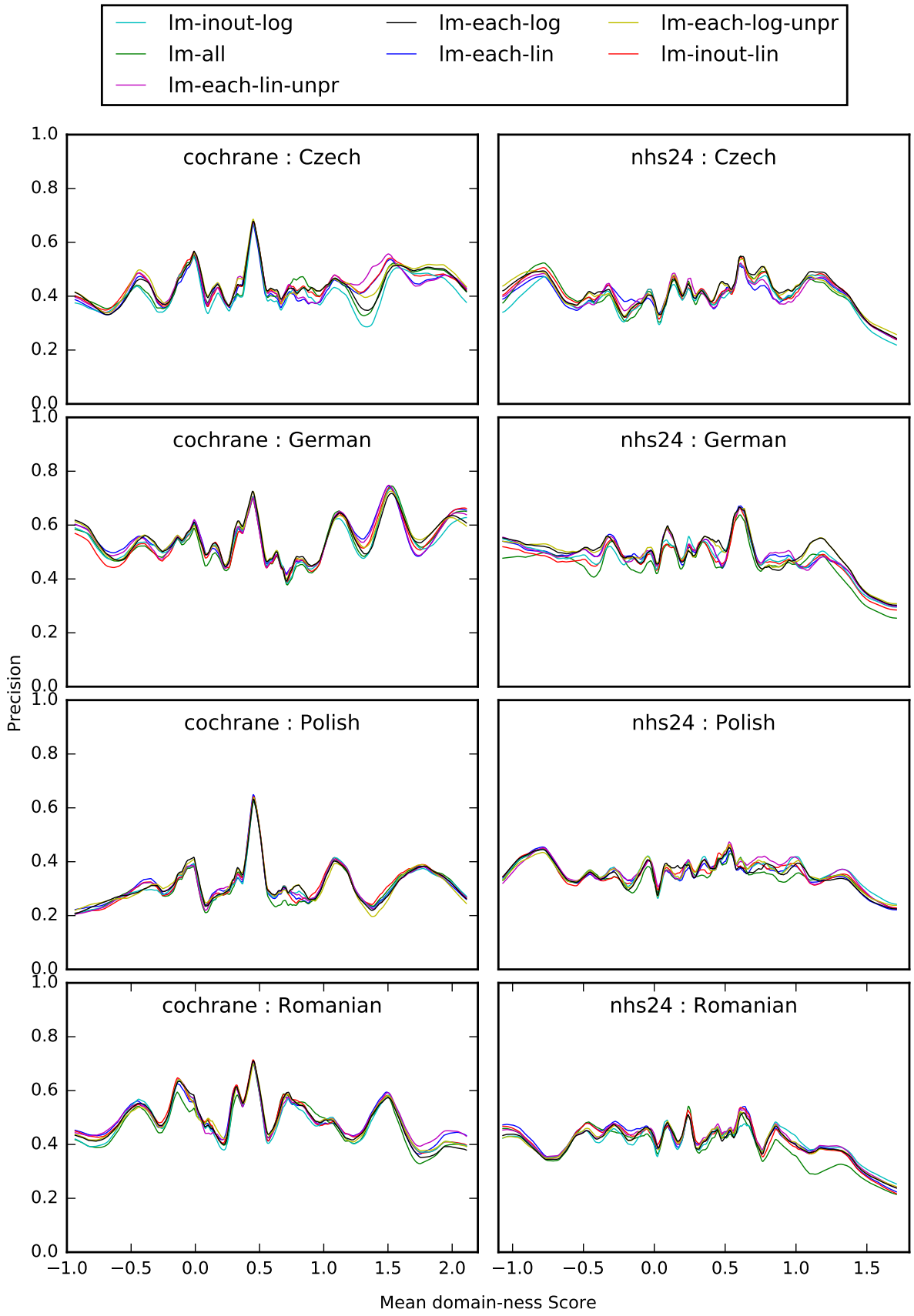

Figure 1: Precision (rolling mean) against mean domain-ness of terms for LM variants

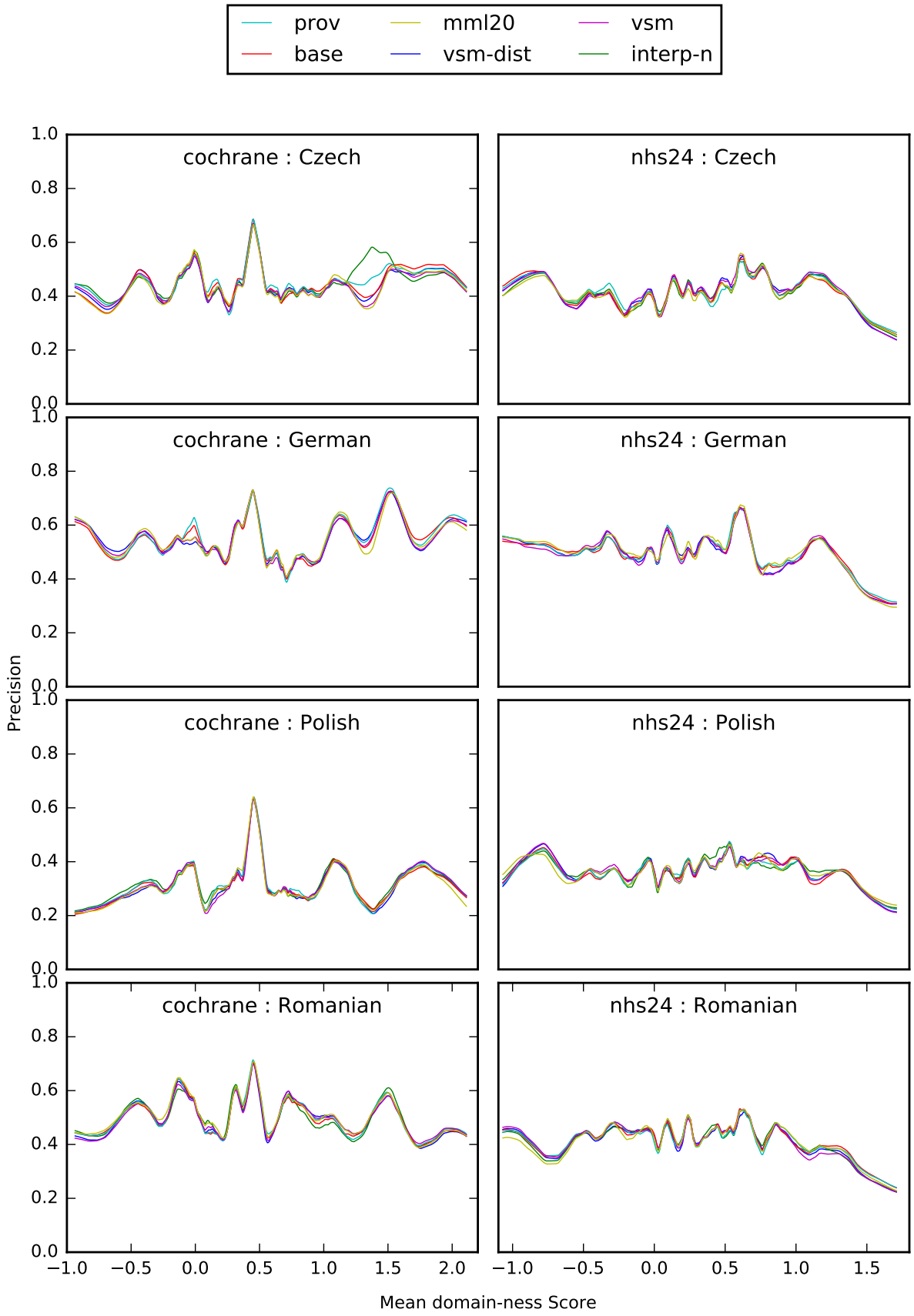

Figure 2: Precision (rolling mean) against mean domain-ness of terms for TM variants