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

D5.4: Report on second year's MT evaluation

Author(s): Alexandra Birch

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	D5.4
Deliverable title	Report on second year's MT evaluation
Type	Report
Status and version	1.0
Number of pages	13
Contributing partners	UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Alexandra Birch
EC project officer	Martina Eydner
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

© 2017 Alexandra Birch

This document has been released under the Creative Commons Attribution-Non-commercial-Share-alike License v.4.0 (http://creativecommons.org/licenses/by-nc-sa/4.0/legalcode).

Contents

1 Executive Summary

This document provides a summary of the work done in the second year of the HimL project. The objectives of this work package are to develop human and automatic accuracy-based evaluation strategies for machine translation, and to evaluate the quality and impact of the innovations we deliver to NHS24 and Cochrane.

The first section in this report (Section 2) describes the creation of the second HimL test set from our use case partners. In the next section (Section 3) we describe experiments using the human semantic MT evaluation metric developed in this project. The HUME metric was applied to the output of three different types of translation systems across the HimL languages to determine if it is able to discern differences between them. The final content section in this deliverable (Section 4) describes a ranking experiment which we performed to compare various different versions of the HimL machine translation systems. We also included a commercial machine translation system, Google, in this evaluation.

We conclude with an outlook in Section 5.

Task	Description	Planned Schedule	Status
5.1	Test corpora for the required language pairs	M1-M6	Complete
5.2	Human Semantic MT Evaluation Metric	M1-M12	Complete
5.3	Automatic Semantic MT Evaluation Metric	M6-M18	Complete
5.4	User acceptance testing	M9-M12 annually	Y2 Complete
5.5	Evaluation of the impact	M6-M36	Delayed, starting 1 Feb 2017

Table 1: Tasks in workpackage 5, their dates and their status as of 31/Jan/2017

Table 1 summarises the status of the tasks in the evaluation workpackage. We can see that all tasks are currently on schedule except for task 5.5. There have been delays in integrating the year 2 models in the HimL translation platform which has meant that NHS24 and Cochrane are only testing the year 2 translation models now. We anticipate that we will be able to perform user surveys and collect website statistics starting from the 1 February 2017 and we will be able to report results of these studies in the upcoming deliverable D5.5 Report on integrated semantic evaluation metric which is due on 31 July 2017.

2 Second HimL Test set

The HimL test set has been heavily used for development over the last two years. This means that its value as a genuinely unseen test set has been compromised. We have gathered a new data set from our user partners to create an unseen test set with which we report the final evaluation numbers for the HimL project. In Table 2 we see the statistics for the new test set. This is currently being translated and will be available shortly.

User Partner	Sentences	Words
NHS24	1044	13531
Cochrane	464	8671

Table 2: Statistics of the English source of the year 2 test set

3 Semantic Evaluation

In the HimL project, we have been developing a human metric of adequacy called Human UCCA-based MT Evaluation (HUME). This was initially described in Deliverable 5.2 Report on first year's MT evaluation, and then in further detail in Deliverable 5.3 Report on preliminary semantic evaluation metric. We provide a brief motivation here and then discuss results of this year's evaluation.

Human evaluation of machine translation normally uses sentence-level measures such as relative ranking or adequacy scales. However, these provide no insight into possible errors, and do not scale well with sentence length. We demonstrated that a semantics-based evaluation, which captures what meaning components are retained in the MT output, provides a more fine-grained analysis of translation quality, and enabling the construction and tuning of semantics-based MT. We presented a human semantic evaluation measure (Birch et al., 2016), building on the UCCA semantic representation scheme. HUME covers a wider range of semantic phenomena than previous methods and does not rely on semantic annotation of the potentially garbled MT output.

In year 1 of the HimL project we established HUME's broad applicability, and reported good inter-annotator agreement rates and correlation with human adequacy scores. In this year's HUME evaluation we experiment with HUME annotation on different machine translation system types, such as phrase-based, syntax-based and neural machine translation models. We determine what a semantic metric like HUME can tell us about the different behaviours of these systems.

3.1 Data

The goal of this evaluation was to see the value of the HUME evaluation metric in detecting differences between translation systems. In order to have access to a wide variety of top performing translation systems we turned to publicly available data. We selected 301 sentences from the shared task in the Conference on Machine Translation from 2016 (WMT16). An added advantage of using this data was so that we could compare our semantic evaluation results with existing ranking results. This data was only available for German, Czech and Romanian as English-Polish was not one of the WMT language pairs last year. So for Polish we selected the 351 sentences from last year's UCCA annotated HimL test set. We then used the HimL year 1 and year 2 systems and the HimL NMT system to translate the sentences to Polish.

German	Czech	Polish	Romanian
Neural MT	Neural MT	Neural MT	Neural MT
Phrase-based MT	Tectogrammatic MT	Phrase-based MT	Phrase-based MT Y1
Syntax-based MT	Chimera System Combination	Phrase-based MT Y2	System Combo

Table 3: Systems used in the comparative study

3.2 Annotators

Cochrane The majority of Cochrane annotators who participated in the human semantic annotation and the ranking tasks have a medical background and are familiar with Cochrane. As such, they are both domain experts, and part of Cochrane's professional audience, but not representative of the general public audience. In addition, some of the annotators are also experienced Cochrane translators, and as such are very familiar with Cochrane content. Cochrane provided one, two or three annotators for each of the target languages.

NHS24 NHS 24 manages a number of websites for NHS Scotland. NHSinform, as used by the EU HimL project has about 20,000 pages and articles. Translations of health information are available from NHS 24 in a number of languages but if a specific document is not already available in any language it will be translated on request. NHS 24 does not have translators amongst its staff but has call-off contracts with a number of translation agencies for this service when required. The nature of the fine grained semantic analysis task involved sentences being deconstructed and the correctness of component words and phrases being assessed. This task required native speakers of the target language, educated to degree level, with excellent written English and if possible experience in translation of health documents. Two agencies were asked to supply 2 candidates for each language (Polish and Romanian) with the intention of selecting one subject for each language for the task. Candidates resident in Central Belt of Scotland were preferred as this kept travel costs down.

Agencies used their databases to select candidates, the deconstructive nature of the task being viewed as important led to candidates with a Linguistics degree being put forward. The candidates with the longest experience of health related translation work were selected. Agencies were paid a fixed sum for 40 hours work including task training, on-line annotation and task review sessions.

3.3 Results

In order to explore the results of our annotations, we first show some basic statistics about the task. In Table 4, we can see the total number of sentences annotated with the number of nodes.

As we can see in Table 4 we have collected a large number of annotations for each of the target languages. German and Czech have two annotators and Romanian and Polish have three.

In Table 5 we can see the HUME score results for the different systems for the different language pairs. The HUME score is defined as the average number of correct nodes. Correct nodes are 'Green' and 'Acceptable' nodes, and 'Orange' nodes count for half. Incorrect nodes are 'Red' and 'Bad' nodes. We can see that the HUME score does highlight interesting differences between systems. It seems like for German the HUME scores are highest for the syntax-based models, and for Czech and Polish the neural model comes highest. For Romanian it is the phrase-based model which scores highest. Looking at the differences

	No. Sentences	No. Nodes
De1	495	12299
De2	496	11949
Cs1	496	12321
Cs2	487	11649
Pl1	287	7664
Pl2	288	7595
Pl3	522	14428
Ro1	311	7611
Ro2	287	7100
Ro3	204	5046

Table 4: Number of annotated sentences and nodes

between the scores for lexical and atomic nodes, it seems that lexical nodes are scored higher for German and Romanian than structural nodes. For Czech this is reversed and the structural nodes score higher. For Polish there is not much difference between structural and lexical nodes. Another point to mention is that the Tecto system seems to score much lower than all other systems. This could be because of one particular annotator who was more strict than the rest. We are investigating this further.

In Table 6 we can see the overall scores given per annotator. It seems clear that annotator 'Cs2' has behaviour which is quite different to the other annotator and scores everything much lower, in particular the lexical nodes. In future analyses we will consider normalising the annotator scores.

In Table 7 we can see the initial inter annotator agreement scores. We can see that for German and Romanian they are quite good, but for Polish and particularly Czech, there is much more disagreement. We will investigate these result further.

3.4 Summary

We have run an experiment comparing three top-ranked systems from WMT with different underlying models, in order to determine if our HUME metric is able to discern patterns of differences between systems. We have collected a substantial number of judgments across the four HimL languages and, in a preliminary analysis, we have shown that the HUME metric can be used to detect interesting differences between machine translation systems.

4 Ranking

In the Evaluation Plan we promised to perform a ranking experiment to get humans to evaluate the HimL systems. Ranking of the output of different machine translation systems is the standard method for judging which system is better quality. In the annual machine translation shared tasks in the Conference on Machine Translation, the official results are provided by a human ranking task. We have performed this task to compare different versions of our HimL machine translation models on HimL data, and to compare our models to a commercially available translation system. In this experiment we chose Google.

4.1 Data

The test set is the HimL test set created in year 1 of the project. The entire test set was converted into xml HITs (Human Intelligence Task) which consist of a ranking task for 3 consecutive sentences. These HITs were then randomly shuffled so that annotators could not guess which system they were annotating by its position in the list. In Table 8 we can see the number of sentences and HITS created for the ranking task.

We took the HimL test sets and we translated them with four models:

- Year 1 HimL system
- Year 2 HimL system
- Google
- NMT model trained for HimL

German

Overall	Score
NMT	84.22
PBMT	78.95
SBMT	90.60
Lexical Nodes	
NMT	88.57
PBMT	85.19
SBMT	91.64
Structural Nodes	
NMT	75.98
PBMT	69.05
SBMT	88.60

Czech

Overall	Score
NMT	78.82
Tecto	36.05
Chimera	68.96
Lexical Nodes	
NMT	76.07
Tecto	29.54
Chimera	68.55
Structural Nodes	
NMT	84.15
Tecto	49.06
Chimera	69.75

Polish

Overall	Score
NMT	68.93
PBMT	68.55
Year1	65.80
Lexical Nodes	
NMT	67.48
PBMT	67.23
Year1	61.52
Structural Nodes	
NMT	71.71
PBMT	71.06
Year1	73.77

Romanian

Overall	HUME Score
NMT	82.87
PBMT	86.62
Combo	80.41
Lexical Nodes	
NMT	84.01
PBMT	89.60
Combo	81.10
Structural Nodes	
NMT	82.87
PBMT	80.96
Combo	79.10

Table 5: Systems plain HUME scores overall, and for lexical and structural nodes.

	HUME Score	**Lexical Score**	**Structural Score**
De1	82.16	88.08	70.84
De2	87.03	88.89	87.03
Cs1	82.32	82.16	82.63
Cs2	40.36	34.40	52.07
Pl1	67.91	65.96	71.84
Pl2	70.15	71.83	66.85
Pl3	66.56	62.10	74.82
Ro1	78.67	78.23	79.23
Ro2	84.61	86.55	80.94
Ro3	87.41	91.59	79.42

Table 6: HUME scores per annotator

	IAA Score	No. overlapping sentences
German	0.63392	100 doubly annotated
Czech	0.25132	100 doubly annotated
Polish	0.44015	30 triply annotated
Romanian	0.60720	50 triply annotated

Table 7: Inter annotator agreement for the different target languages and the number of overlapping sentences for which these IAA numbers have been calculated.

The Google translations were created on the 19 October 2016. We split the files into allowable chunks and then translated the files via the website. At this point Google was actively developing their neural machine translation production model and we are not sure if this had been deployed at this point or not. Unfortunately using a commercial system as a baseline is not very informative as the exact nature of the model is unknown and they are constantly changing their products.

	Cochrane	NHS24
Sentences	672	1257
HITS	224	419

Table 8: Number of sentences and HITS that were used in the ranking task

4.2 Software

We used the Appraise software that has been developed by Christian Federmann. It was initially developed eight years ago, and it has been the used in the four latest WMT (2012-1016) shared tasks. It has been their method for producing the official results of the competition. The tool shows the source sentence, the reference sentence and allows the users to score up to five machine translation system outputs. The translations get a score from Best to Worse on a 4 point scale, allowing ties. Figure 1 shows a screen-shot of this tool.

Figure 1: "Screen-shot of the ranking tool"

4.3 Annotators

The two user partners provided annotators for performing this evaluation. Cochrane annotators were same as those described for the HUME evaluation in Section 3.

NHS 24 ran the ranking task with annotators selected from a pool of people who were as close as possible to the real end users as possible. They ran a community engagement process to recruit Romanian and Polish speakers in Scotland. The number of annotators used in the Ranking Task were:

- 16 Romanian
- 13 Polish

The majority of interest from the public has come after the distribution of EU HimL project posters (See Figure 2) which have been translated to Polish and Romanian. These posters have been distributed via social media and support organisations such as the Polish Family Support Centre and University Student Associations.

The Romanian annotators consist mainly of secondary school pupils aged between 14 and 17. There are approximately 80 pupils studying at Shawlands Academy in Glasgow who are from a Romanian background. The remainder of Romanian annotators were sources from organisations such as Govanhill Housing Association and Crossreach. The Romanian community is not an easy one to engage with, and has well-known issues with poverty and illiteracy.

The Polish annotators come from various sources such as Universities, Health and Social Care and Service Users (via support organisations such as PFSC). The Polish community was easier to engage with as they are better integrated and established in Scotland however, many people were not interested in taking part in an NHS project. Some were disengaged due to the UK's decision to leave the European Union.

Initial contact was via the poster (in Polish or Romanian) and being sufficiently interested in the project to participate. Figure 2 shows the poster used to recruit Polish participants. The selection process for the volunteer annotators consisted of a short interview to confirm their first language and ability to communicate adequately in English, so that instructions for the task would be understood.

NHS 24 purchased a number of High Street 'Love2Shop' Vouchers which were used as a participation fee in lieu of cash. After taking part in tasks, each annotator was given a voucher (£10 for each task completed)

4.4 Results

The ranking evaluation was performed in order to obtain human judgments comparing alternative HimL translation systems. It complements the HUME evaluation because it gives us an overall score and it provides well understood methods for interpreting the ranking results and extracting statistical significance.

4.4.1 BLEU score results

We perform an initial analysis of the these systems performance as reported by the BLEU score. We used the mteval13b script in the MOSES software to compile these numbers. They are taken over untokenised and cased output.

System	German	Czech	Polish	Romanian
Y1	32.07	22.07	19.45	26.86
Y2	30.95	23.49	21.23	**34.93**
NMT	35.21	**29.60**	19.02	31.44
Google	**35.36**	27.06	**21.29**	32.75

Table 9: BLEU score results for the HimL Year 2 evaluation on all test data

In Table 9 we can see the scores assigned to the different systems across the entire test set. The Google translations obtain the highest scores overall for English-German and English-Polish. The NMT system obtains the highest score for English-Czech and English-Romanian is the only language where the year 2 system performs the best. Although the BLEU score results are not considered to be as reliable as human judgment, this table still provides strong support for the conclusion that in order to provide the highest possible translation quality to the HimL project, we should consider switching to neural machine translation models. NMT results are nearly 4.5 BLEU points ahead of year 2 system results for English-German, and more than 6 BLEU points ahead for English-Czech. For Polish and Romanian more experiments need to be run, but there are strong indications that with better data preparation, neural models could also perform better here than the phrase-based models.

System	German	Czech	Polish	Romanian
Y1	33.82	23.68	15.72	27.78
Y2	35.55	25.55	17.01	**37.09**
NMT	**38.52**	**33.41**	17.11	34.95
Google	37.56	29.78	**18.47**	35.39

Table 10: BLEU score results for the HimL Year 2 evaluation on Cochrane test data

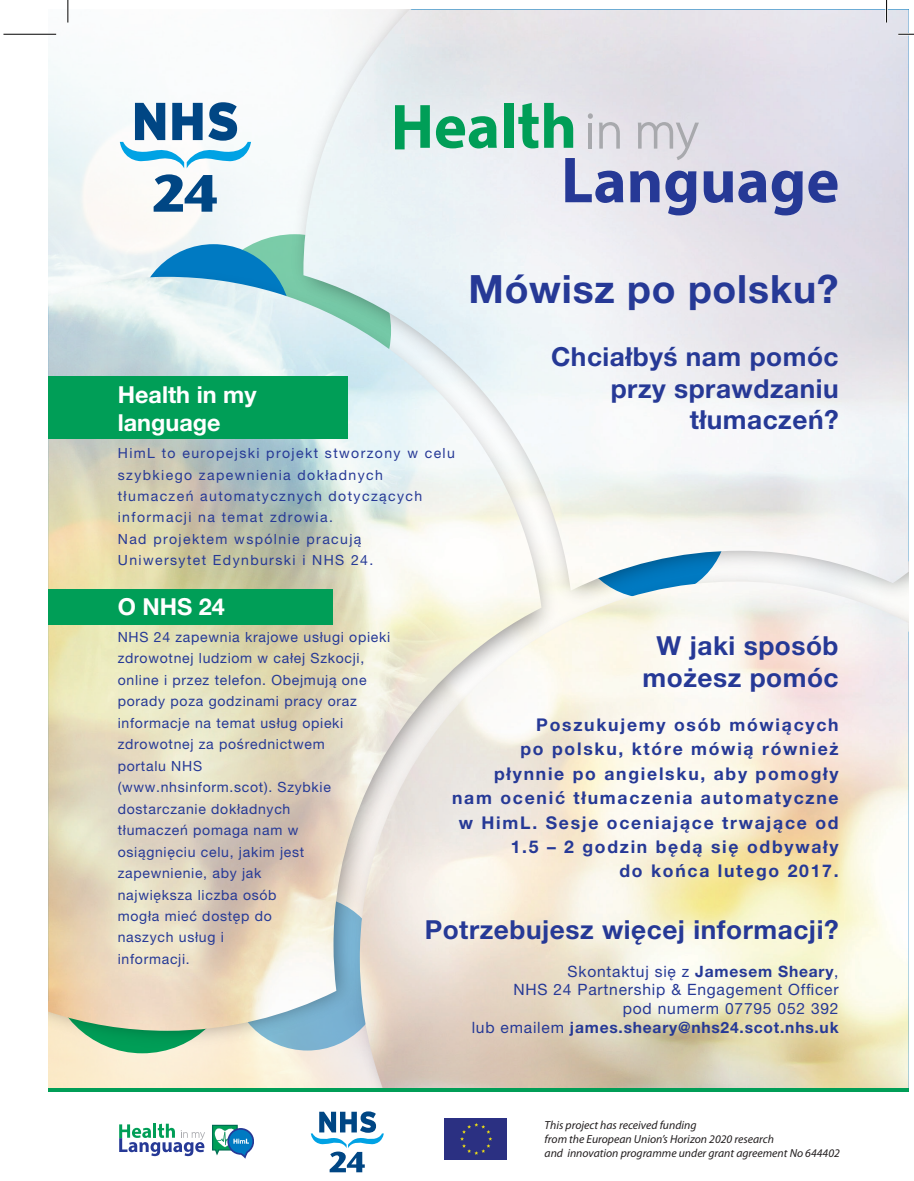

Figure 2: "Leaflet for recruiting Polish ranking annotators"

System	German	Czech	Polish	Romanian
Y1	30.09	20.42	23.58	25.87
Y2	26.23	21.34	**25.37**	**32.55**
NMT	31.69	**25.56**	20.90	27.70
Google	**33.10**	24.17	24.37	29.27

Table 11: BLEU score results for the HimL Year 2 evaluation on NHS24 data using

In order to understand these results better we have broken them down by looking at BLEU scores for the Cochrane and NHS24 test sets separately. In Table 10 we can see results for the Cochrane sentences and in Table 11 we can see results for the NHS24 sentences.

For the Cochrane results we can see that we beat Google on three of the 4 language pairs. It seems that our domain specific models are able to pull ahead of a model trained on a lot of general data. Also notable here is that out German year 2 system here is beating the year 1 system, whereas for the entire test set and for NHS24 data, the year 1 system beats our year 2 system. This is possibly because the year 2 system is more reliant on sentences having correct syntax and the NHS24 data contains many non-grammatical sentences.

For the NHS24 results, again we beat Google on three of the 4 language pairs. Here our English-German NMT system is beaten by Google, but our Polish year 2 system is stronger than all the rest. In fact the year 2 system seems to cope very well with the NHS24 data, coming first for Polish and Romanian. Perhaps the phrase-based models are more robust to ungrammatical sentences and sentence fragments like lists.

Although BLEU score results are very informative, it is still important to use human judgment as the final arbiter of quality. We now report results for a human ranking experiment which we would consider to be the official results for the HimL project.

4.4.2 Quantity

The first analysis of the data was to check how much of the data had been annotated. We removed a number of annotations which had been performed as tests and we can see in Figure 12 the total number of sentences that were ranked for the different test sets.

	Cochrane	NHS24	Total
German	660	1254	1914
Czech	669	1254	1923
Polish	669	1169	1838
Romanian	669	969	1638

Table 12: Number of sentences ranked using Appraise

4.4.3 Ranking results

We follow (Bojar et al., 2016) to extract ranking results from the raw ranking data of four ranked system for each test sentence. From these rankings, we produce pairwise translation comparisons, and then evaluate them with a version of the TrueSkill algorithm adapted to our task. We refer to this approach as the relative ranking approach (RR), so named because the pairwise comparisons denote only relative ability between a pair of systems, and cannot be used to infer their absolute quality. We use the TrueSkill method for producing the official ranking, in the following fashion. We produce 500 bootstrap resampled datasets over all of the available data (i.e., datasets sampled uniformly with replacement from the complete dataset). We run TrueSkill over each dataset. We then compute a rank range for each system by collecting the absolute rank of each system in each fold, throwing out the top and bottom 2.5%, and then clustering systems into equivalence classes containing systems with overlapping ranges, yielding a partial ordering over systems at the 95% confidence level.

We can see the results in Table 13. These results show that on the whole the neural models are preferred by the target users of the translation systems. For three out of the four language pairs the NMT model is first place, and in two of these cases it is significantly better than all the other systems. For English-Romanian the NMT system comes third. We hypothesis that this is due to the fact that the training data for Romanian is of mixed quality. There is inconsistency in the writing of diacritics in the training and testing data and it is likely that this makes a noticeable difference when translating into Romanian.

The other results worth noting is that Google systems perform strongly but that Google only beats our systems for English to Romanian. As we do not know whether the Google system was a neural machine translation model, it is hard to make conclusions other than that the field is moving very quickly during the period of this project, and that there are bound to be some anomalies at certain points in time.

English-German

System	Score
NMT	1.717
Google	0.545
Y1	-0.822
Y2	-1.440

English-Czech

System	Score
NMT	1.398
Google	0.169
Y2	-0.329
Y1	-1.238

English-Polish

System	Score
NMT	0.712
Google	0.493
Y2	-0.579
Y1	-0.626

English-Romanian

System	Score
Google	1.476
Y2	0.284
NMT	-0.174
Y1	-1.586

Table 13: Ranking results for the HimL Year 2 evaluation. Systems are ordered by their inferred system means. Lines between systems indicate clusters according to bootstrap resampling at p-level p $\leqslant$ .05. Systems in a cluster are considered tied.

4.4.4 User specific ranking results

In Table 13 we can see that for German, our year 2 system ranks at the bottom, even though it shares a cluster with the year 1 system and is therefore not significantly worse. In order to understand why this is the case we analyse the results individually for the ranking data from the different user partners. In Table 14 we can see the results.

These tables show us that there are quite different results depending on where the data came from. The year 2 system beats the year 1 system for the Cochrane data but it is significantly behind the year 1 system for the NHS24 data. This could be because the NHS24 data contains many non grammatical sentences either as part of a title, or as part of a list. The year 2 system relies heavily on linguistic tools which are not designed to cope with this kind of data.

All data

System	Score
NMT	1.717
Google	0.545
Y1	-0.822
Y2	-1.440

Cochrane data

System	Score
NMT	1.710
Google	0.535
Y2	-0.837
Y1	-1.407

NHS24 data

System	Score
Google	1.265
NMT	0.993
Y1	-0.556
Y2	-1.702

Table 14: Ranking results for the HimL Year 2 German evaluation.

4.5 Discussion

The results from this section indicate that the neural machine translation models are performing better than previously state-of-the-art phrase-based translation models as judged by humans. The exception is English-Romanian, where we suspect noisy training data may impact the performance of the neural system. Based on this evidence, we need to consider whether NMT is suitable for deployment in Year 3 of the project.

5 Outlook

For year three of the project, we will be pursuing the following activities:

- We will analyse the performance of our models on the HimL test set to determine what improvements are necessary for the consumer health domain.
- We will evaluate research done on domain adaptation, semantics and morphology to determine which deliver performance improvements for the HimL test set.
- We will be continuing development of the human semantic metric based on feedback from the second evaluation experiment.
- We further investigate automating the human semantic metric.
- We will implement comprehensive user acceptance and impact studies including another ranking evaluation and surveys of NHS24 and Cochrane users.

References

Birch, A., Haddow, B., Bojar, O., and Abend, O. (2016). Hume: Human ucca-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., et al. (2016). Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*, volume 2, pages 131–198.