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

D2.2: Intermediate report on employing semantic role labelling and fidelity checking in machine translation

Author(s): Ondřej Bojar, Ondřej Hübsch, Maria Nadejde, David Mareček, Roman Sudarikov, Phil Williams

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	D2.2
Deliverable title	Intermediate report on employing semantic role labelling and fidelity checking in machine translation
Type	Report
Status and version	1.0
Number of pages	14
Contributing partners	CUNI, UEDIN
WP leader	CUNI
Task leader	CUNI
Authors	Ondřej Bojar, Ondřej Hübsch, Maria Nadejde, David Mareček, Roman Sudarikov, Phil Williams
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

Executive Summary

The goal of WP2 Semantically Motivated MT is to improve the accuracy of statistical machine translation. We address the two frequent error types in the first two tasks: T2.1 aimed at preserving semantic roles of verb complements and modifiers, and T2.2 aimed at correct handling of negation.

Two additional tasks complement these efforts in a more direct way: T2.3 aims at avoiding critical translation errors of phrase-based (and hierarchical) models by removing translation units that are likely to introduce errors, and T2.4 aims to make MT benefit from existing large-scale high-quality dictionaries developed by Lingea.

This deliverable 2.2 is devoted to the work on semantic role labelling and on core fidelity in year 2 of the project. We refer the reader to report D2.1 for a description of the first year's work. The final deliverable of the workpackage (D2.3) will cover the results of the remaining tasks.

1 Task 2.1: Modelling Semantic Role Labelling in Machine Translation

Semantic roles specify the relationships between the verbs of a sentence and their arguments. For instance, in the sentence 'John likes Mary,' the arguments 'John' and 'Mary' have different roles: John is the person who loves and Mary is the person who is loved. These distinct roles are often given generalized labels like 'agent' (for the person or thing performing the action of the verb) and 'patient' (for the person or thing to whom the action of the verb is directed). For a translation to faithfully convey the meaning of a source text, it is crucial that the semantic roles of the source text are correctly mapped to the target language. For instance, 'John loves Mary' can be correctly translated to the German 'John liebt Mary,' but if during translation the phrases 'John' and 'Mary' are reordered, then their semantic roles will be flipped.

Most machine translation models have no explicit concept of semantics, although there is now a small but growing body of literature on modelling semantic roles in machine translation. To date, the majority of this work has focused on Chinese-English translation (among other reasons, this partly reflects the availability of suitably annotated corpora and language processing tools for both of those languages).

In the first year of the project, Task 2.1 was focused on shallow syntactic statistical MT models. At the start of the project, this was the natural direction, since almost all pre-existing work on using semantic labels in MT was based on syntax-based statistical approaches. During the second year, we made a transition from syntax-based statistical MT to neural MT (NMT). This was motivated by the following observations and findings:

- The overall field of machine translation is rapidly undergoing a shift from statistical to neural MT on the basis of compelling empirical results. For many language pairs and data sets, the translation quality of neural MT has surpassed that of statistical MT (see, for example, the results of WMT16 (Bojar *et al.*, 2016b)). Where neural MT hasn't yet surpassed statistical MT, it is rapidly gaining ground.
- Parallel work in the research-oriented QT21 project has shown superior results for neural MT over statistical MT in three out of the four HimL language pairs (English-Polish was not tested). Encouragingly it has also demonstrated improvements from the use of shallow syntactic and semantic labels. This is in contrast to statistical MT where (for the HimL pairs) the use of syntax has generally harmed translation quality.
- Looking ahead to building final systems, neural MT is much more straightforward to combine with phrase-based MT (the default model type in HimL) than syntax-based MT. This is due to the ability of neural MT to score arbitrary sentence pairs making it simple to rescore phrase-based n-best lists with a neural MT model.

In this section, experiments in four settings of semantic role labelling (SRL) are explored. The first part is devoted to the use of SRL coming from PropBank-like annotation in both syntax-based SMT (continuation of our work in year 1) as well as NMT. The second part makes use of SRL labels coming from a slightly deeper t-layer style of analysis. Again, both traditional statistical MT (here phrase-based MT) and neural MT are covered. Note that the SRL annotation is used only on the English side so it can be used for all HimL language pairs. The t-layer experiments are limited to English-to-Czech.

In the last part of this section (Section 1.5), we describe our experiments with selectional preferences, a phenomenon very closely related to semantic role labelling. While semantic roles are concerned with morphosyntactic forms expressing syntactic dependencies, selectional preferences provide lexical support for this. By improving either or both, the core relations in the translated sentence are more likely to be preserved.

System	Syntactic		Semantic	
	l-m	r-m	l-m	r-m
English-Chinese (Li et al)	80.7	85.6	70.9	73.5
English-German (reimplementation)	81.6	85.8	63.8	67.9

Table 1: Accuracy of the syntactic and semantic reordering models when tested on gold test sets. Results are given for the leftmost (l-m) and rightmost (r-m) variants of the models (see Li *et al.* (2014) for a full description of the models). English-Chinese results are from Li *et al.* (2014) and English-German results are from our reimplementation.

	en-zh	en-de			en-ro	
	News	Cochrane	NHS24	Khresmoi	Cochrane	NHS24
Phrase-based	11.8	**35.5**	**28.0**	**18.4**	**34.0**	27.1
Hierarchical phrase-based	**11.9**	34.7	27.4	17.9	32.6	**27.3**

Table 2: Bleu scores for phrase-based and hierarchical phrase-based systems.

1.1 Statistical MT with PropBank Semantic Role Labels

As described in D2.1, we experimented in the first year with a number of approaches to incorporating SRL labels. A promising technique was based on previous work by Li *et al.* (2013), in which the authors showed improvements for the Chinese-English language pair. However, when applied to the HimL language pairs, our reimplementation was found to perform less well. Whereas applying hard syntactic constraints (a prerequistite of Li *et al.*'s approach) was found to improve translation quality for English-Chinese, the opposite was found for English-German and English-Romanian. For this reason, we subsequently adopted the closely-related approach taken in the follow-up paper (Li *et al.*, 2014). In brief, the basic idea is to extend a hierarchical phrase-based model by adding independent syntactic and semantic reordering models based on word alignments. This has the advantage that it eliminates the requirement for hard syntactic constraints. Although not fully implemented (due to the neural MT switch), our partial implementation provides further evidence of the difficulty of semantic role reordering in English-German.

Implementing Li *et al.* (2014)'s model within the Moses toolkit requires four main steps: i) adding SRL annotation to training data trees; ii) extracting reordering types and features from the training data; iii) learning the syntactic and semantic reordering models; and iv) implementing a reordering feature function. The first step was implemented earlier in the project and required no additional engineering effort. We implemented the second and third steps, using Vowpal Wabbit[1] as the maximum-entropy classifier, and tested the resulting reordering model. For English-German, we received similar results to those reported by Li *et al.* (2014) for English-Chinese. Table 1 gives the accuracy of the syntactic and semantic reordering models when tested on gold test sets. Interestingly, the accuracy of the semantic reordering model is lower for German than Chinese. A possible interpretation is that prediction is made more difficult by the higher degree of semantic role permutation in English-German. In the analysis of the work in year one, we observed Kendall-Tau distances of 0.79 for English-German and 0.84 for English-Chinese (lower scores indicate a higher degree of reordering, see D2.1 for details).

The fourth implementation step is the most time consuming and has not yet been implemented. Based on the comparative performance of syntactic versus neural MT, we instead focused our efforts on neural MT instead. Table 2 gives Bleu scores for three language pairs, English-German and English-Romanian, plus English-Chinese (en-zh), which we used as a point of comparison. For the two HimL language pairs, using a hierarchical phrase-based model has a negative impact on translation quality overall.

1.2 Neural MT with PropBank Semantic Role Labels

For our neural MT experiments, we used the Nematus toolkit[2] and a similar setup to that reported in Sennrich and Haddow (2016). Thus, our system is an encoder-decoder with an attention mechanism. We use BPE (byte-pair encoding) to segment words into subword units in a preprocessing step, The main difference in our setup is that we do not use back-translated monolingual data. This is due to the need to keep computational requirements low enough for rapid experimentation during initial development. However, we anticipate that final neural systems are likely to use back-translation since it has shown consistently good results, for instance in Sennrich *et al.* (2016).

Prior to BPE segmentation, the English source-side data was annotated using the MatePlus tool[3]. This adds several layers of

[1] http://hunch.net/ vw/

[2] https://github.com/rsennrich/nematus

[3] https://github.com/microth/mateplus

	en-cs		en-de		en-pl		en-ro	
	cochrane	nhs-24	cochrane	nhs-24	cochrane	nhs-24	cochrane	nhs-24
baseline	**30.41**	21.93	33.90	**28.52**	17.08	20.68	**34.71**	26.86
SRL	30.20	21.88	**34.20**	27.85	**17.12**	19.94	34.28	**27.32**
all	30.30	**23.10**	33.50	28.51	16.85	**21.34**	**34.71**	27.00

Table 3: Neural MT results with surface form input only ('baseline' row) verbal-N and nominal-N features only ('SRL' row), and all MatePlus features ('all' row).

annotation, from which we extract 10 factors:

- **form** - surface form of the token (e.g., maintaining)
- **lemma** - surface lemma (e.g., maintain)
- **tag** - part-of-speech tag (e.g., VBG)
- **dep** - dependency label (e.g., PMOD)
- **verbal-N** (for N=1,2,3) - role of word in frame for verbal predicate at depth N (where N=1 is closest to the root of the dependency tree). These factors are PropBank labels (e.g. PRED, A0, AM-TMP).
- **nominal-N** (for N=1,2,3) - As verbal-N but for nominal predicates.

After BPE segmentation, we added a further BIOE tag indicating the position of this subword token within the enclosing word (as in Sennrich and Haddow (2016)).

Due to the large number of experiments, we used a subset of the training data. Specifically, for each language pair, we sampled subcorpora of 4M parallel sentence pairs from the HimL Corpus described in report D1.1. Since the constituent sub-corpora are of wildly differing sizes (for instance, for English-Romanian, there are 80M sentence pairs in the OpenSubtitles subcorpus and 740k sentence pairs in the EMEA subcorpus), we used a balanced sampling strategy that chose (without replacement) one sentence pair from each sub-corpus in a round-robin.

Table 3 gives BLEU scores for the Cochrane and NHS24 test sets. Disappointingly, the results fail to show consistent improvements. While we see a gain of 1 BLEU point for using all factors for English-Czech for the NHS24 test set, there is a (small) loss on the Cochrane test. Across the language pairs, there are no obvious patterns, with BLEU scores sometimes increasing and sometimes decreasing. This is in contrast with the findings in QT21 where consistent improvements were observed for German and Romanian using WMT data. Since the work in QT21 is ongoing, and since QT21 is a research-focused project, we will explore this more deeply in collaboration with that project.

1.3 Neural MT with PDT Tectogrammatical Semantic Role Labels

This section is devoted to the experiments with neural MT which use semantic roles labels as defined the in the tectogrammatical layer, see the Prague Dependency Treebank (Hajič *et al.*, 2006). The tectogrammatical layer is available for both Czech and English, but we focused on the English side only.

We start by briefly introducing the exact data used in these experiments (Section 1.3.1). Note that the same data is then used also in English-to-Czech phrase-based experiments (Section 1.4).

1.3.1 Data Selection and Annotation

Experiments with semantic role labelling require computationally expensive processing of the data and the informed models are in general larger and more difficult to handle. For that reason, we carry out our experiments with only a small subset of all available data. However, we select a subset that is most relevant for our domain, as described in Deliverable 1.1.

English-to-Czech experiments with modeling SRL were based on the following subcorpora, all extracted from CzEng 1.6 (Bojar *et al.*, 2016a):

- CzEngMed – medical section of CzEng 1.6, 1.5 million sentences
- CzEngTop1 – top 1% from CzEng 1.6 scored using XenC data selection tool described in D1.1 with CzEngMed as domain-specific data, 1 million sentences

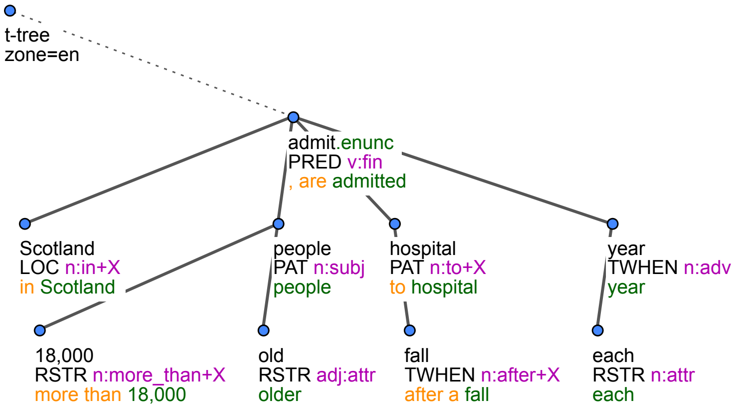

Figure 1: Tectogrammatical representation of sentence "In Scotland, more than 18,000 older people are admitted to hospital after a fall each year". The SRL factors we extract are *functors* (in black in the second line at each tree node) and formemes (in purple in the second line). The *valframe* values are not shown, however they are only identifiers pointing to the valency lexicon.

- CzEngTop5 – top 5% from CzEng 1.6 scored using XenC with CzEngMed as domain-specific data, 4.5 million sentences

The corpora were automatically parsed up to the deep syntactic representation (tectogrammatical layer of Prague Dependency Treebank (Hajič *et al.*, 2006)) using the Treex toolkit [4]. One example of automatically analyzed sentence is in Figure 1.

We extracted the following factors from the tectogrammatical representation:

- **form** - surface form of the token
- **lemma** - surface lemma
- **tag** - part-of-speech tag
- **functor** - represents the semantic value of a syntactic dependency relation; it expresses the function of an individual modification in the sentence. Examples: ACT (actor), PAT (patient), ADDR (addressee), LOC (location), DIR3 (direction to), TWHEN (temporal), ...
- **formeme** - string representation of selected morpho-syntactic features of the content word and selected auxiliary words that belong to the content word. Examples: adj:attr (attributive adjective), n:obj (nominal object), na:v+4 (noun in accusative with preposition *na*), ...
- **valframe** - valency frame of verbs; pointer to the valency lexicon
- **functor-at-verbs** - functor of verbs only
- **formeme-at-verbs** - formeme of verbs only
- **functor-at-deps** - functor of words depending on verbs
- **formeme-at-deps** - formeme of words depending on verbs

The semantic factors are defined only for content words, some of them only on verbs or their dependents. In case the factor is not defined for a particular token, it gets the default value "_". Therefore, e.g. the function words (prepositions, auxiliary verbs, determiners, etc.) have only the basic factors (form, lemma, and tag) filled.

Note that not all the factors are used in all our experiments, see the details below.

1.3.2 Experiments with English-to-Czech NMT

We used Nematus[5] NMT tool on CzEngMedical corpus. We trained two translation models, one without SRL using only *stc*, *lemma*, and *tag* factors and one with additional SRL factors *formeme*, *valframe*, and *functor*. Both models use the same total dimension of word vector 500, distributed across factors as [380, 110, 10] and [360, 100, 10, 10, 10, 10] respectively.

[4] http://ufal.mff.cuni.cz/treex

[5] https://github.com/rsennrich/nematus

Source-side factors	BLEU after 42000 iterations
stc+lemma+tag+formeme+valframe+functor	17.23
stc+lemma+tag	16.82

Table 4: Nematus results for English-to-Czech

Table 4 provides the results of this experiment. The setup with access to the SRL information (formeme, functor and the valency frame) gives better results, allowing us to conclude that target-side SRL information is likely to help. It is however important to say that the overall translation quality of that system is still worse than year 2 HimL system, the complex hybrid of deep-syntactic transfer, phrase-based MT and the final automatic correction of grammatical errors.

1.4 Experiments with English-to-Czech SMT

For SMT-based experiments we used Moses SMT system. The basic configuration is derived from the Chimera system that combines standard phrase phrase table coming from a parallel corpus with a synthetic phrase table produced by a transfer-based system (TectoMT). The common settings for all the experiments are the following:

- Phrase table source-side factors are true-cased word form for baseline adding formeme/valframe/functor for different models.
- Phrase table target-side factors are true-cased word form, lemma, and morphological tag in all setups.
- Three language models are applied to true-cased word forms, lemmas, and morphological tags, resp.
- The phrase tables are interpolated using MERT on HimL development corpus.
- Final evaluation was performed using HimL test corpus.

Table 5 shows the different setups of phrase tables, that were used in the experiments.

System	Phrase Table 1	Phrase Table 2	Phrase Table/VowpalWabbit Model 3
Baseline	CzEngMed (stc)	TectoMT (stc)	-
Run 1	CzEngMed (stc+formeme)	TectoMT (stc+formeme)	-
Run 2	CzEngMed (stc+functor)	TectoMT (stc+functor)	-
Run 3	CzEngMed (stc+valframe)	TectoMT (stc+valframe)	-
Run 4	CzEngMed (stc)	TectoMT (stc)	CzEngMed (stc+formeme)
Run 5	CzEngMed (stc)	TectoMT (stc)	CzEngMed (stc+functor)
Run 6	CzEngMed (stc)	TectoMT (stc)	CzEngMed (stc+valframe)
Run 7	CzEngMed (stc)	TectoMT (stc)	VowpalWabbit CzEngMed (stc+valframe)
Run 8	CzEngMed (stc)	TectoMT (stc)	VowpalWabbit CzEngTop5 (stc+valframe)
Run 9	CzEngMed (stc)	TectoMT (stc)	VowpalWabbit CzEngMed (RICHER)
Run 10	CzEngMed (stc)	TectoMT (stc)	VowpalWabbit CzEngMed (RICHER+valframe)
Run 11	CzEngTop5 (stc)	TectoMT (stc)	VowpalWabbit CzEngTop5 (RICHER)
Run 12	CzEngTop5 (stc)	TectoMT (stc)	VowpalWabbit CzEngTop5 (RICHER+valframe)

Table 5: English-to-Czech SMT phrase tables with source side factors.

Baseline uses just the two phrase tables. Runs 1–3 enrich the source side of the phrases table with one factor carrying one of the variants of the SRL information: formeme, functor or valency frame. In these runs, both the corpus and the TectoMT tables are enriched. Runs 4–6 enrich with SRL only the corpus table and more importantly, add this enriched table as a third option, allowing the model to resort to less sparse (but also less informative) baseline tables.

Runs 7 and 8 do not use the third phrase table with semantic factors, but instead rely upon a discriminative model (Tamchyna and Bojar, 2015) that scores phrase translation candidates based on the source factors of form and valency frame and the target factors of form, lemma, and morphological tag.

Finally, Runs 9–12 (labelled "RICHER") use a richer feature set in VowpalWabbit: not only the true-cased word form (stc) but also lemma and morphological tag. The runs differ in training data size (smaller CzEngMed or larger CzEngTop5) and use or do not use the valframe factor.

System	BLEU	Avg. BLEU
Baseline	24.34 [23.36, 25.32]	24.3
Run 1	23.93 [22.98, 24.96]	-
Run 2	23.45 [22.48, 24.43]	-
Run 3	24.13 [23.15, 25.14]	-
Run 4	24.42 [23.41, 25.48]	24.3
Run 5	24.29 [23.28, 25.26]	24.3
Run 6	24.39 [23.39, 25.46]	24.2
Run 7	24.38 [23.41, 25.45]	24.6
Run 8	24.64 [23.54, 25.74]	24.8
Run 9	24.45 [23.41, 25.51]	24.7
Run 10	24.68 [23.62, 25.73]	24.8
Run 11	24.57 [23.60, 25.61]	25.1
Run 12	24.88 [23.86, 25.99]	25.0

Table 6: Results of English-to-Czech translation with source-side SRL information. 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 results show no improvement when the SRL information is used only to enrich phrase table entries, regardless the particular variant of the information (formeme/functor/valframe) and also regardless if all tables are enriched or if the enriched table is added just as an additional option. On the other hand, adding the discriminative model (Run 7) improves the results on average by 0.3 BLEU. We are able to further improve this result by training on more data (Run 8), adding 0.5 BLEU to the baseline.

Runs 9–12 explore richer source-side features (stc+lemma+tag) and show, that the same improvement can be achieved using smaller training data with richer semantic information for discriminative model. The highest average BLEU is achieved in Run 11 by using larger amount of training data with richer source-side features in the discriminative model. It is worth noting that Runs 9–10 and Runs 11–12 do not significantly differ from each other, so in this richer setting, SRL information (in the form of t-layer valency frame) brings no further improvement, with p-value 0.17 and 0.52, resp. The significance testing was performed using MultEval (Clark *et al.*, 2011).

To conclude, we were able to improve translation quality from English to Czech by adding semantic role information derived from the tectogrammatical annotation. This additional information has to be integrated using a discriminative model, not simply to augment source-side factors. Similar improvements are however achieved with the discriminative model alone, relying on morphological features only.

1.5 Selectional Preferences Model for String-to-tree SMT

Li *et al.* (2013) proposed to improve translation of predicate-argument structures by modeling reordering and deletion of semantic roles. However, the proposed models do not encode information about the lexical semantic affinities between target predicates and their argument fillers. Selectional preferences describe such semantic affinities. For example, the verb “drinks” has a strong preference for arguments in the conceptual class of “liquids”. Therefore the word “wine” can be disambiguated when it appears in relation to the verb “drinks”.

As part of the HimL project, Nadejde *et al.* (2016a) explored whether modeling selectional preferences is useful for translating ambiguous predicates and arguments. The authors propose a selectional preference feature for string-to-tree statistical machine translation based on the information theoretic measure of Resnik (1996). The feature models selectional preferences of verbs for their core and prepositional arguments as well as selectional preferences of nouns for their prepositional arguments. It also uses unsupervised clusters to generalize over seen arguments.

In Table 7 we present the main results of the paper for the German→English language pair. The baseline string-to-tree system trained on WMT data is compared with two augmented string-to-tree systems: one with the Selectional Preferences feature and one with a RDLM feature. The RDLM–P_w (Sennrich, 2015) is a feed-forward neural network which predicts the head word of a syntactic dependent conditioned on a large syntactic context which includes ancestors and siblings. The HWCM metric (Liu and Gildea, 2005) is an f-score over syntactic n-grams which captures the improvement in translation quality for long-distance dependencies. The results on the WMT newstest2013, 2014 and 2015 showed that neither of the features improves automatic evaluation metrics. After further analysis the authors concluded that mistranslated verbs are negatively impacting these features. The authors have therefore addressed the problem of mistranslated verbs with a Neural Verb Lexicon Model (Nadejde *et al.*, 2016b) as part of the more research oriented European project, QT21.

While this pilot study was conducted on the German→English language pair, this method is not language specific. Therefore we can apply it to any target language for which a syntactic parser is available. A future direction could be to combine the Selectional Preferences model and Neural Verb Lexicon model for re-ranking the output of either phrase-based MT or NMT. In the case of NMT, these models might also help to resolve word sense disambiguation errors.

System	BLEU	HWCM
Baseline	26.45	24.47
+ SelPref	$26.48_{+.03}$	$24.54_{+.07}$
+ RDLM–P_w (1, 0, 0)	$26.35_{-.10}$	$24.75_{+.28}$

Table 7: Results for string-to-tree systems with Selectional Preferences (SelPref) and RDLM features on the WMT newstest2013, 2014 and 2015. The number of clusters used with SelPref is 500. The triples in parenthesis indicate the context size for ancestors, left siblings and right siblings respectively. The RDLM configuration (1, 0, 0) captures similar syntactic context as the selectional preference feature.

2 Task 2.3: Improving Core Fidelity of Shallow Models

The goal of Task 2.3 is to make sure that the verity of standard shallow models such as phrase-based translation is not degraded by individual phrase table entries. In the ideal case, every phrase entry in the phrase table would be correct and errors could get introduced only at phrase boundaries. More information can be found in Deliverable 2.1.

In our analysis carried out in months 1-9 and reported in the Interim progress report, we established that there are fewer situations where such phrase-local checks make sense compared to our expectations when writing the project proposal. We thus focused on the most promising type of errors, negation flip.

Some of the possible approaches to avoid a negation flip are:

1. Translation to a simplified target language (without double negative) (details proposed in Section 2.1 below)
2. Refinement of word-alignments in the corpus (experiments described in previous D2.1, brief summary here in Section 2.2)
3. Filtering phrases that flip negation from the phrase table (experiments described in Section 2.3 below)
4. Scoring phrases with an additional score (Section 2.4)

It is also possible to extend the scope back beyond negation flip and devise an additional score for phrases in the phrase table that would reflect some bilingual “similarity”. This approach was presented in Zhang *et al.* (2016) and we describe our attempt to apply it for English-to-Czech medical data in Section 2.5.

Parts of this section were previously reported in the non-public Deliverable 7.4.

2.1 Translation to a Simplified Target Language without Double Negation

Instead of a English-to-Czech translation, we could translate from English to a “simplified” version of Czech that doesn’t contain double negative and then change some forms to negative again. The translation would be dependent on the ability to properly convert to and from the intermediate form of the target language. The dependence of the method on the particular language is the main reason why we decided not to pursue this avenue.

2.2 Refinement of Alignments

To avoid negation flip, we can look for pairs or larger sets of words that jointly express a (single) negation in the target language. If all these words are linked with word-alignment links, the phrase extraction mechanism cannot break them into separate phrase pairs. However, the alignment methods often ignore the polarity of the words. This is particularly likely e.g. for Czech, where the alignment is based on lemmas and lemmas do not express word polarity. The alignment can be artificially augmented by links connecting all words belonging to one negation occurrence, or, as a simpler proxy, negation flag can be preserved on lemmas. We already did some experiments in this direction in D2.1 and we decided not to continue with them for this deliverable.

	Czech	German	Polish	Romanian
Sentence pairs	52.6 M	9.1 M	57.9 M	84.0 M
Words (English)	638 M	179 M	556 M	727 M
Words (the other language)	544 M	167 M	447 M	689 M
Extracted mistranslated word pairs	1279	275	582	612
Phrase pairs (filtered to development set)	29.1M	84.8M	2.4M	27.6M
Number of affected phrase pairs	3406	2817	801	2089

Table 8: Parallel corpus sizes and extracted mistranslated phrase pairs.

2.3 Filtering Phrases that Flip Negation

In our first experiments for English-to-Czech, we detected phrases with likely negation flip by checking the presence of the English word “not” but absence of any Czech word with marked morphological negation. Removing such phrase pairs increased the BLEU score slightly (23.29 → 23.36) and this small gain was also confirmed in a manual check.

This implementation was clearly limited: it relied on the availability of morphological analysis (which recognizes negation) of the target language and it was unable to identify phrase pairs where the opposition in the meaning is not expressed by a morpheme but rather lexically (e.g. “cheap” vs. “expensive”). We thus designed a method that employs knowledge of opposite words in the source language only (English in our case), e.g. by consulting a lexical database like Wordnet[6], or by using simple morphological rules, such as adding the prefix “un-” or “in-”. A word-aligned corpus to any other target language is then sufficient to extract *wrong word pairs*, i.e. pairs of words that are related to each other but express the opposite meaning. Following our example, the good translation pairs are “cheap” = “levný” and “expensive” = “drahý” while the crossed pairs “cheap” ≠ “drahý” and “expensive” ≠ “levný” need to be avoided in the phrase table. The method is based on the observation that current word-alignment methods are likely to ignore the polarity of the word, so they will align the good pairs as well as the bad pairs, but the polarity of individual words will be preserved in the training sentence pairs more often than flipped. Statistical measures like pointwise mutual information will be thus stronger for the good pairs than for the bad pairs.

We implemented this method and employed it for all HimL languages, relying on the mentioned English Wordnet and prefixation rules. Random examples of the extracted word pairs are given in Figure 2. We see that a large majority of the extracted pairs are indeed mistranslations and most often true antonyms. The automatic extraction is controlled by several thresholds. We experimentally fine-tuned these for our English-Czech data, using a small manually constructed list of wrong pairs. More permissive thresholds lead already to a decrease in precision. We then applied the same thresholds to other languages. We see in Table 8 that the number of extracted word pairs is very low, and even lower for languages other than Czech, due to the different linguistic (esp. morphological) properties and corpora sizes and repetitiveness.

The small number of extracted mistranslated pairs leads to a very small number of phrases removed from the phrase table, which in turn leads to negligible or no effect at all at translation time.

Surprisingly, when we used the filtered phrase table for model optimization, the resulting performance was substantially worse (BLEU of 23.74 instead of 24.16); again with no difference if the test set is then translated with the filtered or non-filtered phrase table. This result suggests that it might be actually possible to improve translation quality using the “bad pairs” removed in the core fidelity filtering.

In the following section, we describe our experiments in this direction, augmenting (as opposed to filtering) the phrase table with several novel scores in an attempt to strike the right balance between removing and keeping suspicious phrase pairs.

2.4 Scoring Phrases in the Phrase Table with Additional Scores

This section describes our experiments carried out towards the end of Task 2.3. The goal was to compare filtering of the phrase table with a potentially less harmful method: adding a score or some feature indicating whether a phrase pair is likely to flip a negation. We briefly describe the experiments in the following subsections. All of the experiments use the same set of forbidden words that was described in 2.3.

2.4.1 Indicator 0.5 / 0.8

The intention is to indicate that we think some phrases might have a negation flip problem. This is the most straightforward softening of the hard filtration described in Section 2.3. Instead of strictly removing such a suspicious entry, we mark it with a score of 0.5 (meaning that there is 50% chance that this entry is wrong). Otherwise, we add 0.8 indicating that we believe the

[6] http://wordnetweb.princeton.edu/perl/webwn

English	Czech	Gloss
long	zkráceně	short
eastern	západní	western
~ liability	majetek	property
~ pull	odvézt	carry away
night	svítání	dawn
× pull	netlačit	not-push
~ senior	maturit’ák	graduation ball
× hit	neminout	not-miss
easy	rychle	fast
top	dole	down

English	German	Gloss
~ surgical	medizinisch	medical
insufficient	ausreichen	suffice
~ indoor	Schwimmbad	(indoor) swimming pool
~ employ	einsetzen	deploy
insufficient	reichen	suffice
unsaturated	sättigen	saturate
father	Mutter	mother
~ female	Genitale	genital
~ employ	benutzen	use, deploy
× yes	sicherlich	sure

Figure 2: Random sample from 1279 English-Czech and 275 English-German automatically extracted mistranslated word pairs. We see that for instance for Czech, 2 of these sample 10 entries are wrong (they do not flip the negation, marked with "×") and 3 other entries are not exactly opposite terms (marked "~") but they neverthless distort the meaning so they should be removed from the phrase table.

pair should be correct. We don't add 1 because the weight for this feature in the log-linear model would then be completely ignored for these phrases.

2.4.2 Bucketing

We noticed that our filtering is more reliable for shorter phrases because longer phrases are more likely to contain further words of the negation, correctly preserving polarity of the whole phrase. We thus introduce separate indicators for different lengths of phrases.

A phrase entry belongs to a bucket described by an interval when the length of both the source and target sides belongs to that interval. For a given phrase entry, the indicator value for irrelevant buckets (the buckets that this entry doesn't belong to) is set to 1 which effectively causes that this value is ignored. The value of the feature for the relevant bucket is set in the same way as in Section 2.4.1. In one experiment, we also try to combine both the bucketed and the comprehensive indicators.

2.4.3 Number of wrong word pairs

To reflect the fact that phrase pairs differ in the number of suspicious word pairs, we test a very simple feature: for every phrase pair we add the number of source-target word pairs that were in the list of forbidden word pairs. More precisely, we add the exponential of that value, to make the feature behave as a counter in the log-linear model.

2.4.4 Estimating probability

Our last considered feature builds upon the observation that more suspicious word pairs in a phrase indicate that the phrase is complex and difficult to categorize, rather than the phrase would be more clearly wrong.

For languages like Czech that use a double negative, this reflects that property in some way: a lot of frequent cooccurrences of forbidden words might actually make the phrase pair correct ("nikdy nechci nic" (lit: never do-not-want nothing) and "vždycky chci něco" (always want something) are semantic equivalents while "nikdy"/"vždycky", "nechci"/"chci" and "nic"/"něco" are all pairs of antonyms).

The score is set as follows: For every phrase without any forbidden source-target word pair, we set the score to 0.99. For other phrases, we count the number c of forbidden source-target word pairs among all word pairs. Then we set $\min(0.99, c/100)$ as the score for this entry.

The results for all considered scorings of phrase pairs are listed below.

2.5 Semantic Similarity Score Using Neural Networks

Zhang *et al.* (2016) propose a method that uses bidimensional attention based recursive autoencoder (BattRAE) to get bilingual phrase representations. BattRAE aims to embed phrases into a vector space where all phrases with a similar meaning are close to one another according to some measure. Then they use a bilinear neural model to measure bilingual semantic similarity and use the calculated score as a feature in an existing SMT model. The results seem very promising: a significant BLEU improvement on both MT06 (31.55→33.19) and MT08 (23.66→25.29) test data for the Chinese-English translation task.

	CzMed1	CzMed2	Y2CZ	Y2DE	Y2PL	Y2RO
Baseline	20.93	20.88	22.17	24.83	19.03	26.14
Table filtering 2.3	20.84	20.88	22.09	25.24	18.67	25.77
Indicator 2.4.1	20.97	20.89	22.18	**25.41**	18.67	26.09
Buckets 2.4.2 (2, 3, 5, rest)	20.83	–	–	–	–	–
Buckets 2.4.2 (2, 3, 5, rest + 2.4.1)	20.62	–	–	–	–	–
Buckets 2.4.2 (≤ 4 and rest)	–	**21.17**	22.06	DB	19.02	25.39
Number of wrong pairs 2.4.3	–	20.86	–	–	–	–
Estimated probability 2.4.4	–	20.99	**22.21**	25.06	18.71	25.95
BattRAE 2.5	20.95	–	22.09	–	–	–

Table 9: Experimental results of core fidelity checking for English-to-Czech using both hard and soft methods of phrase table filtering. Improvements over the baseline in bold.

Such a semantic measure could help our task of core fidelity checking in a more general way than the negation flip examined above.

The sources are publicly available at https://github.com/DeepLearnXMU/BattRAE and one needs to provide examples of good and bad phrases to train the model. We followed the technique of the authors: We used force-decoded phrases from our baseline translation to generate positive samples for the test data. BattRAE expects to receive all positive samples complemented with a negative sample. We generated negative samples by replacing every word in the positive sample by a random word from the same language. Initial word embeddings were trained on a big bilingual corpus using word2vec.

We tried to replicate the results of Zhang *et al.* (2016) on our Czech medical corpus (1M sentences) but there wasn't any improvement on the HimL test set: the results were worse than baseline in multiple runs. This failure might be attributed to overtraining or bad configuration of the BattRAE toolkit but more investigation would be needed.

2.6 Results

Table 9 summarizes our results. Not all possible setup combinations were performed due to time constraints.

The first two columns (CzMed1 and CzMed2) are based on our smaller Czech medical corpus (around 1M sentences) using two different lists of forbidden words. When we used the method from 2.3 on that corpus using our finetuned thresholds, we got just 71 mistranslated pairs. This corresponds to column CzMed2. We also tried more permissive thresholds, yielding 1226 word pairs with lower quality (column CzMed1). Altogether, neither option makes any significant difference for any of the filtering or scoring methods: all the results differ only slightly and such a small difference can be easily caused just by the randomness of model optimization (MERT).

The remaining four columns of Table 9 cover all HimL languages, trying to apply core fidelity checks for the respective year 2 systems. As before, the results are rather inconclusive, although some improvements over the baseline were obtained (and in the case of German by up to 0.6).

The preliminary experiments with BattRAE failed to reproduce the reported improvements and it is not yet clear if the main reason is the difference of languages, the domain, or possibly some technical error in the application of the provided toolkit.

Conclusion

In Tasks 2.1 and 2.3, we have carried out a number of experiments in search for methods that improve semantic correctness of machine translation in the medical domain. We highlight the more promising results here.

Experiments in Task 2.1 make use of semantic role labels on the source side (English). Our experiments span from traditional statistical approaches (both phrase-based and syntax-based) to the recent neural machine translation. We were able to improve translation quality with semantic role labels for English-to-Czech (neural MT as well as phrase-based MT) with semantic role labels derived from the tectogrammatical layer of annotation. Similar improvements are however achieved with a discriminative model using source morphology only, with no further benefit from semantic roles. For other HimL languages, our experiments relied on PropBank-style labels and the results were inconclusive: the additional linguistic annotation is helpful for some languages and some sub-domains (Cochrane vs. NHS-24).

In Task 2.3 we focused mainly on prevention of negation flip and our initial experiments filtered out phrases based on lists of forbidden word pairs. The forbidden word pairs are extracted automatically from parallel corpora but rely on manual optimization of extraction thresholds. Unfortunately, there is not a clear balance between the precision and recall of forbidden word pair identification to reach an improvement in BLEU after all phrases with forbidden word pairs are removed. We thus experimented with a few additional approaches to utilize the lists of forbidden word pairs and also with a completely different approach to identifying bad phrase pairs based on recursive autoencoders. The results show that the BLEU score improved slightly for Czech and German using our lists of forbidden words and some of the variants of soft filtering. While the BattRAE method seems promising from a theoretical view, we could not reproduce the original results in our setting.

No further work on semantic role labels and core fidelity of MT is planned within the project HimL, but as mentioned, we are in close touch with the research project QT21 which provides one more year for these directions. In case QT21 comes up with methods that significantly improve phrase-based MT, we are ready to incorporate these findings into HimL systems. We are however expecting that more scientific attention will now be given to neural MT. As of now, it is still unclear whether it will be possible to deploy neural MT for HimL production systems. We will discuss this in WP4.

References

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, 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. 2016b. “Findings of the 2016 conference on machine translation.” *Proceedings of the First Conference on Machine Translation*, 131–198. Berlin, Germany.

Clark, Jonathan H, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. “Better hypothesis testing for statistical machine translation: Controlling for optimizer instability.” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 176–181.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. 2006. “Prague Dependency Treebank 2.0.” CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Li, Junhui, Yuval Marton, Philip Resnik, and Hal Daumé III. 2014. “A unified model for soft linguistic reordering constraints in statistical machine translation.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1123–1133.

Li, Junhui, Philip Resnik, and Hal Daumé III. 2013. “Modeling syntactic and semantic structures in hierarchical phrase-based translation.” *Proceedings of NAACL-HLT*, 540–549.

Liu, Ding and Daniel Gildea. 2005. “Syntactic features for evaluation of machine translation.” *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 25–32. Ann Arbor, MI.

Nadejde, Maria, Alexandra Birch, and Philipp Koehn. 2016a. “Modeling selectional preferences of verbs and nouns in string-to-tree machine translation.” *Proceedings of the First Conference on Machine Translation*, 32–42. Berlin, Germany.

Nadejde, Maria, Alexandra Birch, and Philipp Koehn. 2016b. “A neural verb lexicon model with source-side syntactic context for string-to-tree machine translation.” *Proceedings of the International Workshop on Spoken Language Translation (IWSLT).*

Resnik, Philip. 1996. “Selectional constraints: an information-theoretic model and its computational realization.” *Cognition*, 61:127–159.

Sennrich, Rico. 2015. “Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation.” *Transactions of the Association for Computational Linguistics*, 3:169–182.

Sennrich, Rico and Barry Haddow. 2016. “Linguistic input features improve neural machine translation.” *CoRR*, abs/1606.02892.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. “Edinburgh neural machine translation systems for WMT 16.” *CoRR*, abs/1606.02891.

Tamchyna, Aleš and Ondřej Bojar. 2015. “What a Transfer-Based System Brings to the Combination with PBMT.” *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, 11–20. Stroudsburg, PA, USA.

Zhang, Biao, Deyi Xiong, and Jinsong Su. 2016. “Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings.” *CoRR*, abs/1605.07874.