This document is part of the Research and Innovation Action "Quality Translation 21 (QT21)".
This project has received funding from the European Union's Horizon 2020 program for ICT
under grant agreement no. 645452.

Deliverable D2.1

Intermediate Report: Morphologically Rich Languages

Jan Niehues (KIT), Jan-Thorsten Peter (RWTH), Liane Guillou (UEDIN),
Matthias Huck (UEDIN), Rico Sennrich (UEDIN), Stella Frank (UvA),
Ondřej Bojar (CUNI), Tom Kocmi (CUNI), Franck Burlot (LIMSI-CNRS),
Inguna Skadina (TILDE), Daiga Deksne (TILDE)

Dissemination Level: Public

Final, 30th January, 2016

Grant agreement no.	645452
Project acronym	QT21
Project full title	Quality Translation 21
Type of action	Research and Innovation Action
Coordinator	Prof. Josef van Genabith (DFKI)
Start date, duration	1st February, 2015, 36 months
Dissemination level	Public
Contractual date of delivery	31st January, 2016
Actual date of delivery	30th September, 2016
Deliverable number	D2.1
Deliverable title	Intermediate Report: Morphologically Rich Languages
Type	Report
Status and version	Final
Number of pages	24
Contributing partners	KIT, RWTH, UEDIN, UvA, CUNI, LMSI-CNRS, TILDE
WP leader	KIT
Author(s)	Jan Niehues (KIT), Jan-Thorsten Peter (RWTH), Liane Guillou (UEDIN), Matthias Huck (UEDIN), Rico Sennrich (UEDIN), Stella Frank (UvA), Ondřej Bojar (CUNI), Tom Kocmi (CUNI), Franck Burlot (LIMSI-CNRS), Inguna Skadina (TILDE), Daiga Deksne (TILDE)
EC project officer	Susan Fraser
The partners in QT21 are:	▪ Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany ▪ Rheinisch-Westfälische Technische Hochschule Aachen (RWTH), Germany ▪ Universiteit van Amsterdam (UvA), Netherlands ▪ Dublin City University (DCU), Ireland ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ Centre National de la Recherche Scientifique (CNRS), France ▪ Univerzita Karlova v Praze (CUNI), Czech Republic ▪ Fondazione Bruno Kessler (FBK), Italy ▪ University of Sheffield (USFD), United Kingdom ▪ TAUS b.v. (TAUS), Netherlands ▪ text & form GmbH (TAF), Germany ▪ TILDE SIA (TILDE), Latvia ▪ Hong Kong University of Science and Technology (HKUST), Hong Kong

For copies of reports, updates on project activities and other QT21-related information, contact:

Prof. Stephan Busemann, DFKI GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany

stephan.busemann@dfki.de
Phone: +49 (681) 85775 5286
Fax: +49 (681) 85775 5338

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.qt21.eu/`

Contents

1 Executive Summary

This deliverable describes our work on morphologically rich languages within the QT21 project. This work is part of work package 2 (WP2) during the first year of the project.

Several European languages with a rich morphology have been addressed. QT21's focus is on morphologically rich languages, such as German, Czech, and Latvian. Some of the work done for this deliverable began before the start of QT21 and was done on other morphologically rich languages as those chosen for QT21. As the work done on those languages (e.g. Russian, Finnish) is very relevant to QT21 and in order not to stop the analysis done, we decided to finalise this work what will then be applied to the QT21 language pairs during the next reporting period. In order to improve the translation quality for these languages, new ways to represent words in a translation system were evaluated, along with modifications in the modelling of the word agreement.

One area of work is the representation of words in neural network models. While neural network models have shown a great success recently in various areas, the word representation for morphologically rich languages is still a challenge. In Chapter 2, different techniques to improve the word representation in neural network models were investigated.

The second direction of our research work aims at improving the representation of morphologically rich languages within phrase-based statistical machine translation systems. Due to the increased vocabulary size, the translation systems have a higher out-of-vocabulary rate and the model size increases. The work in this area is described in Chapter 3. We investigated the representation of compounds as well as word segmentation. Furthermore, we analysed strategies to model the different morphological forms of a stem in statistical phrase-based systems. In addition, the phrase table size increases dramatically with the vocabulary size in morphologically rich languages. We analysed phrase table filtering techniques to handle this challenge.

Finally, when translating into morphologically rich languages generating translation with correct word agreement is especially difficult. We investigated different techniques to improve the modelling of the word agreement, as described in Chapter 4. Models for phrase-based as well as syntax-based translation systems were evaluated for this purpose. Furthermore, the modelling of pronoun translation was addressed.

2 Word Representation in Neural Networks

Recently, neural network models have shown great success in machine translation. In order to increase their impact on morphologically rich languages we investigated different word representation techniques.

2.1 Neural Network Translation Models

The commonly used one-hot encoding of input words for neural network translation models hides the internal structure of words by reducing it to a single index. Two words that only differ by a letter are treated the same way as two words that have nothing in common. RWTH tested prefix, suffix and class features as additional input features to overcome this limitation (unpublished). The experiments were conducted on the German to English IWSLT 2013 translation task. The basic setup is a phrase-based system similar to the system used in Wuebker et al. (2013). We reranked 500-best lists with a Neural Network Translation Model (NNTM) similar to the Neural Network Joined Model in Devlin et al. (2014) without the language model part and used this as our strong baseline.

We started a first experiment on improving the baseline + NNTM system by adding a prefix feature containing the first three letters (P_3), a suffix feature containing the last two letters (S_2) and the rest (R) which was not covered by the prefix and suffix in addition to the one-hot encoding. This resulted in a TER improvement of 0.3% absolute on the eval set.

In a second experiment we provide the neural network model with more information from the morphologically rich language by using a clustering algorithm to create word classes. These word classes can be used in addition to the word feature to get more reliable inputs since the classes are seen more frequently than the individual words. This yielded an improvement of 0.2 BLEU points absolute compared to the NNTM model and a TER improvement of 0.3% on the eval set (Table 1).

IWSLT	test		eval	
	BLEU[%]	TER[%]	BLEU[%]	TER[%]
Phrase-based	30.7	49.1	36.0	44.0
+ NNTM (Baseline)	31.9	47.4	36.7	43.0
+ P_3-R-S_2	32.1 (+0.2)	**47.2 (-0.2)**	36.8 (+0.1)	**42.7 (-0.3)**
+ Class Features	32.1 (+0.2)	**47.0 (-0.4)**	36.9 (+0.2)	42.7 (-0.3)

Table 1: Experimental results of neural network translation model with word morphology and class features. The German to English IWSLT 2013 translation task was used for these experiments. *test* is the evaluation set from 2010 and *eval* the evaluation set from 2011. A bold font indicates results that are significantly better than the baseline system (Phrase-based + NNTM) with $p < 0.05$.

2.2 Neural Machine Translation of Rare Words with Subword Units

Neural machine translation (NMT) models typically operate with a fixed vocabulary but translation is an open-vocabulary problem. Until now translation of out-of-vocabulary words is performed using back-off models. UEDIN has introduced a simpler and more effective approach (Sennrich et al., 2015), making the NMT model capable of open-vocabulary translation by encoding rare and unknown words as sequences of subword units, based on the intuition that various word classes are translatable via smaller units than words, for instance names (via character copying or transliteration), compounds (via compositional translation), and cognates and loanwords (via phonological and morphological transformations). UEDIN has investigated the suitability of different word segmentation techniques, including simple character n-gram models and a segmentation based on the byte pair encoding compression algorithm. UEDIN has empirically shown that subword models improve over a back-off dictionary baseline for the

system	en-de	en-ru
phrase-based (Haddow et al., 2015)	22.8	24.3
NMT baseline (large-vocabulary model with backoff dictionary)	24.2	22.8
character bigrams	25.3	24.1
byte pair encoding	24.7	24.1

Table 2: Experimental results of neural machine translation with subword units (BLEU scores).

WMT 2015 translation tasks English-to-German and English-to-Russian by up to 1.1 and 1.3 BLEU, respectively (cf. Table 2).

2.3 SubGram

One of the popular vector embeddings for words is word2vec based on the so-called Skip-gram model (Mikolov et al., 2013a). Before applying this model in machine translation, CUNI wanted to avoid one of the clear limitations of the model: treating word forms as atomic units.

CUNI proposed a substring-oriented extension of Skip-gram model called "SubGram" which induces vector embeddings from character-level structure of individual words (submission under review). This approach gives the NN more information about the examined word without any drawbacks in data sparsity or reliance on explicit linguistic markup. In future work, we will compare this totally linguistically uninformed approach with a setup where morphological features are provided to the NN in various ways.

We evaluate the new model on the same set of "semantic" and "syntactic" word tuples created by Mikolov et al. (2013a). Each tuple checks, if the model is able to identify the word related to a given input word in the same way as another pair of words suggests. For example, given the pair "boy–girl" and the query "father", the model is expected to find the word "mother" without any supervision and instead inducing it from observing word co-occurrences in a monolingual corpus.

While the "semantic" relations are idiomatic and not related to the spelling of the words, the "syntactic" part of the test set covers several phenomena clearly described by the processes of word formation and our SubGram can benefit from word-form similarities. This is confirmed by the experiments.

The full details of these experiments are included in Annex A.1.

2.4 Character-based Neural Network Language Models

Language models over characters, rather than words, have the potential to alleviate the problems of large vocabulary/data sparsity that arise when dealing with morphologically rich languages. UvA investigated whether neural languages models over characters implicitly capture morphemes, and how they can then be used to generate new words, i.e. via novel combinations of latent morphemes. Initial experiments were performed on languages with highly-concatenative morphology, in order to evaluate on segmentation points (English, Finnish, and Turkish). Future experiments will include languages with more syncretic morphology (German, Czech).

3 Word Representation in Statistical Machine Translation

When using a phrase-based machine translation system to translate from a morphologically rich language, the different surface forms lead to a higher OOV rate. We therefore investigated techniques that use word segmentation or clustering methods to learn translations for the unseen surface forms. In the reverse direction, the model size is increased due to the high number of surface forms. We used filtering techniques to reduce its size.

3.1 Compounds

At UvA, our research (published as Daiber et al. (2015)) looked into using distributional semantics for derivational morphology. Dealing with productive word formation processes, such as noun compounding, is a key issue for MT. Compound splitting is a common preprocessing step when translating from languages with productive compounding, for example German or Hungarian; commonly used approaches use surface features of the word string, such as component frequency (Koehn and Knight, 2003), or morphological analyzers (Fritzinger and Fraser, 2010). Distributional semantics capture the usage of the compound as well as proposed components and represent them in high dimensional semantic space, such that analogical relations like "handmade is to made as handwriting is to writing" hold. Exploiting these to identify compound splits can be done in a completely unsupervised way (Soricut and Och, 2015) and leads to far more accurate splits than the similarly unsupervised frequency-based methods, as well as improved translation performance between de-en.

3.2 Word Segmentation to Address Sparsity

At the LIMSI-CNRS, a pilot study was run with the intend to improve translation from and into Finnish. The agglutinative character of Finnish morphology makes the task of translating into English very challenging, notably due to the very high number of forms that are rare or even unseen in the training set. These difficulties are common with other morphologically rich languages in which many forms correspond to one single English word. Such forms plague the word alignment and generate a lot of noise during the translation process. A simple workaround is to segment the Finnish text into smaller units that make both languages look more similar. However, finding the right level of segmentation of Finnish words into morpheme-like units so as to produce more robust alignments with English is challenging, and can probably not be performed by looking at only the Finnish side. These early experiments were performed in the context of the WMT 2015 (Marie et al., 2015).

While the experiments carried out for the WMT 2015 evaluation (Finnish:English) did not result in a clear improvement, another approach is being investigated. To address the segmentation in morphological units, a Pitman-Yor Hidden Semi-Markov Model (Uchiumi et al., 2015) is investigated to provide a more efficient model, notably enhanced with an easier control of the segmentation process. While this unsupervised model was, in previous work, proposed in the context of automatic Part-of-Speech tags induction, its generative capabilities can be adapted to the morphological segmentation task. Preliminary experiments on the morpho-challenge for Turkish showed promising results, but some issues need further investigation. For instance, ongoing experiments aim at designing a suitable method for infering the hyper-parameters. We also plan in a near future to evaluate this model in the context of Machine Translation tasks on the QT21 language pairs.

3.3 Source Words Representation in Phrase-based Machine Translation

One of the main problem when translating from a morphologically rich languages is the number of Out-Of-Vocabulary (OOV). Every word stem has many different surface forms and many of them will only be seen rarely. When translating into a less morphologically rich language we often do not require all the morphological information. Therefore, KIT developed and evaluated different techniques to integrate word stem-ming into the translation system. While only relying

on the word stems hurts the performance of the MT system, we could improve the translation of rare words by using a combination of stemmed and surface word representation in a Phrase-based SMT system. By using this methods it is also easy to still integrate any advanced model into the log-linear combination. A detailed description of the work can be found in Annex A.2.

3.4 Phrase Table Filtering for English-Latvian Translation System

The phrase tables built during the training process of a phrase-based statistical machine translation system (PBSMT) are big in size and tend to contain a large portion of low quality data, which influences the quality of translation. For instance, the size of the phrase table of the English-Latvian general domain PBSMT system trained on 4,874,892 unique parallel sentences is 17.5 GB. It contains 143,894,027 lines (bilingual pairs of phrases). Due to the rich morphology of the Latvian language, many phrases in the phrase table have several thousands of translations. In our experiments we made changes to the different modules and outputs of the Moses training pipeline to improve translation into morphologically rich language Latvian:

- we added an extra function to the phrase table creation module to filter out some entries of the phrase table;
- we modified the phrase extraction module;
- we modified the symmetrized alignment file;
- we modified the alignment files prior to symmetrization.

With the phrase table filtering, we tried to overcome the consequences of both an imperfect alignment as well as deficiencies in the phrase pair extraction algorithm. This approach did not produce the hoped-for results – the BLEU scores of the experimental systems were close but did not reach the baseline (0.42–0.73 BLEU points below the baseline), although there is also a positive side – the size of the phrase table could be reduced by 18% without significant loss of translation quality.

By modifying the symmetrized word alignment file, we were trying to prevent extraction of bad phrase pairs. The results are promising: BLEU scores in most experiments exceeded the baseline (by 0.48 BLEU points maximum).

By modifying the extraction module's algorithm, we gained in flexibility. We extracted only the good fragments from the sentence pairs with a low correlation of word order. We reduced the overgeneration of the phrase pairs containing unaligned words. We ignored pairs where a single word phrase is translated as a single character. We accepted only a single, predefined translation for some functional words. The best result is achieved by simple modification – ignoring single word phrases translated as single characters or vice versa. The baseline is exceeded by 0.27 BLEU points. All experiments together yield a result exceeding baseline by 0.15 BLEU points for the test corpus.

In all experiments the source to target and target to source alignment BLEU score exceeded the baseline (by 0.02–0.46 BLEU points for the test corpus). Every separate experiment addresses a different issue of the alignment – common multiword phrase alignment, the negative forms of the verbs, comma alignment mismatch, alignment of the clitics, etc.. We had high expectations for the cumulative experiment, which tries to solve all these issues, unfortunately the BLEU score for the cumulative experiment did not exceed the best single-issue experiment, although it still exceeds the baseline by 0.41 BLEU points.

We can draw several conclusions from these experiments. Alignment file modification yields better results than direct phrase table filtering or changes in the algorithm of the phrase pair extraction. Alignment points should be fixed prior to phrase extraction. Although experiments described in this deliverable were performed for English-Latvian pair, we believe that they are also applicable to other morphologically rich under resourced languages. More details can be found in Annex A.3

4 Modelling Morphological Agreement

Morphologically rich languages have an additional challenge of their complex word agreement, in addition to the higher vocabulary size. We therefore investigated different techniques which are devised to improve the agreement between the words on the target side.

4.1 Pronoun Translation

Pronouns and mophology of morphologically rich languages interact each other closely, since the choice of an anaphoric pronoun needs to satisfy agreement with the morphological realisation of the noun to which it refers (its *antecedent*) and the verb for which it is an argument. Other agreement requirements may also exist for other pronoun functions. Such agreement requirements pose a problem for SMT, not only within individual sentences, but also at the discourse level.

Research in discourse-aware SMT at UEDIN includes work on the specific problem of pronoun translation. Recent work includes the development of an automated post-editing system submitted to the shared task on pronoun translation at the 2nd Workshop on Discourse in Machine Translation (Guillou, 2015b). The system categorises pronouns according to their function and uses simple, hand-crafted rules to detect and amend pronoun translation in SMT output. With under-performance observed for all shared task systems, pronoun translation remains yet an unsolved task. In order to better understand the problem, UEDIN conducted analyses of manual and automated translation (Guillou, 2015a). We used the ParCor corpus (Guillou et al., 2014) of parallel English-German texts and their annotations to categorise pronouns according to their function. We identified significant differences in pronoun usage between English and German for the anaphoric type in manual translation. In addition to this, we made recommendations as to further sub-categorisation of this function with respect to system development and evaluation.

4.2 A Dependency Model of Morphological Structure

When translating between two languages that differ in their degree of morphological synthesis, syntactic structures in one language may be realized as morphological structures in the other, In this case, SMT models require a mechanism to learn such translations. Morpheme segmentation is commonly employed to allow the translation of morphologically complex words. In a flat representation, i.e. as a sequence of morphemes, however, the hierarchical structure of words is lost, which is especially damaging when translating into a morphologically complex language. For models operating on sequences of surface tokens, such as n-gram language models, splitting morphemes increases the distance between tokens which are inter-dependent. For example, this will be the case for the morphosyntactic agreement between a determiner and the head of a German compound, which is the last segment at the same time. We wish to utilize the morphological structure to better model morphosyntactic constraints and selectional preferences. In Sennrich and Haddow (2015), UEDIN has presented a quasi-syntactic dependency annotation of selected morphological structures in German, which allows for a translation model that models the translation of both syntactic and morphological structures. It informs the dependency language model described previously in Sennrich (2015) about the relationship between morphemes, and facilitates the modelling of agreement and selectional preferences involving morphologically complex words. Improvements in translation quality of up to 1.8 BLEU are achieved on the WMT English-to-German translation task.

4.3 Source Discriminative Word Lexicon

The correct word form in a morphologically rich language often has long-range dependencies. In order to better disambiguate the translations, KIT propose modeling the translation as a prediction task. The source discriminative word lexicon was developed to predict the correct

translation given the source word or contextual information. For every source word, a maximum entropy classifier was trained using local context features and dependency information. We investigated different features as well as different word representation. Using this additional information, we could improve the prediction of the correct morphological form of words in the target language. When adding the model to a phrase-based machine translation, the performance of the English to German translation task of IWSLT was improved by up to 0.6 BLEU points (Ha et al., 2015). A detailed description of the technique can be found in Annex A.4.

4.4 Translating into a Synthetic Language

The LIMSI-CNRS has been addressing the problem of translating from a language with analytic tendencies, such as English, into a rather synthetic language, such as Russian or Czech. The lack of parallelism between these two types of language leads to a difficulty to predict the correct case when translating from English to Russian. To address this problem, we developed and evaluated a two-step translation scenario. After normalizing the Russian part of the data by removing case markers, we trained a system that proceeds as follows: 1) translation from English to normalized Russian using a conventional phrase-based approach; 2) prediction of the case, using a cascade of Conditional Random Fields, and generation of the correct word form. This system was evaluated as part of LIMSI's contribution to the WMT 2015 Machine Translation task (Marie et al., 2015) and did not improve the translation over a direct translation baseline, supposedly because no information was taken from the source sentences during the second step.

When aligning a synthetic language with an analytical one many target (function) words remain unaligned. This is an issue during the phrase extraction phase, as null-aligned words tend to generate many spurious phrases. We have proposed a factored alignment model designed to complete word-to-word alignments by adding morpheme-to-word links. Technically, this model is a simple extension of IBM model 2, which can accomodate bundle of morphological features, rather than words, on the target side. Using this model, we were able to greatly reduce the number of non-aligned words on the target side, yielding more compact and less noisy translation models, with however hardly any impact so far on the translation quality as measured by automatic metrics such as BLEU. Since the last report about this work in Burlot and Yvon (2015), we have obtained results for other languages as described in A.5 and used dependency trees, although this brought little improvement.

4.5 Interaction Between Morphology and Word Order

Modelling morphological agreement correctly often requires knowledge about the structure of the entire sentence, which can also be used to improve word order in translation. UvA has been investigating the interaction between morphology and word order for English-German translation. New models have been developed aiming at exploiting this interaction for improved performance and experiments exploring the potential advantages are underway.

For the WMT 2016 shared task, the LIMSI and the UvA-ILLC teams are collaborating on system submissions for English-German, English-Czech, as well as the other translation directions. In this task, we aim at exploiting improved word-order and morphological markings building on the LIMSI system. We will investiage the preordering approach described in Stanojević and Sima'an (2015) and ongoing work on the joint prediction of morphology and syntax by UvA-ILLC and LIMSI. Efforts to improve these systems still continue and the results are expected by WMT 2016 submission deadline.

Reference

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Franck Burlot and François Yvon. 2015. Morphology-aware alignments for translation to and from a synthetic language. In *International Workshop on Spoken Language Translation (IWSLT 2015)*, pages 188–195, Da Nang, Vietnam, December.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University, Stanford, CA, USA.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland, June.

Fabienne Fritzinger and Alexander Fraser. 2010. How to avoid burning ducks: Combining linguistic analysis and corpus statistics for german compound processing. In *Proceedings of the ACL 2010 Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pages 224–234, Uppsala, Sweden, July. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Liane Guillou. 2015a. Analysing ParCor and its Translations by State-of-the-art SMT Systems. In *Proceedings of the Workshop on Discourse in Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Liane Guillou. 2015b. Automatic Post-Editing for the DiscoMT Pronoun Translation Task. In *Proceedings of the Workshop on Discourse in Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, Eunah Cho, Mohammed Mediani, and A lex Waibel. 2015. The KIT Translation Systems for IWSLT 2015. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam, 3-4 december.

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *EMNLP 2015 Tenth Workshop on Statistical Machine Translation (WMT 2015)*, pages 126–133, Lisbon, Portugal, September. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Proceedings of NIPS 2002*, Vancouver, Canada.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL 2003*, Sapporo, Japan.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL 2003, 10th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17, 2003, Agro Hotel, Budapest, Hungary*, pages 187–194.

Benjamin Marie, Alexandre Allauzen, Franck Burlot, Quoc-Khanh Do, Julia Ive, elena knyazeva, Matthieu Labeau, Thomas Lavergne, Kevin Löser, Nicolas Pécheux, and François Yvon. 2015. LIMSI@WMT'15 : Translation task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 145–151, Lisbon, Portugal, September. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Jan Niehues and Alex Waibel. 2013. An MT Error-driven Discriminative Word Lexicon using Sentence Structure Features. In *Proceedings of WMT 2013*, Sofia, Bulgaria.

J. Niehues, T. Herrmann, S. Vogel, and A. Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK.

Rico Sennrich and Barry Haddow. 2015. A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2081–2087, Lisbon, Portugal, September. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *CoRR*, abs/1508.07909.

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

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1627–1637.

Miloš Stanojević and Khalil Sima'an. 2015. Reordering grammar induction. In *PProceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 44–54. The Association for Computational Linguistics.

Kei Uchiumi, Hiroshi Tsukahara, and Daichi Mochihashi. 2015. Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1774–1782, Beijing, China, July. Association for Computational Linguistics.

Ekaterina Vylomova, Laura Rimmel, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *arXiv preprint arXiv:1509.01692*.

Joern Wuebker, Stephan Peitz, Tamer Alkhouli, Jan-Thorsten Peter, Minwei Feng, Markus Freitag, and Hermann Ney. 2013. The rwth aachen machine translation systems for iwslt 2013. In *International Workshop on Spoken Language Translation*, pages 88–93, Heidelberg, Germany, December.

A Annex

A.1 SubGram Details

This section provides the details of our adaptation of the Skip-gram model for obtaining word embeddings.

A.1.1 Generation of Substrings

We append the characters ^ and $ to the word to indicate its beginning and end. In order to generate the vector of substrings, we take all character bigrams, trigrams etc. up to the length of the word. This way, even the word itself is represented as one of the substrings. For the NN, each input word is then represented as a binary vector indicating which substrings appear in the word.

A.1.2 Adjusting the Skip-gram

The original Skip-gram model by Mikolov et al. (2013a) uses one-hot representation of a word in vocabulary as the input vector. This representation makes training fast because no summation or normalization is needed. The weights w_i of the input word i can be directly used as the output of hidden layer h (and as the distributed word representation):

$$h_j = w_i j \tag{1}$$

In our approach, we provide the network with a binary vector representing all substrings of the word. In the classic NN, one would have to compute:

$$h_j = \sigma(\sum_{i=1}^{|X|} x_i * w_{ij}) \tag{2}$$

The sigmoid function would introduce a non-linearity in the model and slow down the training process. More importantly, we would lose the nice property of the Skip-gram model, namely the suitability for linear vector operations such as $vec(\text{king}) - vec(\text{man}) + vec(\text{woman}) \approx vec(\text{woman})$. To solve the we replaced the sigmoid function with the mean value:

$$h_j = \frac{\sum_{i=1}^{|X|} x_i * w_{ij}}{|S|} \tag{3}$$

where $|S|$ is the number of substrings of the word x.

A.1.3 Evaluation and Data Sets

We train our NN on words and their contexts extracted from the English side of the parallel corpus CzEng 1.0 (Bojar et al., 2012). The training data consist of 217M running words with the vocabulary size of 1072555 distinct word forms. We consider only the 180239 most frequent word forms to simplify the training of the NN. The remaining word forms fall out of vocabulary (OOV) of the NN, so the original Skip-gram cannot provide them with any vector representation. Our SubGram relies on known substrings and always provides at least some approximation.

We test our model on the exact test set by Mikolov et al. (2013a). The test set consists of 19544 "questions", of which 8869 are called "semantic" and 10675 are called "syntactic" and further divided into 14 types, see Table 3. Each question contains two pairs of words (x_1, x_2, y_1, y_2) and captures relations like "What to 'woman' (y_1) is like 'king' (x_2) to 'man' (x_1)?", together with the expected answer 'queen' (y_2). The model is evaluated by finding the word whose representation is the nearest (cosine similarity) to the vector $vec(\text{king}) - vec(\text{man}) + vec(\text{woman})$. If the nearest neighbour is $vec(\text{queen})$, we consider the question answered correctly.

In this work, we use Mikolov's test set because it is used in many subsequent papers, but after a closer examination we came to the conclusion, that it does not test what the broad

Question Type	Sample Pair
capital-countries	Athens – Greece
capital-world	Abuja – Nigeria
currency	Algeria – dinar
city-in-state	Houston – Texas
family	boy – girl
adjective-to-adverb	calm – calmly
opposite	aware – unaware
comparative	bad – worse
superlative	bad – worst
present-participle	code – coding
nationality-adjective	Albania – Albanian
past-tense	dancing – danced
plural	banana – bananas
plural-verbs	decrease – decreases

Table 3: Mikolov's test set question types, the upper part are "semantic" questions, the lower part are "syntactic".

terms "syntactic" and "semantic relations" suggest. "Semantics" is covered by questions of only 3 types: guess a city based on a country or state, predict currency name from the country and predict the feminine variant of nouns denoting family relations. As Vylomova et al. (2015) also points out, many other "semantic" relationships could be tested, e.g. walk-run, dog-puppy, bark-dog, rain-cloud, cook-eat and others.

"Syntactic" questions cover a wider range of relations at the boundary of morphology and syntax. The problem is that all questions of a given type are constructed from just a few dozens of word pairs, comparing pairs with each other. Overall, there are only 313 distinct pairs throughout the whole syntactic test set of 10675 questions, which means only around 35 different pairs per question set. Moreover, of the 313 pairs, 286 pairs are regularly formed (e.g. by adding the suffix 'ly' to change an adjective into the corresponding adverb). We find such a test set extremely small and unreliable to answer the question whether the embedding captures semantic and syntactic properties of words. Therefore, we plan to work on a larger and more robust test set.

In order to test our extension of the Skip-gram on out-of-vocabulary words, we extended the original question sets with new questions where at least one of x_1, x_2 and y_1 is not among the 180k known word forms. Since we are interested in morphosyntactic relations, we extended only the questions of the "syntactic" type with exception of nationality adjectives.

We constructed the pairs more or less manually, taking inspiration in the Czech side of the CzEng corpus where explicit morphological annotation allows to identify various pairs of Czech words (different grades of adjectives, words and their negations, etc.). The word-aligned English words often shared the same properties. For verb tense, we relied on a freely available list of English verbs in their morphological variations. The questions were constructed from the pairs similarly as by Mikolov: generating all possible pairs of pairs. When needed, we downsampled the generated sets of questions to 1000 instances per set.

A.1.4 Experiments and Results

We used a Python implementation of Mikolov's word2vec[1] as the basis for our SubGram.[2]

[1] http://radimrehurek.com/gensim

[2] Gensim implements the model twice, in Python and an optimized version in C. For our prototype, we opted to modify the Python version, which unfortunately resulted into a code about 100 times slower and forced us to train the model only on the 0.2 gigaword-magnitude corpus as opposed to Mikolov's $6 * 10^9$ word2vec training data.

	Skip-gram	SubGram	
Type	**In Vocabulary**	**OOV**	
capital-countries	3.2%	0%	-
capital-world	1.2%	0%	-
currency	0.4%	0%	-
city-in-state	0.5%	0%	-
family	49.4%	4.9%	-
Overall semantic	4.16%	0.3%	-
adjective-to-adverb	6.1%	60.0%	9.4%
opposite	10.1%	66.1%	19.3%
comparative	53.8%	47.3%	0.1%
superlative	22.9%	47.4%	0.1%
present-participle	22.1%	32.1%	5.1%
nationality-adjective	1.4%	9.6%	-
past-tense	18.7%	12.1%	7.9%
plural	24.9%	37.8%	1.7%
plural-verbs	24.3%	63.1 %	1.2%
Overall semantic	20.7%	37.7%	5.6%

Table 4: Results on Mikolov's test set questions ("In Vocabulary") and our additional questions ("OOV").

We limit the vocabulary, requiring each word form to appear at least 10 times in the corpus and each substring to appear at least 500 times in the corpus. This way we get the 180239 unique words and 268090 (+180239 words, as we downsample words separately) unique substrings.

Our word vectors have the size of 100. The size of the context window is 5.

In contrast to Mikolov, we do not compare answers against only the 30000 most common words and use the whole vocabulary instead. This choice leads to worse results (it is harder to pick the correct neighbour from a denser set of possible options), but we consider this evaluation more useful.

Table 4 reports the results. The first two columns are based on Mikolov's original test set where questions with OOV words are removed. The set of questions is identical for both models here so they can be directly compared. The last column contains questions from our extension of the dataset.

Comparing the first two columns, we see that our SubGram outperforms the Skip-gram in nearly every morphosyntactic question. On the other hand, it does not properly capture the tested semantic relations. One possible explanation is that our training corpus is smaller and does not contain as many named entities. The same limited coverage could also be the reason behind the very low performance of both models on the nationality-adjective question.

The last column suggests that the performance of our model on OOV words is not very high, but it is still an improvement over flat zero of the original Skip-gram model. The performance on OOV words is expected to be lower, since the model has no knowledge of the exceptions and can only benefit from regularities in substrings.

A.2 Source Words Representation in Phrase-based Machine Translation

When we translate from a highly inflected language into a less morphologically rich language, not all syntactic information encoded in the surface forms may be needed to produce an accurate translation. For example, verbs in French must agree with the noun in case and gender. When we translate these verbs into English, both the case and gender information may be safely discarded. In order to address this sparsity problem, KIT tried to cluster words that have the same translation probability distribution, leading to higher occurrence counts and therefore more reliable translation statistics. After clustering the words into groups that can be translated in the same or at least in a similar way, there are different possibilities to use them in the translation system. A naive strategy is to replace each word by its cluster representative,

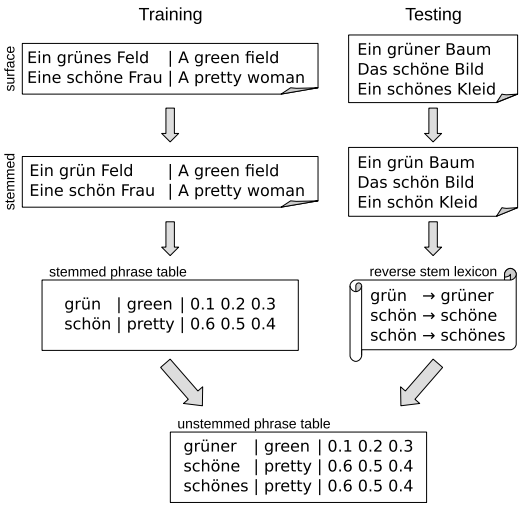


Figure 1: Workflow for unstemming the PT.

called *hard decision stemming*. However, this carries the risk of discarding vital information. Therefore techniques to integrate both, the surface forms as well as the word stems, into the translation system were investigated. In the *combined input*, the stemmed adjectives were added as translation alternatives to the preordering lattices. Since this poses problems for the application of more advanced translation models during decoding, a novel *hidden combination* technique was proposed.

A.2.1 Combined Input

Mistakes made during hard decision stemming cannot be recovered. Soft integration techniques avoid this pitfall by deferring the decision on whether to use the stem or surface form of a word until decoding. The system is able to choose by combining both the surface form based (default) phrase table and the word stem based (stemmed) phrase table log-linearly. The weights of the phrase scores are then learned during optimization. In order to be able to apply both phrase tables at the same time, the input of the decoder needs to be modified. Our baseline system already uses preordering lattices, which encode different reordering possibilities of the source sentence. Every edge in the lattice containing an adjective is replaced by two edges: one containing the surface form and the other the word stem. This allows the decoder to choose which word form to use depending on the word and its context.

A.2.2 Hidden Combination

While it is possible to modify our phrase table to use both surface forms and stems in the last strategy, other models in our log-linear system suffer from the different types of source input. For example, the bilingual language model (Niehues et al., 2011) is based on tokens of target words and their aligned source words. In training, either the stemmed corpus or the original one can be used, but during decoding a mixture of stems and surface forms occurs. For the unknown word forms the scores will not be accurate and the performance of our model will suffer. Similar problems occur when using other translation models such as neural network based translation models.

Therefore a novel strategy to integrate the word stems into the translation system was developed. Instead of stemming the input to fit the stemmed phrase table, the stemmed phrase table is modified so that it can be applied to the surface forms. The workflow is illustrated in Figure 1. All the stem mappings are extracted from the development and test data and compiled a stem lexicon. This maps the surface forms observed in the dev and test data to their corresponding stems. Then this lexicon is applied in reverse to our stemmed phrase table, in effect duplicating every entry containing a stemmed adjective with the inflected form replacing

System	Dev	Test	
Baseline	28.91	30.25	
Hard Decision	29.01	30.30	(+0.05)
Combined Input	29.13	30.47	(+0.22)
Hidden Combination	29.25	**30.62**	(+0.37)

Table 5: TED low-resource small systems results.

System	Dev	Test	
Baseline	29.73	31.33	
Hard Decision	29.74	30.84	(-0.49)
Combined Input	29.97	31.22	(-0.11)
Hidden Combination	29.87	31.61	(+0.28)

Table 6: TED extended features systems results.

the stem. Afterwards this "unstemmed" phrase table is log-linearly combined with the default phrase table and used for translation.

This allows us to retain our generalization won by using word clusters to estimate phrase probabilities, and still use all models trained on the surface forms. Using the hidden combination strategy, stemming can easily be implemented into current state-of-the-art SMT systems without the need to change any of the advanced models beyond the phrase table. This makes our approach highly versatile and easy to implement for any number of system architectures and languages.

A.2.3 Evaluation

The problem of a high number of OOV-words is especially problematic, when translating from a morphologically rich language on a task with only limited training data. We tested this approach on the task of translating German TED lectures into English. The systems were only trained on the TED corpus.

The results for the systems built only on the TED corpus are summarized in Table 5 for the small system and Table 6 for the extended system. A bold font indicates results that are significantly better than the baseline system with $p < 0.05$. The baseline systems reach a BLEU score on the test set of 30.25 and 31.33 respectively.

The small system could be improve slightly to 30.30 using only stemmed adjectives. Adding the stemmed forms as alternatives to the preordering lattice leads to an improvement of 0.2 BLEU points over the small baseline system. This strategy does not tap the full potential of our extended system, as there is still a mismatch between the combined input and the training data of the advanced models.

The hidden combination strategy rectifies this problem, which is reflected in the results. Using the hidden combination the best BLEU score could be achieved for both systems. We could improve it by almost 0.4 BLEU points over the small baseline system and 0.3 BLEU points on the system using extended features.

A.3 Phrase Table Filtering for English-Latvian Translation System

The phrase tables built during the training process of a phrase-based statistical machine translation system (PBSMT) are big and tend to contain a large portion of low quality data, which influences the quality of translation. In this section we explore the possibilities of how to filter out the noisy data during the training process. The aim of this study is to improve the quality of an MT system's output for the morphologically rich Latvian language. We explored four approaches for phrase table filtering - we have filtered the phrase table directly, we have modified a symmetrized word alignment file that is then used for phrase pair extraction and creation of a

phrase table, we have changed the algorithm of the phrase pair extraction module, and we have edited the source to target and target to source alignment files prior to the symmetrization.

A.3.1 Filtering a Phrase Table

By examining the phrase table we noticed several phrase patterns which seems to be incorrect or could unreasonably enlarge the size of the phrase table:

- Every token of the source phrase is a single character, e.g., _ *a* _ *a* _ *a*;
- The source phrase starts with a partial word (clitic), e.g., *' ve* or *s the*;
- The source phrase starts but does not end with a double quote, e.g., *" and*;
- The source phrase ends but does not start with an apostrophe, e.g., *we '*;
- A single symbol is translated as a word or vice versa;
- The number of tokens in a source phrase is much bigger than in a target phrase or vice versa;
- The probability of translation is very small;
- The lexical probability is very small;
- The probability of direct translation is much smaller that the probability of inverse translation or vice versa.

Based on these observations we defined different rules for discarding a phrase pair from the phrase table. The BLEU score of the baseline system is 28.14, the size of the phrase table is 17.5 GB. Our experimental systems show BLEU sores ranging from 27.41 to 27.72 and the size of the phrase tables varies from 14.4 GB to 16.7 GB. We can conclude that the filtering of the phrase table allows decreasing the size of the phrase table notably, in the same time it does not affect the translation quality significantly.

A.3.2 Modifying a Symmetrized Word Alignment File

The experiments with the phrase table allowed us to decrease the size of the phrase table, but did not result in increase of BLEU score, further experiments were performed on a symmetrized word alignment file.

1. Although word order in Latvian is rather free, in general, word order of Latvian and English sentences is similar – subject-verb-object (SVO). This allows us to assume that there is a strong correlation between the word position in source and target sentences. Although there are some differences in word ordering, for example, genitive noun phrase of type 'noun1 of noun2' in English is translated as a 'noun2(in genitive) noun1' in Latvian, the changes in word order are not big. If a sentence contains several instances of the same functional word or punctuation mark, sometimes it is aligned to the wrong instance of this word on a target side, causing bad alignment of the whole sentence. We calculate the Pearson correlation coefficient for the alignment points of every sentence pair. By exploring the values of the correlation coefficient, we can detect the sentences with potentially bad alignment. Such sentences are not used for the phrase extraction.

2. Every five consecutive alignment points we observe if the first two and the last two are in linear order and if the middle one differs by more than four positions, we remove it.

3. We do not extract the phrases from sentences with five or more consecutive unaligned words in order to avoid overgeneration of phrases that differ only by a single word on one side and are the same on the other side.

4. Usually sentences end with a punctuation mark that is correctly aligned. However, if one or several words before the last token are not aligned, there is an overproduction of incorrect phrase pairs, such as one phrase as a single punctuation symbol with a corresponding

System	BLEU	Size of the phrase table (GB)
Baseline	28.14	17.5
Experiment 1	28.32	17.2
Experiments 1 and 2	28.22	17.3
Experiments 1 and 3	28.23	**15.7**
Experiments 1 and 4	27.81	17.6
Experiments 1 and 5	**28.62**	**15.7**

Table 7: BLEU score and phrase table size of the baseline system and systems with the modified word alignment file (the best BLEU score is in bold).

phrase as a punctuation symbol and some unaligned words. We remove the last alignment point if the previous source or target word is not aligned.

5. We add out-of-range alignment points for unaligned words that are between five or more consecutive unaligned words. Those words will be ignored during the phrase extraction step. Several unaligned points in a row may contain information that is not in the other language, and translations produced by a system trained on such phrases may be inadequate.

Table 7 shows evaluation results for systems trained using the modified alignment file.

A.3.3 Modifying a Phrase Extraction Module

Some issues could not be addressed just by changing the symmetrized word alignment file. Therefore, we made the following changes in the extraction module and evaluated their impact on MT output:

1. Sentences with low correlated alignment points were not discarded completely. The fragments with a good correlation were kept (three or more tokens in a linear order in the source and in the target phrase).
2. Ignore a single word phrases translated as a single character or vice versa (except the single symbol token *I*).
3. If a source token is unaligned but one token before it is aligned with two adjacent target tokens then move the alignment point to make a linear alignment (the first source token to the first target token, the second source token to the second target token).
4. Ignore phrases with five or more unaligned tokens.
5. If a single word phrase is preposition, article, conjunction, adverb, pronoun or auxiliary verb (*a*, *an*, *are*, *in*, *of*, *the*, *to*, *at*, *away*, *back*, *before*, *below*, *but*, *by*, *for*, *he*, *his*, *if*, *now*, *on*, *or*, *she*, *then*, *their*, *then*, *therefore*, *under*, *very*, *when*, *with*) then only a single predefined translation is accepted.

Table 8 shows evaluation results for systems where different extraction modules where used in training.

A.3.4 Modifying Alignment Files prior to Symmetrization

By exploring the source to target and target to source alignment files, we have noticed some alignment issues which we try to solve before symmetrization:

- The clitics *'s*, *'t*, *'re*, *'m* do not have adjacent alignment points with the words/tokens to which they are connected. As a result, they are not extracted together with a related word/token. By adding missing points we unite *'s* and *'t* with a previous word, but *'re* and *'m* is united with a next word, if it is a present participle.

System	BLEU (test corpus)	BLEU (dev. corpus)
Baseline	28.09	22.63
Discarded sentences with a low correlation	27.92	**22.86**
Experiment 1	27.84	**22.90**
Experiment 2	**28.36**	21.30
Experiment 3	27.94	22.50
Experiment 4	27.57	**23.11**
Experiment 5	27.97	**22.70**
Experiments 1 to 5 together	**28.24**	22.54

Table 8: BLEU scores for the systems where different extraction modules are used (the scores exceeding the baseline are in bold).

System/issue	BLEU (test corpus)	BLEU (dev. corpus)
Baseline	28.09	22.63
1. Clitics	**28.52**	**22.75**
2. Common phrases	**28.36**	**23.02**
3. Negative expressions	**28.55**	**22.75**
4. Several commas	**28.47**	22.33
5. Several quotes	**28.20**	22.42
6. The lists of the words	**28.47**	**23.02**
7. 'x of y' type phrases	**28.28**	**22.64**
8. Functional/content word alignment	**28.11**	**22.73**
9. Cumulative system	**28.50**	**22.86**

Table 9: BLEU scores for the systems with fixed alignment issues (the scores exceeding the baseline are in bold).

- Sometimes for common phrases, the alignment results is a two-word source and target phrase pair, while separate words are not included in alignment. We change the alignment points in a way that both – the two-word phrase pair and each word separately – are extracted.
- We review and fix alignment for the phrases expressing negation.
- If sentences have several commas, then a single comma on a source side could be aligned to the several commas on a target side or vice versa. We remove the extra alignment points to solve this problem.
- If sentences contain several quotes, then there could be a similar problem as with a several comma alignment. We remove the extra alignment points to solve this problem.
- The lists of the words can be separated by the conjunctions or commas. If there are several such item separators there could be problems with an alignment. We make the list alignment consistent by shifting points or removing extra points.
- In English genitive noun phrases with a structure 'x of y' are very common. These correspond to the Latvian noun phrase with the structure 'y (in genitive) x'. We change alignment to tie together all the words of such phrases.
- A functional word could have a wrong alignment to a content word. We have defined the lists of functional words for the source and for the target language. If the functional word is aligned to non-functional word and it does not have an adjacent both-sided alignment point, the alignment for this functional word is removed.

Table 9 shows evaluation results for the systems where different alignment issues were addressed.

A.4 Source Discriminative Word Lexicon

The correct word form in a morphologically-rich language often has long-range dependencies. In order to better disambiguate the translations, KIT propose modeling the translation as a prediction task. The prediction is motivated by the discriminative word lexicon (DWL) (Niehues and Waibel, 2013). While the DWL operates on the target side and learns to predict for each target word whether it should occur in a given target sentence, the source discriminative word lexicon (SDWL) operates on the source side. For every source word a classifier is trained to predict its translation in the given sentence. A multi-class classification task is performed by identifying for every source word the 20 most frequent translations, which are provided by the word alignment generated using GIZA++. All words in the target language that occur less frequently than the 20 most frequent words are assigned to one class, called **other**. Alignments to the NULL word on the target side are treated in the same way as if NULL were a word. The source vocabulary is limited to the words occurring in the test data and train up to 20 classifiers for each source word. In reality, most words have much fewer alternative translation options than 20. The SDWL uses binary maximum entropy classifiers trained using the one-against-all scheme. That means a maximum entropy model is used to estimate $p(e|f, c(f))$, where e is the target word that should be predicted given source word f and its context/dependency features $c(f)$. During training, the maximum entropy models for the individual classes for each source word are trained based on the given set of features extracted from the source sentence and the correct class of each training example. For the prediction, the test data is first separated into words. For each word the features are extracted from the source sentence it stems from. Then all binary maximum entropy models for the multiple classes are applied and each of them produces a prediction. The final prediction corresponds to the class with the highest prediction probability.

A.4.1 Structural Features

The training examples and test data for the classifiers are represented by a set of features and the class each example belongs to. Different types of features representing the structure of a sentence to varying degrees were evaluated.

Bag-of-Words A straight forward way to represent the source sentence for this classification task is to use the bag-of-words approach. This is the least structural informative feature which does not provide any knowledge about the sentence beyond the mere existence of the words in it.

Context The context feature adds structural information about the local context of the modeled source word in the sentence. In addition to the context words themselves, their position is encoded in the feature such that the same word occurring at a different position (relative to the source word in question) would result in a different feature. Up to six context words were included, three on each side of the source word. Hence, this feature type provides structural information by means of sequential order within a limited context.

Dependency Relations The feature contributing the most information about the sentence structure is based on the relations between the source sentence words in a dependency tree. In order to obtain the dependency relations, a dependency tree is extracted from a constituency parse tree using the Stanford Parser (Klein and Manning, 2002; Klein and Manning, 2003). Then the dependency relations between the source word and its parent and children in the dependency tree are included as features. That means, a feature consisting of the governance relation (parent or child of the source word), the dependency relation type (from the set of dependency relations described in (de Marneffe and Manning, 2008) e.g., `nsubj`, `dobj`, `vmod`, etc.) and the connected word itself is created. This type of feature allows to capture structure by means of semantic dependencies that can range over longer distances in the sentence, but

Sentence:		*Well it obviously is not.*
Features:	bag-of-words	not is it obviously well .
	context	−1_well +1_obviously +2_is
	dependency	dep_parent_nsubj_is

Example A.1: Representation of the source word *it* by the different features.

are relevant due to the semantic connection to the current source word. An example for the features for the word *it* in a given sentence is presented in Example A.4.1.

A.4.2 Word Representation

In this work, two methods to represent the words in the features were compared: word IDs and word vectors.

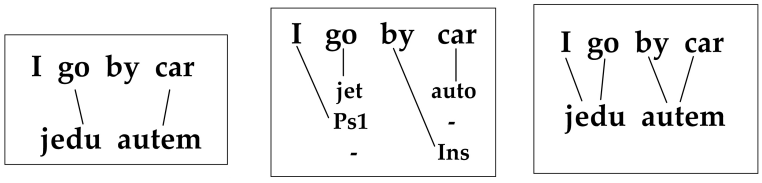
Word IDs When representing words by word IDs, the source vocabulary size V_{source} is used as the dimension of the feature space, a word's ID in the vocabulary as a feature. The feature is set to 1 if it is used in the example. All other features are set to 0. For accommodating the context features (**context**), the size of the features space is extended such that $V_{context} = c{*}V_{source}$ where c equals the size of the context. Each position of a word in the context hence has its own range in the feature space, and words in different context positions can be distinguished accordingly. The features representing dependency relations (**dep**) are included in a similar fashion. Again, a new feature space is defined as $V_{dep} = d * V_{source}$ where d equals the amount of all dependency relations, where parent and child relations are counted separately. The feature types can be combined by simply concatenating the individual feature spaces. That means when all three types of features are used the size of the feature space amounts to $V_{source} + V_{context} + V_{dep}$. It is obvious that with this strategy for design the feature space grows relatively big, possibly leading to data sparseness problems. In order to reduce dimensions, the representation via word vectors is used as a more appropriate measure.

Word Vectors The word vectors for word representation are generated using word2vec (Mikolov et al., 2013b) with the number of dimensions set to 100. That means each word is represented by a 100-dimensional vector. However, it is not straight forward how multiple words should be expressed in this representation, so that the representation by word vectors is not applied for the bag-of-words features, but only for the context and dependency features. In case of the vector representation of the context features (**contextVec**), each position in the context words receives its own range in the feature space. Hence, the size of the feature space equals to $V_{contextVec} = c * dim$, where c is the context size and dim the dimension of the vector representation. This reduces the dimension significantly compared to $V_{context}$ used in the word ID-based representation. The feature space for dependency relations using word vectors (**depVec**) equals to $V_{depVec} = d * dim$ with d being the inventory of dependency relations. Compared to V_{dep}, again a huge reduction can be achieved. In addition to the **depVec** feature, further variants of the dependency feature are compared as followings.

parentDepVec
: For this feature, only the dependency relation to the parent word is represented in vector representation.

parentWordVec
: This feature consists of the vector representation of the parent word and an additional binary feature that is 1 if the parent word is the root of the dependency tree.

parentWordVec+DepRel
: In addition to the **parentWordVec** feature, the dependency relation to the parent word is encoded as a vector.



(a) (b) (c)

Figure 2: Morphological alignments. (a) Lexical alignments missing the English pronoun and preposition that are encoded in the Czech endings. (b) The source 1st person tag is aligned to the target pronoun *I* and the instrumental case tag to the preposition *by*. (c) Lemma and tag alignments are merged to provide links between word forms.

As for the word-based features, word vector features can be combined by concatenation of feature spaces.

A.5 Morphology-aware Alignments

The LIMSI-CNRS has been addressing the problem of translating from a language with analytic tendencies, such as like English, into a rather synthetic language, such as Russian or Czech. Indeed, there are cases where the grammatical information encoded in a Czech or Russian word seems unnecessary, like the accusative case marker that expresses the object function of a noun phrase, a grammatical information that has no equivalent at the word level in English. A well-known strategy in this situation is to remove the case marker from the morphologically complex word in order to make it more similar to the English word it should align to.

One should notice that such a rough normalization exposes us to removing information that is useful to make a correct translation, despite the fact that this information is not encoded in the English word. For instance, when English uses a preposition to express the idea of belonging, Czech language uses a genitive case marker that plays a similar role, as in *the engine* ***of*** *the car - motor aut****a***. Removing the genitive case from the Czech word leaves no possible correct alignment for the English preposition.

Thus, when aligning a synthetic language to an analytical one, many target (function) words remain unaligned (see figure 2.a). This is an issue during the phrase extraction phase, as null-aligned words tend to generate many spurious phrases (in phrase-based systems) or grammar rules (in hierarchical systems). We have proposed a factored alignment model designed to complete word-to-word alignments by adding morpheme-to-word links (see figure 2.b). Technically, this model is a simple extension of IBM model 2, which can accomodate bundle of morphological features (rather than words) on the target side. Using this model, we were able to greatly reduce the number of non-aligned words on the target side, yielding more compact and less noisy translation models, with however hardly any impact so far on the translation quality as measured by automatic metrics such as BLEU. For the sake of comparison, we have also developed a similar system, based on hand-crafted rules; the rule-based approach proved to yield the worst results. We finally ran our model on Russian and Romanian data, for which we got results that were similar to Czech, showing slightly better results in terms of BLEU when translating into the synthetic language (see Table 10).

These tendencies are confirmed by the observation of the output and seem reasonable, since grammatical case is the major morphological category ignored by baseline alignments. The new phrase table we obtain with morphological alignments helps to better predict case inflection, mainly according to the preposition in the source sentence. Indeed, Table 11 shows the wrong translation of the English preposition *by* in the baseline system where the Czech noun phrase is in nominative case. Our system successfully translates the preposition by the instrumental case needed for such passive constructions.

The reported improvement over the baseline systems is not confirmed by a straight BLEU improvement. However we showed that one-to-many alignments from a synthetic language to

	Baseline		Morph_align	
Language pair	BLEU	Phrase Table Size	BLEU	Phrase Table Size
Cs-En	20.34	22,799,794	20.26	21,247,701
En-Cs	14.09		14.21	
Ru-En	25.27	44,051,989	24.98	42,053,080
En-Ru	21.34		21.16	
Ro-En	34.56	31,914,134	34.70	28,795,129
En-Ro	29.21		29.38	

Table 10: Results in BLEU and phrase table size (Moses).

source	who are captured by Ukrainian soldiers
baseline	kteří zadržený ukrajinští vojáci
	who-Plur captured-Passive-Sing Ukrainian-Nom soldiers-Nom
+ morph-alignments	kteří jsou zajati ukrajinskými vojáky
	who-Plur are captured-Passive-Plur Ukrainian-Ins soldiers-Ins

Table 11: Better case prediction and agreement (English-Czech).

English help to better take into account the specificities of each language. While the English output has more words than in the baseline system, such as negative adverbs, auxiliaries, pronouns, the Czech, Russian and Romanian outputs are more concise, showing e.g. fewer incorrect verbal constructions and more reliance on inflection, which leads to better agreement.