BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **17**, No 2
Sofia • 2017 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2017-0014

Machine Translation: Phrase-Based, Rule-Based and Neural Approaches with Linguistic Evaluation

Vivien Macketanz[1], *Eleftherios Avramidis*[1], *Aljoscha Burchardt*[1], *Jindrich Helcl*[2], *Ankit Srivastava*[1]

[1]*German Research Center for Artificial Intelligence (DFKI), Language Technology Lab, Berlin, Germany*
[2]*Institute of Formal and Applied Linguistics, Charles University, Czech Republic*
E-mails: vivien.macketanz@dfki.de eleftherios.avramidis@dfki.de aljoscha.burchardt@dfki.de helcl@ufal.mff.cuni.cz Ankit Srivastava@dfki.de

Abstract: *In this article we present a novel linguistically driven evaluation method and apply it to the main approaches of Machine Translation (Rule-based, Phrase-based, Neural) to gain insights into their strengths and weaknesses in much more detail than provided by current evaluation schemes. Translating between two languages requires substantial modelling of knowledge about the two languages, about translation, and about the world. Using English-German IT-domain translation as a case-study, we also enhance the Phrase-based system by exploiting parallel treebanks for syntax-aware phrase extraction and by interfacing with Linked Open Data (LOD) for extracting named entity translations in a post decoding framework.*

Keywords: *Machine translation, Parallel treebanks, Entity linking, Manual evaluation, Neural approaches.*

1. Introduction to deep Machine translation and evaluation

With the recent appearance of neural approaches to Machine Translation (MT), we are dealing with three main MT paradigms: Rule-Based MT systems (RBMT), “classical” phrase-based Statistical MT systems (SMT) and Neural MT systems (NMT), the latest type of statistical systems. Translating between two languages requires substantial modelling of knowledge about the two languages, about translation, and about the world. Interestingly enough, little effort has been spent in the past on understanding what exactly MT systems learn, or to ask a simpler question – what aspects of language they can deal with and what remains challenging.

Unfortunately, today’s automatic measures for MT quality are not able to detect and model these aspects of translation in a detailed and analytical way. As one consequence, particular differences in the translations of different systems or system variants that may or may not constitute improvements remain undetected. Therefore, we have argued for an evaluation approach that extends the current MT evaluation practice by steps where language experts inspect systems outputs [1]. We have started to use this extended evaluation approach in our contribution to the WMT2016 IT task [2] and presented it also at the Workshop on Deep Language Processing for Quality Machine Translation (DeepLP4QMT) in Varna, Bulgaria. In this contribution, we will provide a much extended description of our novel evaluation method driven by linguistic phenomena assembled in a test suite.

While test suites are a well-known tool that has often been used in Natural Language Processing (NLP), e.g., to test the performance of a parser, they are employed in MT only to a minor degree. One of the reasons might be that the complexity of languages makes it difficult to evaluate the MT output and draw conclusions from the findings. Nevertheless, in narrow domains there seems to be interest in detecting differences between systems and within the development of one system, e.g., in terms of pronouns [3] or verb-particle constructions [4]. A related fertile area of research is the series of shared tasks on cross-lingual pronoun prediction wherein similar to our linguistically-driven evaluation the discourse phenomenon (pronouns) is evaluated on competing MT systems using a “test suite” of lemmatised target-language human-authored translations [5].

In this paper we want to show to what extent a linguistically-driven evaluation may grant interesting insights into the nature of different MT systems and how these observations may help to improve the systems. In order to achieve this aim, we present a domain-specific as well as a domain-independent analysis.

Machine translation like other language processing tasks is confronted with the Zipf’ian distribution of relevant phenomena. Although surface-data-driven systems have enlarged the head considerably over the last years, the tail still remains a challenge. Many approaches have therefore tried to include various forms of linguistic knowledge in order to systematically address chunks of the tail [6].

One goal of this paper is to show how we can extend the classical phrase-based SMT systems in this direction. Adding to previous work [2], we will report more in-depth on “deeper”, more knowledge-driven ingredients of our work, namely (i) exploiting parallel treebanks for syntax-aware phrase extraction in SMT, and (ii) using Linked Open Data (LOD) for extracting named entity translations as a post-decoding module. Both parallel treebanks and LOD have been integrated in SMT systems previously. Syntactically annotated corpora have been used directly in syntax-based models [7-9] as well as indirectly as an augmentation to the non-linguistic phrase pairs [10-12]. In this paper, we follow the latter approach by extracting linguistically motivated phrase pairs from aligned and parsed corpora and appending them to the standard phrase-based SMT models. We extended the aforementioned works (primarily focusing on parliamentary proceedings) to new domains (IT-domain). There have also been several attempts to exploit linked data (resources stored on the web and connected via web links) into translating nouns and named entities in SMT systems [13, 14]. We implement a similar approach and enrich our phrase-based SMT system with translations from semantically linked knowledge bases.

2. Method

2.1. Baseline Machine translation systems

The extensions and evaluations we describe below start from three baseline systems:

The **phrase-based SMT** baseline is a domain-enhanced version of several state-of-the-art phrase-based systems, as indicated in the Shared task of Machine translation in WMT [15]. As the best system UEDIN-SYNTAX [16] included several components that were not openly available, we proceeded with adopting several settings from the next best system UEDIN [17], also given the fact that the difference of their ranking position is minimal (0.587 vs 0.614 BLEU score for English-German which was not statistically significant as a difference). The generic parallel training data (Europarl [18], News Commentary, MultiUN [19], Commoncrawl [20]) are augmented with domain-specific data from the IT domain (Libreoffice, Ubuntu, Chromium Browser [21]). The monolingual target side of the above corpora, along with the WMT News Corpus, is used for training one language model per corpus, whereas all of these intermediate language models are interpolated on in-domain data to form the final model used within the phrase-based decoding. In this paper, we describe two enhancements to the phrase-based SMT baseline, namely syntax-aware phrase extraction and linked-data-aware post-processing in Sections 3 and 4 respectively. In the examples we refer to this system as “SMT”.

The **rule-based** baseline is Lucy [22], a system that has shown state-of-the-art performance in many shared tasks. In this method, translation occurs in three phases, namely analysis, transfer, and generation. All three phases consist of hand-written linguistic rules that can capture the structural and semantic differences between German and other languages. Additionally, manual inspection has shown that it provides better handling of complex grammatical phenomena, such as long distance dependencies, particularly when translating into German.

Our **neural MT algorithm** represents the state of the art. It follows the description of [23]. The input sequence is processed using a bidirectional RNN encoder with Gated Recurrent Units (GRU) [24] into a sequence of hidden states. The final backward state of the encoder is then projected and used as the initial state of the decoder. Again, our decoder is composed of an RNN with GRU units. In each step, the decoder takes its hidden state and the attention vector (a weighted sum of the hidden states of the encoder, computed separately in each decoding step), and produces the next output word.

In addition to the attention model, we use byte pair encoding [25] in the preprocessing step. This ensures that there are no out-of-vocabulary words in the corpus and, at the same time, enables open-vocabulary decoding.

We trained our model on the same data as the phrase-based SMT baseline system and used the first 1000 segments of the QTLeap corpus (**http://metashare.metanet4u.eu/go2/qtleapcorpus**) for validation during training. In the experiments, the sentence length was limited to 50 tokens. The size of the hidden state of the encoder was 300 units, and the size of the hidden state of the decoder was 256 units. Both source and target word embedding vectors had 300 dimensions. For training, a batch size of 64 sentences was used. We used dropout and L2 for regularization.

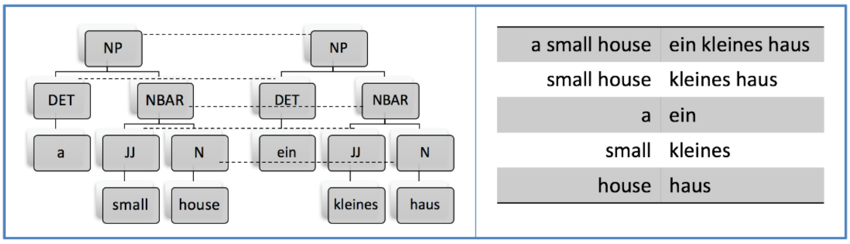
Our model was implemented using Neural Monkey [26], a sequence to sequence learning toolkit built on top of the Tensorflow framework [27]. In the examples we refer to this system as “neural”.

2.2. Syntax-aware phrase extraction

Herein we define a linguistic enhancement to the phrase-based SMT baseline system described in Section 2. Under standard configuration such as in the baseline phrase-based SMT system, phrase pairs are extracted from parallel (sentence-aligned) corpora by obtaining word alignment in both directions and using heuristics such as the Grow-Diag-Final (GDF) algorithm [28].

The phrase pairs in the baseline system are not linguistically motivated which in turn leads to a number of errors in translation such as missing verbs. We extract linguistically motivated phrase pairs by obtaining phrase structure parse trees for both the source and target languages (on the same data as the baseline system) using monolingual constituency structure parsers such as the Berkeley Parser [29], and then aligning the subtrees using a statistical tree aligner [30]. These phrase pairs (illustrated with an example in Fig. 1) are then merged with the phrase pairs extracted in the baseline SMT system into one translation model. Thus we are merely using syntax to constrain the phrase boundaries and enabling SMT decoder to pick syntax-aware phrases, thereby ensuring noun phrases and verb phrases remain cohesive.

Through experimentation detailed in [31], we have discovered that non-linguistic phrase-based models (baseline phrase-based SMT) have a long tail (of coverage) and syntax-aware phrases underperform, if not concatenated with non-linguistic phrase pairs. We observed the syntax-aware system scored 0.8 BLEU points over the baseline system. Note that this system is referred to as the “SMT-syntax” system hereafter in the evaluations.



a small house	ein kleines haus
small house	kleines haus
a	ein
small	kleines
house	haus

(a) (b)

Fig. 1. Example of a: parallel treebank entry (a); associated set of extracted phrase pairs (b)

2.3. Named entity translation using linked data

In this section, we describe another enhancement to the baseline system: named entity translation. Named entities are terms (usually nouns like people names, places, organizations, locations or technical terms) which have a fixed (consistent) translation. SMT systems often translate them inconsistently or are unable to translate them on account of the named entity being absent in the models (unknown words).

One technique to address this deficiency is to integrate the SMT system with a Named Entity Recognition (NER) system, i.e., annotate all words and phrases in the source language which are identified as named entities. These named entities are then linked with a bilingual dictionary to retrieve translations in the target language which are then inserted into the translation in a post-decoding process.

For the dictionary in our experiments, we exploit multilingual terms semantically linked with each other in the form of freely available linguistic linked data on the web such as DBpedia (**http://wiki.dbpedia.org**) to identify named entities in our dataset in the same vein of [13]. These entities and their linked translations are then integrated with the translations of the baseline system such that the translations from DBpedia overwrite the baseline system translations. A step-by-step procedure for translating named entities in this manner is detailed in [32].

Note that although many unknown words are correctly identified and translated, DBpedia is a user-generated dictionary sourced from Wikipedia and is prone to contain errors or a different term altogether ("Microsoft Paint" versus "MS Paint") which may reflect poorly in automatic evaluation metrics. This is another motivation to exercise the deep manual evaluation on enhancements to the baseline phrase-based SMT system. Hereafter "linked data" is used to refer to this MT system in the linguistic evaluations.

3. Manual linguistic evaluation

The evaluation of our systems is comprised of a deep manual analysis, performed by a professional German linguist (Following the general practice in industry, only one trained person, in this case the linguist, does the quality assurance). The goal of the manual evaluation was to validate the systems' capabilities of specific linguistic phenomena. Apart from gaining insights into the nature of the errors, this method can also provide guidance for setting priorities for future extensions and improvements of the systems.

The manual evaluation has been performed on a variety of MT systems so far, the interested reader is referred to [2, 33].

3.1. Manual evaluation procedure

For the human-based analysis the following procedure was found to be a good practice: In a first step, the linguist browses through the outputs of the different systems and detects errors related to linguistic phenomena that seem prevalent and systematic. Additionally, we have consulted professional translators that provided us with a list of possible (machine) translation errors in the technical domain. With this approach we make sure that we do not miss any important linguistic categories that might lead to errors. To this end, we use the domain corpora of the WMT 2016 IT-translation task, namely the QTLeap corpus. The result of this first step is a short list of phenomena that require closer inspection.

Note that we understand "linguistic phenomenon" in a pragmatic sense covering a wide range of issues that impact translation quality. This can include not only common morpho-syntactic and semantic categories, but also formatting issues, issues of style, etc.

In order for this manual evaluation to be transferable to other contexts and domains as well, we are currently creating an expansive test suite (English <> German) that will be published elsewhere, containing a wide range of various linguistic phenomena (cf. Section 5.3). By selecting only the categories that are needed in a given context or domain, this test suite can serve as a basis for evaluation in various settings.

As it would be too time-consuming to perform a deep analysis of the complete corpus, 100 source segments that contain the respective linguistic phenomenon are randomly selected. Based on the source sentences, all the instances (by "instance" we refer to each occurrence of the phenomenon, e.g., each verb, term, etc.) of the respective phenomenon are counted in the 100 selected target segments of each system. Consequently, the occurrences of correctly translated phenomena in the system outputs are counted. The percentage of correctly translated phenomena is calculated by dividing the overall number of correctly translated instances by the overall number of instances in the source sentences. As one segment may consist of several sentences and contain several phenomenon instances, overall instance numbers can be greater than 100.

Additionally, certain key rules are followed in the evaluation process: First of all, the translation of the linguistic phenomena does not have to be equivalent to the reference translation, as there may be several correct translations. Furthermore, if a linguistic phenomenon is realized in a different structure that correctly translates the meaning, the output is counted as a correctly translated instance, cf. Example 1, in which the compound (*F11 key*) can either be translated as a compound like in the reference translation (*F11-Taste*) or as a (stylistically slightly dispreferred) noun modifier-construction like in the MT output (*Taste F11*) (The fact that the MT output produces the unnecessary verb "angezeigt" is being ignored here, as the focus always lies on only one phenomenon at a time).

Example 1.

Source: *Try pressing the **F11 key**.*
Syntax: *Drücken Sie die **Taste F11** angezeigt.*
Reference:*Betätigen Sie einfach die **F11-Taste**.*

Note that the reference in this case introduces a spurious adverb (*einfach* – *simply*). This is one of several issues that we detected in the given corpus. It can be affiliated to the fact that – as has frequently been observed – human reference translations are sometimes not of perfect quality, depending on the circumstances of their creation.

The linguistic phenomena we identified as particularly prone to translation errors in the given corpus include imperatives, compounds, quotations marks, menu item separators (">"), missing verbs, phrasal verb and terminology (as the segments were from the technical helpdesk domain). All these phenomena were analyzed separately, which means that the correctness of phenomena occurring within other phenomena (e.g., phrasal verbs within imperative constructions) is ignored when analyzing the latter.

The central idea of the test suite evaluation approach is to focus on certain phenomena and aspects of translation at a given time. For the test item, the criterion is not whether the whole phrase has been translated correctly. It depends on the test item what needs to be present in the translation to count as a correct treatment of the test item at hand.

Below, we will amplify how we treated the evaluation of the various phenomena in detail, as the different categories required individual regulations in addition to the universal key rules mentioned above.

Imperatives. Imperatives were classified as correct when the translation contained the proper imperative verb form and formal address pronoun (*Sie*). It is not required that the translation of the verb itself is correct, as the only important requirement was that the imperative construction is realized (correctly). As mentioned above, missing elements of phrasal verbs are irrelevant for the correctness of the construction.

The following example shows a segment that contains two imperatives (*slide* and *Tap*), of which only the second is translated correctly by the SMT system (*Tippen Sie*), whereas the first imperative is mistranslated as the noun *Folie* (because *slide* is a homonym that can mean both). The neural system leaves out the verbs completely. The reference translation contains the two correct translations (*gleiten Sie* and *Tippen Sie*).

Example 2.

Source: *On the Home screen, quickly* **slide** *left to the Applications list.* ***Tap*** *Settings> About> More info.*

SMT: *Über die innere Bildschirm, schnell* **Folie** *links nach der Anwendungsliste auswählen.* **Tippen Sie** *Einstellungen > About > Mehr.*

Neural: *Auf dem Home Bildschirm links links auf die Applications Liste. Einstellungen > Über > Weitere Informationen.*

Reference: *Auf dem Startbildschirm,* **gleiten Sie** *schnell nach links zur Anwendungsliste.* **Tippen Sie** *auf Einstellungen> Über> Mehr Infos.*

Compounds. A compound is counted as correctly translated when its components either form a German compound, which means that they have to be realized in one word, or when the components form a morphological construction that expresses the compound (e.g., adjective + noun, or genitive construction).

Example 3 contains three compounds (*cartridge contact – Patronenkontake*, *ink cartridges – Tintenpatronen* and *connectivity issues – Verbindungsprobleme*). The RBMT system correctly translates all three compounds (*Patronenkontakte*, *Tintenpatronen*, *Konnektivitätsfragen* (the latter being a literal translation instead of the correct term, a fact which is being ignored in this case as the focus lies on the compound itself). The SMT-syntax (in the examples referred to as "syntax") system correctly translates the first two compounds (*Patronen-Kontakte* (the variant with the hyphen is not common but not incorrect either) and *Tintenpatronen*), while it mistranslates the third compound (*Anbindungsmöglichkeiten Fragen*).

Example 3.

Source: *Try to remove it and blow into the* ***cartridge contacts*** *as well as the printer's. If they are recycled* ***ink cartridges****, there could be* ***connectivity issues****.*

RBMT: *Versuch, um es zu beseitigen und in die* ***Patronenkontakte*** *ebenso wie den Drucker zu blasen. Wenn sie wiederverwertete* ***Tintenpatronen*** *sind, könnte es* ***Konnektivitätsfragen*** *geben.*

Syntax: *Versuchen, um er zu entfernen und Schlag in die* ***Patronen-Kontakte*** *sowie die Drucker. Wenn sie wiederaufbereitet werden* ***Tintenpatronen****, es könnte* ***Anbindungsmöglichkeiten Fragen.***

Reference: *Versuchen Sie, es zu entfernen und blasen in die* ***Patronenkontakte*** *als auch die des Druckers. Wenn es recycelte* ***Tintenpatronen*** *sind, könnte es* ***Verbindungsprobleme*** *sein.*

Quotation marks. The quotation marks need to be placed around the right word in order to be counted as correct. They are not counted pairwise but separate as it may be the case that only one of a pair is placed correctly while the second one is missing or placed somewhere else. When there are more instances of quotation marks in the MT output than in the source, every redundant quotation mark is subtracted from the overall count of the respective segment, as it is the case in Example 4.

The source sentence in Example 4 comprises four quotation marks, but the RBMT system produces an output with five quotation marks. The first one of them is placed correctly (before (*Advanced*) – (*Fortgeschritten*)) while the second one is misplaced (after *an* instead of *Privacy* – *Privatsphäre*). Furthermore, the other two quotation marks from the reference around *Clear browsing data* are placed correctly around *Klare Browsingdaten* but the system added an additional quotation mark after *Klare*. Hence, even though the MT system achieves three correct instances, subtracting the redundant quotation mark results in two correct instances. The SMT system on the other hand places the right amount of quotation marks at the right positions.

Example 4.

Source: [...] *Touch "*(*Advanced*) *Privacy". Select "Clear browsing data".*

RBMT: […] *Fassen Sie* „(*Fortgeschritten*) *Privatsphäre an*". *Auserlesene* „*Klare*" *Browsingdaten*".

SMT: […] *Touch* „(*Advanced*) *Datenschutz*". *Wählen Sie* „*Browserdaten löschen*".

Reference: […] *Berühren* „(*Erweitert*) *Datenschutz*". *Wählen Sie* „*Browserdaten löschen*".

Menu item separators. The menu item separator ">" is counted in the same way as the quotation marks: The placement between two words needs to be correct in order for the menu item separator to be counted as correctly translated. Furthermore, the same rule concerning additional separators holds, meaning that those will be subtracted from the segment count.

Example 5 demonstrates the incorrect and correct translation of the menu item separator “>”: The source sentence contains two separators. Even though the RBMT system places the two separators in its output between the right words, it adds hyphens before and after the separators, converting the three words around the separators into one long compound. The linked data system represents the correct placement of the separators.

Example 5.

Source: *Go to Settings > General > Code Blocking.*
RBMT: *Gehen Sie zu Einstellungs->-General->-Code-Blockierung.*
Linked data: *Gehen Sie zu Einstellungen > Allgemein > Code Blocking.*
Reference: *Gehen Sie auf Einstellungen > Allgemein > Codesperre.*

Verbs. For the translation of a verb in order to be counted as correct it is important that the verb is present in the MT output. The verb needs to be translated correctly or at least partly correctly as for example incomplete phrasal verbs are counted as correct. Every occurring verb form is counted separately. The conjugation does not need to be correct and verbs realized as nominalizations are also counted as correct. As has been said above, we allow ourselves a certain freedom what we call a linguistic phenomenon as our goal is not to create a linguistic theory, therefore verbs in fixed commands are not counted as they rather belong to terminology. Furthermore, it needs to be taken account of the fact that English progressive constructions (consisting of two verb forms) do not exist in German and are translated into a single verb which means that those constructions should be counted as one instance instead of two in the source sentence.

The source sentence in Example 6 contains four instances of verbs (*have*, *go*, *choose* and *are programming*) as the progressive construction *are programming* counts as one instance. The SMT system leaves out the verb *go – gehen* and mistranslates the progressive construction as verb + noun (*sind Programmierung*) which is a frequently occurring error. The RBMT system does not produce either of those errors as it correctly translates all four verbs (*müssen*, *gehen*, *wählt* and *programmieren*). Note that the conjugation of the verb *wählt* is incorrect (cf. reference *auswählen*) but as mentioned above this translation counts as correctly translated (see below for the case of *wählen* vs. *auswählen*).

Example 6.

Source: [...] *You **have** to **go** to the Language menu and there **choose** the language in which you **are programming**.*
SMT: [...] *Sie **haben**, um das Language Menü und **wählen** Sie die Sprache, in der Sie **sind Programmierung**.*
RBMT: [...] *Sie **müssen** zum Sprach-Menü **gehen** und es **wählt** die Sprache, in der Sie **programmieren**.*
Reference: [...] *Sie **müssen** ins Sprachen Menü **gehen** und die Sprache **auswählen**, in der Sie **programmieren**.*

Phrasal verbs. German phrasal verbs have the characteristic that their prefixes move to the end of the sentence in certain constructions. The moved prefix is prone to getting lost in a machine translation or not moving to the end of the sentence but instead staying in its initial position. Therefore, only translations that contain the verb as well as its prefix (in the expected position) are counted as correct. Nevertheless, the evaluation of this phenomenon is not always easy as there are often cases where the English verb can be translated with a phrasal verb or a regular verb, which means that if a regular verb is present it needs to be counted as a correctly translated phrasal verb. Moreover, there are phrasal verbs that are acceptable with and without their suffix (e.g., ***aus****wählen* vs. *wählen*). Hence, the analysis of the translation of the phrasal verbs needs to be treated with caution.

The verb *depend* in the source sentence in Example 7 translates into German as *abhägen*. In the given sentence, the prefix *ab* moves to the end of the sentence, as is it the case in the reference and the SMT-syntax system. In the baseline SMT translation on the other hand the prefix stays in its initial position which is incorrect.

Example 7.

Source: *It* ***depends***. [...]
SMT: *Es* ***abhängt***. [...]
Syntax: *Das* ***hängt*** *davon* ***ab***. [...]
Reference: *Es* ***hängt*** *davon* ***ab***. [...]

Terminology. In order to be counted as correct, a translation of a term either needs to match the reference or the translation needs to be found in Microsoft's Language Portal for Terminology (**https://www.microsoft.com/Language/en-US/Search.aspx**). Commands consisting of more than one word (e.g., *Save as…*) are counted as one single term. Compounds on the other hand are counted as separate terms (e.g., *router page* is counted as two instances). Moreover, proper terms also belong to terminology. Case sensitivity needs to be taken into account.

In Example 8 the source sentence contains the three terms *desktop, right-click* and *icons* that should be translated into *Desktop, klicken Sie mit rechten Maustaste* and *Symbole* in German, as can be seen in the reference. While the SMT system correctly translates *desktop – Desktop* and *icons – Symbole*, it leaves out the verb and the pronoun (*klicken Sie*) in *right-click – klicken Sie mit der rechten Maustaste*, resulting in two correct instances. The linked data system correctly translated *icons – Symbole* and also leaves out the verb and pronoun in the second term. Additionally, it translates *desktop* as *Schreibtisch* – which is not an incorrect translation in general, but is incorrect in this technical domain.

Example 8.

Source: *On the* ***desktop****,* ***right-click*** *in the area without icons* [...].
SMT: *Auf dem* ***Desktop, mit der rechten Maustaste*** *auf dem Gebiet ohne* ***Symbole*** [...].
Linked data: *Auf dem* ***Schreibtisch, mit der rechten Maustaste*** *auf dem Gebiet ohne* ***Symbole*** [...].
Reference: *Auf dem* ***Desktop klicken Sie mit der rechten Maustaste*** *in den Bereich ohne* ***Symbole*** [...].

3.2. Manual Evaluation results

For the seven linguistic categories depicted in the previous section, 657 source segments were extracted for the human-based analysis (for the category of phrasal verbs only 57 instead of 100 segments could be found in the given corpus, leading to a total of 657 instead of 700). As described above, each source segment contains at least one instance of the respective phenomenon, in many cases several instances could be found within one segment in this analysis, resulting in 2104 phenomena overall (Table 1).

As it can be seen in Table 1, the overall *average* performance of the systems is very similar for all systems. The SMT, RBMT and neural system slightly outperform the other systems with a 0.95 confidence level on the average performance. This is an interesting observation as the performance on the linguistic phenomena is quite diverse: While a shallow evaluation would render the systems more or less identical, this view makes it possible to identify and study their strengths and weaknesses in detail. Note that none of the systems was optimised to perform particularly well on these phenomena, although it is expected that the RBMT system already contained hand-written rules to handle linguistic phenomena.

Table 1. Translation accuracy on manually evaluated sentences focusing on particular phenomena. Boldface indicates best system on each phenomenon (row) with 0.95 confidence level

Phenomenon	#	SMT	RBMT	SMT-syntax	Linked data	Neural
Imperatives	247	68%	79%	68%	68%	74%
Compounds	219	55%	87%	55%	56%	51%
“>”-separators	148	99%	39%	97%	97%	93%
Quotation marks	431	97%	94%	93%	94%	95%
Verbs	505	85%	93%	81%	85%	90%
Phrasal verbs	89	21%	67%	7%	11%	38%
Terminology	465	63%	50%	53%	51%	55%
Sum	2104					
Average		76%	76%	71%	72%	75%

While the baseline SMT system outperforms the other systems on terminology and (except for the neural) on quotation marks, all three SMT systems outperform the RBMT system on the “>”-separators, but only the baseline is significantly better than the neural in this category. The RBMT system shows a complementary performance compared to the SMT baseline system, as it outperforms all other systems on compounds, verbs and phrasal verbs, and outperforms the three SMT systems on imperatives.

It can be stated that the RBMT system shows the tendency to perform better on the morpho-syntactic linguistic categories (i.e., imperatives, compounds, verbs and phrasal verbs), while the baseline SMT systems seem to be able to handle the remaining categories better (namely the “>”-separators, quotation marks and terminology). The tendency on the performance regarding these categories is similar for the other two SMT systems (SMT-syntax, linked data) but generally less pronounced.

Concerning the neural system, the individual categories indicate that it performs very close to the SMT system in overall, but it improves significantly on it concerning verbs and phrasal verbs. Furthermore, it is remarkable that the neural system is the only system that reaches 90% accuracy or more on three categories. Nevertheless, our neural system is rather premature and there may still be implementation issues. Particularly the fact that compounds are not properly formed despite the byte pair encoding indicates that further work needs to focus on the performance and integration of this module.

The generally lower scores on phrasal verbs and terminology for all systems might be an indication that at the present time these are the categories (at least of those categories we inspected) causing the most difficulties for all systems – regardless of the nature of the system. Thus, future work might in a first step focus on tackling these problems.

Lastly, it is interesting to mention that there were cases in which the translation of the MT system was found to be better than the reference translation, as can be seen in Example (9) in which the reference contains the English spelling of the term *email*. The correct German spelling can be seen in the SMT-syntax output (*E-Mail*).

Example 9.

Source:	*Send an* ***email*** *to* [...].
Neural:	*Senden Sie eine* ***E-Mail*** *an* [...].
Reference:	*Senden Sie eine* ***email*** *an* [...].

3.3. Evaluation of test suite data

In addition to the evaluation of our systems on the seven domain-specific categories, we also evaluated the systems on a small-scale generic test suite by creating 100 test sentences of 50 general linguistic categories (two sentences per category). These 50 linguistic categories can be condensed to fourteen super-categories.

The evaluation process was conducted the same way as in the domain-specific analysis: The correctly translated phenomena per category were counted and thus the overall sum of correctly translated phenomena was divided by the overall number of instances in the phenomenon.

Even though we are aware that the analysis on such few instances per category is not necessarily representative, the evaluation of this data still provides interesting insights into the distinct nature of the systems. Table 2 shows the behaviour of the systems on the different super-categories.

The best-performing system on this data selection is the RBMT system, as it shows an average percentage of correct translations more than twice of the SMT, SMT-syntax and linked data systems. While the latter three systems have very similar average scores ranging 28-31%, the neural system has the second-highest average score, namely 48%.

The three SMT systems do not only have similar overall average scores but also behave similarly regarding various phenomena: In six of the fourteen super-categories, the baseline SMT, SMT-syntax and linked data system correctly translate the same percentage of test sentences (on false friends, function words, composition, Named Entity (NE) & terminology, negation and punctuation). On four of these super-categories, all three systems reach 50% or more, the SMT baseline and SMT-syntax additionally have 50% or more on two categories.

The neural system reaches 50% or more on eight of the fourteen categories while the RBMT system shows this property on eleven systems.

Table 2. Translation accuracy on test suite sentences focusing on particular phenomena. Boldface indicates best system on each phenomenon (row) with a 0.95 confidence level when significant

Supercategory	#	SMT	RBMT	SMT-syntax	Linked data	Neural
Ambiguity	2	50%	50%	50%	0%	0%
Coordination & ellipsis	8	13%	13%	0%	0%	13%
False friends	2	100%	50%	100%	100%	50%
Function word	4	50%	75%	50%	50%	100%
LDD & interrogative	16	25%	69%	19%	25%	63%
MWE	10	40%	40%	50%	50%	10%
Composition	2	0%	100%	0%	0%	0%
NE & terminology	6	50%	67%	50%	50%	33%
Negation	2	50%	100%	50%	50%	50%
Non-verbal agreement	8	50%	88%	38%	38%	25%
Punctuation	2	0%	0%	0%	0%	50%
Subordination	10	20%	90%	30%	30%	60%
Verb tense/mood/asp.	18	33%	89%	17%	11%	78%
Verb valency	10	10%	80%	30%	30%	50%
Sum	100					
Average		31%	69%	29%	28%	48%

Even though we know that this small-scale study is not fully representative, we calculated the statistical significance of the best system per phenomenon. We found the RBMT and neural system to be the best systems on the function words, Long Distance Dependency (LDD) & interrogative and verb tense/mood/aspect. The SMT-syntax and the linked data system are outperforming the neural system on the MultiWord Expressions (MWE), but not the baseline SMT or the RBMT. The RBMT is additionally the best system on composition and non-verbal agreement. Furthermore, the RBMT is better than the SMT-syntax and linked data system on subordination and verb valency. In these two categories, the neural additionally outperforms the SMT system. It is also worth noting that out of two samples containing negation, only the RBMT system translated both of them correctly, whereas all of the statistical systems missed one.

The general conclusions that can be drawn from this small test suite evaluation are that the RBMT seems to handle the given linguistic phenomena better than the other systems. Moreover, the neural system is not as good as the RBMT system but still better than the SMT systems.

Among the SMT systems, the baseline SMT system treats the phenomena a little bit better than the other two SMT systems, just like in the domain-specific analysis. In addition to that, the RBMT system is in both analyses one of the best systems/the best system.

4. Conclusion and outlook

In this paper we have described several ways of making machine translation more linguistically aware. We have attempted to introduce linguistically aware phrases in the models as well as show improvements in the translation of named entities by linking with semantic web resources such as the DBpedia. Our detailed evaluation of relevant linguistic phenomena has shown that the performance of the MT systems differs considerably with respect to these phenomena while their overall performance in terms of errors made on these phenomena is very much the same. While the extended systems had previously shown performance improvements in automatic tests on larger corpora, we could not find such indications in the selected test items. However, the systems were not optimized for performance on the test suite. In this sense, this approach can really be seen as a "stress test". Moreover, the manual evaluation of SMT-syntax and linked data systems highlight the limitations inherent in such approaches dependent on external tools with their own set of errors. For example, the SMT-syntax system is sensitive to errors in the respective language parsers as well as the statistical tree aligner employed to extract the linguistically motivate phrase pairs. The linked data system obtains its translations from user-generated knowledge bases and is also limited by the performance of the Named Entity Recognition system employed to identify the named entities.

Interestingly enough, the more general evaluation shows first indications that the neural system is capable of learning several aspects of the language that are coded in the rules of the RBMT in a better way than the phrase-based SMT systems. They certainly lack abstraction (and generalization) in this respect. We are convinced that this test-suite based approach will lead to more insights in the future and will become important, e.g., in the area of machine teaching for neural MT.

Given this detailed method and results, it is now possible to select/improve systems with respect to a given task (an extension of this work for the purpose of the WMT16 Shared Task is presented in [2]). For example, if there is a post-editor involved, one would focus on fixing issues that are hard to post-edit. If the goal is to provide information to end users, one would focus on those issues that affect readability most. This prioritization would not be possible when using today's automatic measures. One obvious way for improving statistical systems would be to create targeted training material focusing on the relevant aspects such as imperatives starting from the test items.

In order to adapt the evaluation for other language pairs, it might be helpful to draw inspiration from the evaluation done in the context of this paper, but it would also include extensive manual work due to this approach being language-dependant. Furthermore, the choice of the phenomena is a subjective decision, which means that many more/different categories could be investigated, as for instance lexical choice, modal verbs, etc.

Adaptation of the evaluation to a different task is a manual step involving human expertise. Once the community (including industry) has come up with a set of test suites for certain tasks/requirement, it will be easier to put together tests for new tasks from these sources.

***Acknowledgements*:** This article has received support from the EC's Horizon 2020 research and innovation programme under Grant agreements No 645452 (QT21), from FP7 (2007-2013) under Grant agreement No 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches", and by Project Digitale Kuratierungstechnologien (DKT), supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument "Wachstumskern-Potenzial" (No 03WKP45). We also thank the two anonymous reviewers for their valuable comments.

References

1. Burchardt, A., K. Harris, G. Rehm, H. Uszkoreit. Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation. – In: Proc. of LREC 2016 Workshop Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (LREC'16), Located at International Co, 2016.
2. Avramidis, E., A. Burchardt, V. Macketanz, A. Srivastava. DFKI's System for WMT16 IT-Domain Task, Including Analysis of Systematic Errors. – In: Proc. of 1st Conference on Machine Translation, 2016, pp. 415-422.
3. Guillou, L., C. Hardmeier. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. – In: 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia, 2016.
4. Schottmüller, N., J. Nivre. Issues in Translating Verb-Particle Constructions from German to English. – In: 10th Workshop on Multiword Expressions, Gothenburg, Sweden, 2014, pp. 124-131.
5. Guillou, L., C. Hardmeier, P. Nakov, S. Stymne, J. Tiedemann, Y. Verslay, M. Cettolo, B. Webber, A. Popescu-Belis. Findings of the 2016 WMT Shared Task on Cross-Lingual Pronoun Prediction. – In: Proc. of 1st Conference on Machine Translation, 2016, Berlin, Germany, pp. 525-542.
6. Steedman, M. Romantics and Revolutionaries. – Linguistic Issues in Language Technology, Vol. **6**, 2011, No 11, pp. 1-20.
7. Chiang, D. A Hierarchical Phrase-Based Model for Statistical Machine Translation. – In: Proc. of 45th Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, 2005, pp. 263-270.
8. Quirk, C., A. Menezes, C. Cherry. Dependency Treelet Translation: Syntactically-Informed Phrasal SMT. – In: Proc. of 45th Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, 2005, pp. 271-279.
9. Galley, M., et al. Scalable Inference and Training of Context-Rich Syntactic Models. – In: Proc. of 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'06), Sydney, Australia, 2006, pp. 961-968.
10. Tinsley, J., M. Hearne, A. Way. Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. – In: Proc. of 6th International Workshop on Treebanks and Linguistic Theories (TLT'07), Bergen, Norway, 2007, pp. 175-187.
11. Hearne, M., S. Ozdowska, J. Tinsley. Comparing Constituency and Dependency Representations for SMT Phrase-Extraction. – In: 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08), Avignon, France, 2008.

12. Srivastava, A. K., A. Way. Using Percolated Dependencies for Phrase Extraction in SMT. – In: Proc. of Machine Translation Summit XII, Ottawa, Canada, 2009, pp. 316-323.
13. McCrae, J. P., P. Cimiano. Mining Translations from the Web of Open Linked Data. – In: Proc. of Joint Workshop on NLP, LOD and SWAIE, Hissar, Bulgaria, 2013, pp. 8-11.
14. Du, J., A. Way, A. Zydron. Using BabelNet to Improve OOV Coverage in SMT. – In: Proc. of 10th International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, 2016.
15. Bojar, O., et al. Findings of the 2013 Workshop on Statistical Machine Translation. – In: 8th Workshop on Statistical Machine Translation, 2013.
16. Nadejde, M., P. Williams, P. Koehn. Edinburgh's Syntax-Based Machine Translation Systems. – In: Proc. of 8th Workshop on Statistical Machine Translation, 2013, pp. 170-176.
17. Durrani, N., B. Haddow, K. Heafield, P. Koehn. Edinburgh's Machine Translation Systems for European Language Pairs. – In: Proc. of 8th Workshop on Statistical Machine Translation, 2013, pp. 114-121.
18. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. – In: Proc. of 10th Machine Translation Summit, Vol. **5**, 2005, pp. 79-86.
19. Eisele, A., Y. Chen. MultiUN: A Multilingual Corpus from United Nation Documents. – In: Proc. of 7th Conference on International Language Resources and Evaluation (LREC'10), 19-21 May 2010, La Valletta, Malta, pp. 2868-2872.
20. Buck, C., K. Heafield, B. Van Ooyen. N-Gram Counts and Language Models from the Common Crawl. – In: Proc. of Language Resources and Evaluation Conference, 2014.
21. Tiedemann, J. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. – In: Advances in Natural Language Processing. Vol. **V**. N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov, Eds. Borovets, Bulgaria. Amsterdam/Philadelphia, John Benjamins, 2009, pp. 237-248.
22. Alonso, J. A., G. Thurmair. The Comprendium Translator System. – In: Proc. of 9th Machine Translation Summit, 2003.
23. Bahdanau, D., K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. – In: 3rd International Conference on Learning Representations, 2015.
24. Cho, K., B. Van Merrienboer, D. Bahdanau, Y. Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. – In: Proc. of SSST-8, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 2014, pp. 103-111.
25. Sennrich, R., B. Haddow, A. Birch. Neural Machine Translation of Rare Words with Subword Units. – CoRR, Vol. **abs/1508.0**, 2015.
26. Helcl, J., J. Libovický. Neural Monkey: An Open-Source Tool for Sequence Learning. – Prague Bulleting of Mathematical Linguistics, Vol. **107**, 2017, pp. 5-17.
27. Abadi, M., et al. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv Preprint arXiv:1603. 04467, 2016.
28. Koehn, P., F. J. Och, D. Marcu. Statistical Phrase-Based Translation. – In: Proc. of 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 48-54.
29. Petrov, S., D. Klein. Improved Inference for Unlexicalized Parsing. – In: Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, New York, 2007, pp. 404-411.
30. Zhechev, V. Unsupervised Generation of Parallel Treebank through Sub-Tree Alignment. – Prague Bulletin of Mathematical Linguistics, Vol. **91**, 2009, pp. 89-98.
31. Srivastava, A. K. Phrase Extraction and Rescoring in Statistical Machine Translation. Dublin City University, 2014.
32. Srivastava, A. K., F. Sasaki, P. Bourgonje, J. M. Schneider, J. Nehring, G. Rehm. How to Configure Statistical Machine Translation for Linked Open Data. – In: Proc. of 38th Annual Conference on Translating and Computer, London, United Kingdom, 2016, pp. 138-148.
33. Avramidis, E., V. Macketanz, A. Burchardt, J. Helcl, H. Uszkoreit. Deeper Machine Translation and Evaluation for German. – In: Proc. of 2nd Deep Machine Translation Workshop. Deep Machine Translation Workshop (DMTW'16), 21 October 2016, Lisbon, Portugal, pp. 29-38.