SN Computer Science (2021) 2:330
https://doi.org/10.1007/s42979-021-00723-4

SURVEY ARTICLE

A Survey of Orthographic Information in Machine Translation

Bharathi Raja Chakravarthi[1] · **Priya Rani**[1] · **Mihael Arcan**[2] · **John P. McCrae**[1]

Received: 13 August 2020 / Accepted: 22 May 2021 / Published online: 7 June 2021
© The Author(s) 2021

Abstract
Machine translation is one of the applications of natural language processing which has been explored in different languages. Recently researchers started paying attention towards machine translation for resource-poor languages and closely related languages. A widespread and underlying problem for these machine translation systems is the linguistic difference and variation in orthographic conventions which causes many issues to traditional approaches. Two languages written in two different orthographies are not easily comparable but orthographic information can also be used to improve the machine translation system. This article offers a survey of research regarding orthography's influence on machine translation of under-resourced languages. It introduces under-resourced languages in terms of machine translation and how orthographic information can be utilised to improve machine translation. We describe previous work in this area, discussing what underlying assumptions were made, and showing how orthographic knowledge improves the performance of machine translation of under-resourced languages. We discuss different types of machine translation and demonstrate a recent trend that seeks to link orthographic information with well-established machine translation methods. Considerable attention is given to current efforts using cognate information at different levels of machine translation and the lessons that can be drawn from this. Additionally, multilingual neural machine translation of closely related languages is given a particular focus in this survey. This article ends with a discussion of the way forward in machine translation with orthographic information, focusing on multilingual settings and bilingual lexicon induction.

Keywords Orthography · Under-resourced languages · Machine translation · Rule-based machine translation · Statistical machine translation · Neural machine translation

Introduction

Natural language processing (NLP) plays a significant role in keeping languages alive and the development of languages in the digital device era [1]. One of the sub-parts of NLP is machine translation (MT). MT has been the most promising application of artificial intelligence (AI) since the invention of computers, which has been shown to increase access to information by the native language of speakers in many cases. One of such critical cases is the spread of vital information during a crisis or emergency [2, 3]. Recently, translation accuracy has increased, and commercial systems have gained popularity. These systems have been developed for hundreds of languages, and hundreds of millions of people gained access. However, some of the less common languages do not enjoy this availability of resources. These under-resourced languages lack essential linguistic resources, e.g. corpora, POS taggers, computational grammars. This is more pertinent for MT since most common systems require

✉ Bharathi Raja Chakravarthi
bharathi.raja@insight-centre.org

Priya Rani
priya.rani@insight-centre.org

Mihael Arcan
mihael.arcan@insight-centre.org

John P. McCrae
john.mccrae@insight-centre.org

[1] Unit for Linguistic Data, Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Galway, Ireland

[2] Unit for Natural Language Processing, Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Galway, Ireland

large amounts of high-quality parallel resources or linguists to make a vast set of rules. This survey studies how to take advantage of the orthographic information and closely related languages to improve the translation quality of under-resourced languages.

The most common MT systems are based on either Rule-Based Machine Translation (RBMT) or Corpus-Based Machine Translation (CBMT). RBMT systems [4–10] are based on linguistic knowledge which are encoded by experts. On the other hand, CBMT [11, 12] depends on a large number of aligned sentences such as Statistical Machine Translation (SMT) [13–18] and Neural Machine Translation (NMT) [19–22]. Unlike RBMT systems, which require expertise of linguists to write down the rules for the language, CBMT-based systems rely on examples in the form of sentence aligned parallel corpora. CBMT systems such as SMT and NMT have alleviated the burden of writing down rules, which is not feasible for all languages since human languages are more dynamic in nature.

However, CBMT systems suffer from the lack of parallel corpora for under-resourced languages to train machine translation systems. A number of the methods have been proposed to address the non-availability of parallel corpora for under-resourced languages, such as pivot-based approaches [23–25], zero-shot translation [26–30] and unsupervised methods [31–33], which are described in detail in following sections. A large array of techniques have been applied to overcome the data sparsity problem in MT, and virtually all of them seem to be based on the field of transfer learning from high-resource languages in recent years. Other techniques are based on lexical and semantic similarities of closely related languages which are more relevant to our survey on orthographic information in machine translation.

The main goal of this survey is to shed light on how orthographic information is utilised in the MT system development and how orthography helps to overcome the data sparsity problem for under-resourced languages. More particularly, it tries to explain the nature of interactions with orthography with different types of machine translation. For the sake of simplicity, the analysis presented here in this article is restricted to those languages which have some form of internet resources. The survey is organised as follows: second section explains the background information to follow this article. We present orthographic information in subsection. Third section describes the challenges of automatically using orthographic information in RBMT outputs. Fourth section presents an analysis of orthographic information in SMT systems. Fifth section presents an analysis of orthographic information in NMT systems. This survey ends with a discussion of the future directions towards utilising the orthographic information.

Background

In this section, we explain the necessary background information to follow the paper, different types of MT systems and the orthographic information available for MT.

Under-resourced Languages

Worldwide, there are around 7000 languages [34, 35]. However, most of the machine-readable data and natural language applications are available for very few popular languages, some of these are: Chinese, English, French, or German etc. For other languages, resources are scarcely available and, for some languages, not at all. Some examples of these languages do not even have a writing system [36–38], or are not encoded in major schemes such as Unicode. Due to the unavailability of digital resources, many of these languages may go extinct. With each language that is lost, we lose connection with the culture of the people and characteristics of the languages.

Alegria et al. [36] proposed a six-level language typology to develop language technologies that could be useful for several hundred languages. This classifies the world's languages based on the availability of Internet resources for each language. According to the study, the term resource-poor or under-resourced is relative and also depends on the year. The first level is the most resourced languages; the second level is languages in the top 10 languages used on the web. The third level is languages which have some form of resources in NLP. The fourth level considers languages which have any lexical resources. Languages that have a writing system but not in digital form are in the fifth level. The last level is significant, including oral languages which do not have a writing system of its own. For the purpose of this work, we define under-resourced languages to be those at the third and fourth levels as the challenges are purely technical rather than social in nature. Languages that lack extensive parallel corpora are known as under-resourced or low-resourced languages [39].

Languages that seek to survive in modern society need NLP, which requires a vast amount of data and linguistic knowledge to create new language technology tools for languages. Mainly, it is a big challenge to develop MT systems for these languages due to the scarcity of data, specifically sentence aligned data (parallel corpora) in large amounts to train MT systems. For example, Irish, Scottish Gaelic, Manx or Tamil, Telugu, and Kannada belonging to the Goidelic and the Dravidian languages, respectively are considered as under-resourced languages

due to scarcely available machine-readable resources as mentioned in Alegria et al. [36].

Orthographic Information

Humans are endowed with a language faculty that is determined by biological and genetic development. However, this is not true of the written form of the language, which is the visual representation of the natural and genetically determined spoken form. With the development of orthography, humans have not only overcome limitations with human short-term memory, and brain storage capacity, but also this development allows communication through space and time [40]. Orthography is a linguistic factor of mutual intelligibility which may facilitate or impede inter-comprehension [41].

The orthographic information of languages does not only represent the information of the language but also the psychological representation of the world of the users. Chinese orthography is unique in its own in the sense that it uses a logo graphic writing system. In such a system, each Chinese character carries visual patterns along with rich linguistic information. These characters are visualised in square space, which depends on the number of strokes a character has. Each character can be decomposed in two parts. *Radicals*, which carries the semantic meaning, whereby the other part tells about the pronunciation. According to Shuo WenJie Zi[1] new Chinese characters consist of 540 radicals but only 214 in modern Chinese [42]. The problems lie when the decomposition strategy does not comply with some of the characters. On the other hand, other Asian languages such as Korean and Japanese, have two different writing systems. Modern-day Korea uses the Hangul orthography, which is part of the syllabic writing system, and the other is known as Hanja, which uses classical Chinese characters. Like the history of writing in Korea, Japan to have two writing systems, Kana and Kanji, where Kanji is identified as Classical Chinese characters, and Kana represents sounds where each kana character is recognized as a syllable. As both Korean and Japanese are very different from Chinese and morphologically-rich languages, the adoption of Chinese characters was rather difficult. These problems also posed great difficulty in the field of translation and transliteration. Irrespective of all the differences and challenges these three Asian languages share common properties which could be significant advantages in MT.

Closely related languages share similar morphological, syntactic, orthographic properties. Orthographic similarity can be seen from two major sources. First one is based on the genetic relationship between languages such as based on language families, Germanic, Slavic, Gaelic and Indo-Aryan languages. The second one is based on the contact though geographical area Indo-Aryan and Dravidian languages in the Indian subcontinent [43]. Two languages posses orthographic similarity only when these languages have the following properties: overlapping phonemes, mutually compatible orthographic systems and similar grapheme to phoneme mapping. Tables 1 and 2 shows the example difference and similarities in writing systems in the same language family.

Table 1 The table categorises the languages of the Indo-European language family which share the same and have different orthographies

Similar orthography	Different orthography
English	Russian
Spanish	Hindi
French	Punjabi
German	Bangla
Dutch	Urdu
Portuguese	Persian

Table 2 The table categorises the languages of the Dravidian language family which have different orthographies

Different orthography
Tamil
Telugu
Malayalam
Kannada

The widespread and underlying problem for the MT systems is variations in orthographic conventions. The two languages written in two different orthography leads to error in MT outputs. Orthographic information can also be used to improve the machine translation system. In the following subsection, we describe the different orthographic properties related to MT.

Spelling and Typographical Errors

Spelling or typographical errors are to be handled very carefully in MT task as even a minor spelling error could generate an out-of-vocabulary error in the training corpus. The source and the target languages highly influenced the methodology used to correct orthographic errors. As these languages vary in use of the same orthographic conventions very differently. These problems can be solved with different methods which basically depend upon the type and source of the problem, for example [44] came up with some solutions to overcome the errors in Catalan-Spanish language pairs

[1] https://en.wikipedia.org/wiki/Shuowen_Jiezi

such as the incorrect use of geminated l, the apostrophe, and the coordinating conjunctions *y* and *o*.

True-casing and Capitalization

The process of restoring case information to badly cased or not cased text is true-casing [45]. To avoid orthographical errors, it is a popular method to lower-case all words, especially in SMT. This method allows the system to avoid the mismatching of the same words, which seems different due to differences in casing thus keeping all the text in the lower-case is one of the methods to avoid the error. In most MT systems, both a pre-processing and post-processing is carried out. Post-processing of the text involves converting all the lower case to its original case form and generating the proper surface forms. These is done mostly in case of Latin and Slavic languages, where the same words with different case could me overgeneralised as different by the models for example the word *cat* and the word *CAT* could be put in different semantic category just because of the case. Therefore, to avoid such mistake True-casing is necessary.

Normalization

The use of the same words with different orthographic spellings such as *colour* and *color* give rise to different errors while building a translation model. In such cases, orthographic normalization is required. There are several other issues which require orthographic normalization, which could be language-specific such as Arabic diacritization, or contextual orthographic normalization. This approach needs some linguistic knowledge and can be adapted easily to other languages as well. Normalization is a process which is carried out before most of the natural language processing task; similarly, in machine translation, language-specific normalization yields a good result. Some of the examples of text normalization carried out for SMT system are removal of HTML contents, extraction of tag contents, splitting each line after proper punctuation marks as well as correction of language-specific word forms [46]. Normalization reduces sparsity as it eliminates out-of-vocabulary words used in the text [47].

Tokenization and Detokenization

The process of splitting text into smaller elements is known as tokenization. Tokenization can be done at different levels depending on the source and the target language as well the goal which we want to achieve. It also includes processing of the signs and symbols used in the text such as hyphens, apostrophes, punctuation marks, and numbers to make the text more accessible for further steps in MT. Like normalization, tokenization also helps in reducing language sparsity. The most commonly used words are assigned specific ids in sub-word tokenization technique, whereas less frequently used words are broken into sub-words that better reflect the context separately. If the word few appears regularly in the language, it will be given a special ID, while fewer and fewest, which are more unusual words that occur infrequently in the text, will be broken into sub words such as few, er, and est. This prevents the language model from misinterpreting less and fewest as two distinct terms. This helps the unknown terms in the data collection to be identified during preparation.

Detokenization is the process of combining all the token to the correct form before processing the main output. Tokenization and detokenization are not linked directly to orthographic correction, rather, they are more about morphological linking and correction, especially towards morphological rich languages like Irish and Arabic [48]. Orthography plays a major role in tokenization and detokenizations as each orthography has different rules on how to tokenize and detokenize.

Transliteration

Transliteration is the conversion of the text from one orthography to another without any phonological changes. The best example of transliteration is named entities and generic words [49]. Data collected from social media are highly transliterated and contain errors, thus, using these data for building a machine translation system for resource-poor languages cause errors. One of the primary forms that have a high chance of transliteration is cognates. Cognates are words from different languages derived from the same root. The concept cognate in NLP approaches are the words with similar orthography for example *family* in English and *familia* in Spanish. In the conventional approaches to automatic cognate detection, words with similar meanings or forms are used as probable cognates. From such sets, the ones that reveal a high phonological, lexical and/or semantic similarity, are investigated to find true cognates. Therefore, cognates have a high chance of transliteration. Though machine translation has progressed a lot in recently, the method of dealing with transliteration problem has changed from a language-independent manner to cognates prediction when translating between closely related languages, transliteration of cognates would help to improve the result for under-resourced languages.

Code-Mixing

Code-mixing is a phenomenon which occurs commonly in most multilingual societies where the speaker or writer alternate between more than one languages in a sentence [50–53]. Most of the corpora for under-resourced languages came

from the publicly available parallel corpora which were created by voluntary annotators or aligned automatically. The translation of technical documents such as KDE, GNOME, and Ubuntu translations have code-mixed data since some of the technical terms may not be known to voluntary annotators for translation. Code-mixing in the OpenSubtitles corpus is due to bilingual and historical reasons of native speakers [51, 54]. Different combinations of languages may occur while code-mixing, for example, German-Italian and French-Italian in Switzerland, Hindi-Telugu in state of Telangana, India, Hokkien-Mandarin Chinese in Taiwan [55]. As a result of code-mixing of the script are also possible from a voluntary annotated corpus. This poses another challenge for MT

Orthographic Information in RBMT

RBMT was one of the first approaches to tackle translation from the input of the source text to target text without human assistance by means of collection of dictionaries, collection of linguistics rules and special programs based on these dictionaries and rules. It also depends on rules and linguistic resources, such as bilingual dictionaries, morphological analysers, and part-of-speech taggers. The rules dictate the syntactic knowledge while the linguistic resources deal with morphological, syntactic, and semantic information. Both of them are grounded in linguistic knowledge and generated by linguists [7, 10, 56, 57]. The strength of RBMT is that analysis can be done at both syntactic and semantic level. However, it requires a linguistic expert to write down all the rules that cover the language.

An open-source shallow-transfer MT engine for the Romance languages of Spain such as Spanish, Catalan and Galician was developed by Armentano-Oller et al. [58]. They were regeneration of existing non-open-source engines based on linguistic data. The post-generator in the system performs orthographical operations such as contraction and apostrophes to reduce the orthographical errors. Dictionaries were used for string transformation operations to the target language surface forms. Similarly, the translation between Spanish-Portugues used a post-generation module to performs orthographical transformations to improve the translation quality [59, 60].

Manually constructed list of orthographic transformation rules assist in detecting cognates by string matching [61]. Irish, Scottish and Gaelic belong to the Goidelic language family and share similar orthography and cognates. Scannell [62] developed ga2gd software which translates from Irish to Scottish Gaelic. In the context-sensitive syntactic rewriting submodule, the authors implemented transfer rules based on orthography, which are stored in a plain text. Then each rule is transformed into a finite-state recogniser for the input stream. This work also uses simple rule-based orthographic changes to find cognates by orthography.

A Czech to Polish translation system also followed the shallow-transfer method at the lexical stage. A set of collective transformation rules were used on a source language list to produce a target language list of cognates [63]. Another shallow-transfer MT system used frequent orthographic changes from Swedish to Danish to identify cognates and transfer rules are based on orthography [64]. A Turkmen to Turkish MT system [65, 66] uses the finite-state transformer to identify the cognate even though the orthography is different for these languages.

Orthographic Information in SMT

Statistical Machine Translation (SMT) [15, 16, 67–69] is one of the CBMT based systems. SMT systems assume that we have a set of example translations$(S^{(k)}, T^{(k)})$ for $k = 1 \ldots .n$, where $S^{(k)}$ is the kth source sentence, $T^{(k)}$ is the kth target sentence which is the translation of $S^{(k)}$ in the corpus. SMT systems try to maximize the conditional probability $p(t|s)$ of target sentence t given a source sentence s by maximizing separately a language model $p(t)$ and the inverse translation model $p(s|t)$. A language model assigns a probability $p(t)$ for any sentence t and translation model assigns a conditional probability $p(s|t)$ to source / target pair of sentence [70]. By Bayes rule

$$p(t|s) \propto p(t)p(s|t) \tag{1}$$

This decomposition into a translation and a language model improves the fluency of generated texts by making full use of available corpora. The language model is not only meant to ensure a fluent output, but also supports difficult decisions about word order and word translation [68].

The two core methodologies used in the development of machine translation systems—RBMT and SMT—come with their own shares of advantages and disadvantages. In the initial stages, RBMTs were the first commercial systems to be developed. These systems were based on linguistic rules and have proved to be more feasible for resource-poor languages with little or no data. It is also relatively simpler to carry out error analysis and work on improving the results. Moreover, these systems require very little computational resources.

On the contrary, SMT systems need a large amount of data, but no linguistic theories, and so especially with morphologically rich languages such as Irish, Persian, and Tamil, SMT suffers from out-of-vocabulary problems very frequently due to orthographic inconsistencies. To mitigate the problem, orthographic normalization was proposed to improve the quality of SMT by sparsity reduction [71]. SMT

learns from data and requires less human effort in terms of creating linguistics rules. SMT systems, unlike RBMT system, does not cause disambiguation problems. Even though SMT has lots of advantages over rule-based, it also has some disadvantages. Its is very difficult to conduct error analysis with SMT and data sparsity is another disadvantage faced by SMT [72].

Spelling and Typographical Errors

The impact of spelling and typographical errors in SMT has been studied extensively [73–75]. Dealing with random, non-word error or real-word error can be done in many ways; one such method is the use of a character-level translator, which provides various spelling alternatives. Typographical errors such as substitution, insertion, deletion, transposition, run-on, and split can be addressed with edit-distance under a noisy channel model paradigm [76, 77]. Error recovery was performed to correct spelling alternatives in the input before the translation process.

True-casing and Capitalization, Tokenization and Detokenization

Most SMT systems accept pre-processed inputs, where the pre-processing consists of tokenising, true-casing, and normalising punctuation. Moses [16] is a toolkit for SMT, which has pre-processing tools for most languages based on hand-crafted rules. Improvement has been achieved for recasing and tokenization processes [78]. For a language which does not use Roman characters, linguistically-motivated tokenization has shown to improve the results on SMT [79]. Byte Pair Encoding (BPE) avoids out-of-vocabulary issues by representing more frequent sub-word as atomic units Sennrich et al. [80]. A joint BPE model based on the lexical similarity between Czech and Polish identified cognate vocabulary of sub-words. This is based on the orthographic correspondences from which words in both languages can be composed [81].

Normalization

Under-resourced languages utilise corpora from the user-generated text, media text or voluntary annotators. However, SMT suffers from customisation problems as tremendous effort is required to adapt to the style of the text. A solution to this is text normalization, that is normalising the corpora before passing it to SMT [75] which has been shown to improve the results. The orthographies of the Irish and Scottish Gaelic languages were quite similar due to a shared literary tradition. Nevertheless, after the spelling reform in Irish, the orthography became different. Scannell [82] proposed a statistical method to normalise the orthography between Scottish Gaelic and Irish as part of the translation for social media text. To able to use the current NLP tool to deal with historical text, spelling normalization is essential; that is converting the original spelling to present-day spelling which was studied for historical English text by Schneider et al. [83] and Hämäläinen et al. [84]. For dialects translation, spelling normalising is an important step to take advantage of high-resource languages resources [85, 86]

Transliteration (Cognate)

As we know, closely related languages share the same features; the similarities between the language would be of much help to study the cognates of two languages. Cognates can also exist in the same language and different language families. Several methods have been obtained to manipulate the features of resource-rich languages to improve SMT for resource-poor languages. Manipulation of the cognates to obtain transliteration is one of the methods adopted by some of the authors to improve the SMT system for resource-poor languages.

Language similarities and regularities in morphology and spelling variation motivate the use of character-level transliteration models. However, to avoid the character mapping differences in various contexts Nakov and Tiedemann [87] transformed the input to a sequence of character *n*-grams. A sequence character of *n*-grams increases the vocabulary as well as also make the standard alignment models and their lexical translation parameters more expressive.

For the languages which use same or similar scripts, approximate string matching approaches, like Levenshtein distance [88] are used to find cognate and longest common subsequence ratio (LCSR) [89]. For the languages which use different scripts, transliteration is the first step and follow the above approach. A number of studies have used statistical and deep learning methods along with orthographic information [90, 91] to find the cognates. In reference to the previous section we know that cognates can be used for mutual translation between two languages if they share similar properties, it is essential to know the cognateness between the two languages of a given text. The word "cognateness" means how much two pieces of text are related in terms of cognates. These cognates were useful to improve the alignment when the scoring function of the length-based alignment function is very low then it passes to the second method, a cognate alignment function for getting a proper alignment result [92].

One of the applications of cognates before applying MT is parallel corpora alignment. A study of using cognates to align sentences for parallel corpora was done by Simard et al. [93]. Character level methods to align sentences [94] are based on a cognate approach [93].

As early as Bemova et al. [95], researchers have looked into translation between closely-related languages such as

from Czech-Russian RUSLAN and Czech-Slovak CESILKO [96] using syntactic rules and lexicons. The closeness of the related languages makes it possible to obtain a good translation by means of more straightforward methods. However, both systems were rule-based approaches and bottlenecks included complexities associated with using a word-for-word dictionary translation approach. Nakov and Ng [97] proposed a method to use resource-rich closely-related languages to improve the statistical machine translation of under-resourced languages by merging parallel corpora and combining phrase tables. The authors developed a transliteration system trained on automatically-extracted likely cognates for Portuguese into Spanish using systematic spelling variation.

Popović and Ljubešić [98] created an MT system between closely-related languages for the Slavic language family. Language-related issues between Croatian, Serbian and Slovenian are explained by Popović et al. [99]. Serbian is digraphic (uses both Cyrillic and Latin Script), the other two are written using only the Latin script. For the Serbian language transliteration without loss of information is possible from Latin to Cyrillic script because there is a one-to-one correspondence between the characters.

In 2013 a group of researchers used a PBSMT approach as the base method to produce cognates. Instead of translating the phrase, they tried to transform a character sequence from one language to another. They have used words instead of sentences and characters instead of words in the transformation process. The combination of the phrase table with transformation probabilities, language model probabilities, selects the best combination of sequence. Thus the process includes the surrounding context and produces cognates [100]. A joint BPE model based on the lexical similarity between Czech and Polish identifies a cognate vocabulary of sub-words. This is based on the orthographic correspondences from which words in both languages can be composed [81]. It has been demonstrated that the use of cognates improves the translation quality [17].

Code-Switching

An SMT system with a code-switched parallel corpus was studied by Menacer et al. [101] and Fadaee and Monz [102] for Arabic–English language pair. The authors have manually translated or used back translation method to translate foreign words. The identification of the language of the word is based on the orthography. Chakravarthi et al. [103] used the same approach for Dravidian languages; they used the improved MT for creating WordNet, showing improvement in the results. For English–Hindi, Dhar et al. [104] manually translated the code-switched component and shown improvements. Machine translation of social media was studied by Rijhwani et al. [105] where they tackle the code-mixing for Hindi–English and Spanish–English. The same approach translated the main language of the sentence using Bing Translate API [106].

Back transliteration from one script to native script in code-mixed data is one of the challenging tasks to be performed. Riyadh and Kondrak [107] adopted three different methods to back transliterate Romanised Hindi-Bangla code-mixed data to Hindi and Bangla script. They have used Sequitur, a generative joint *n*-gram transducer, DTLM, a discriminate string transducer and the OpenNMT[2] neural machine translation toolkit. Along with these three approaches, they have leveraged target word lists, character language models, as well as synthetic training data, whenever possible, to support transliteration. Finally, these transliterations are provided to a sequence prediction module for further processing.

Pivot Translation

Pivot translation is a translation from a source language to the target language through an intermediate language which is called a pivot language. Usually, pivot language translation has large source-pivot and pivot-target parallel corpora [25, 108]. There are different levels of pivot translation, the first one is the triangulation method where the corresponding translation probabilities and lexical weights in the source-pivot and pivot-target translation are multiplied. In the second method, the sentences are translated to the pivot language using the source-pivot translation system then pivoted to target language using a pivot-target translation system [109]. Finally, using the source-target MT system to create more data and adding it back to the source-target model, which is called back-translation [80, 110]. Back-translation is simple and easy to achieve without modifying the architecture of the machine translation models. Back-translation has been studied in both SMT [111–113] and NMT [23, 80, 110, 114–116].

The pivot translation method could also be used to improve MT systems for under-resourced languages. One popular way is training SMT systems using source-pivot or pivot-target language pair using sub words where the pivot language is related to source or target or both. The subwords units consisted of orthographic syllable and byte-pair-encoded unit. The orthographic unit is a linguistically motivated unit which occurs in a sequence of one or more consonants followed by a vowel. Unlike orthographic units, BPE (Byte Pair Encoded Unit) [80] is motivated by statistical properties of the text. It represents stable and frequent character sequences in the texts. As orthographic syllable and BPE are variable-length units and the vocabularies used

[2] https://opennmt.net/

are much smaller than morpheme and word-level model, the problem of data sparsity does not occur but provides an appropriate context for translation between closely related languages [117].

Orthographic Information in NMT

Neural Machine Translation is a sequence-to-sequence approach [21] based on encoder-decoder architectures with attention [19, 118] or self attention encoder [119, 120]. Given a source sentence $\mathbf{x} = x_1, x_2, x_3, ...$ and target sentence $\mathbf{y} = y_1, y_2, y_3, ..$, the training objective for NMT is to maximize the log-likelihood $\mathcal{L}$ with respect to θ:

$$\mathcal{L}_\theta = \sum_{(\mathbf{x},\mathbf{y})\in \mathrm{C}} \log p(\mathbf{y}|\mathbf{x};\theta) \tag{2}$$

The decoder produces one target word at a time by computing the probability

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{m} p\left(y_j|y_{<j},\mathbf{x};\theta\right) \tag{3}$$

where m is the number of words in $\mathbf{y}$, y_j is the current generated word, and $y_{<j}$ are the previously generated words. At inference time, beam search is typically used to find the translation that maximises the above probability. Most of NMT models follows the *Embedding* → *Encoder* → *Attention* → *Decoder* framework.

The attention mechanism across encoder and decoder is calculated by c_t as the weighted sum of the source-side context vectors:

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \tag{4}$$

$$\alpha_{t,i} = \frac{\exp\left(e_{t,i}\right)}{\sum_{j=1}^{m} \exp\left(e_{t,j}\right)} \tag{5}$$

$\alpha_{t,i}$ is the normalized alignment matrix between each source annotation vector h_i and word y_t to be emitted at a time step t. Expected alignment $e_{t,i}$ between each source annotation vector h_i and the target word y_t is computed using the following formula:

$$e_{t,i} = a(\mathbf{s_{t-1}}, h_i) \tag{6}$$

$$\mathbf{s_t} = g\left(\mathbf{s_{t-1}}, \mathbf{y_{t-1}}, \mathbf{c_t}\right) \tag{7}$$

where g is an activation decoder function, s_{j-1} is the previous decoder hidden state, y_{j-1} is the embedding of the previous word. The current decoder hidden state s_j, the previous word embedding and the context vector are fed to a feedforward layer f and a softmax layer computes a score for generating a target word as output:

$$P\left(y_j|y_{<j},\mathbf{x}\right) = \mathrm{softmax}\left(f\left(\mathbf{s_j}, \mathbf{y}_{j-1}, \mathbf{c_j}\right)\right)$$

Multilingual Neural Machine Translation

In recent years, NMT has improved translation performance, which has lead to a boom in NMT research. The most popular neural architectures for NMT are based on the encoder-decoder [19, 21, 121] structure and the use of attention or self-attention based mechanism [119, 122]. Multilingual NMT created with or without multiway corpora has been studied for the potential for translation between two languages without any direct parallel corpus. Zero-shot translation is translation using multilingual data to create a translation for languages which have no direct parallel corpora to train independently. Multilingual Neural Machine Translation with only monolingual corpora was studied by [123, 124]. In Ha et al. [125] and [28], the authors have demonstrated that multilingual NMT improves translation quality. For this, they created a multilingual NMT without changing the architecture by introducing special tokens at the beginning of the source sentence indicating the source language and target language.

Phonetic transcription to Latin script and the International Phonetic Alphabet (IPA) was studied by Chakravarthi et al. [103] and showed that Latin script outperforms IPA for the Multilingual NMT of Dravidian languages. Chakravarthi et al. [126] propose to combine multilingual, phonetic transcription and multimodal content with improving the translation quality of under-resourced Dravidian languages. The authors studied how to use the closely-related languages from the Dravidian language family to exploit the similar syntax and semantic structures by phonetic transcription of the corpora into Latin script along with image feature to improve the translation quality [127]. They showed that orthographic information improves the translation quality in multilingual NMT [128].

Spelling and Typographical Errors

Spelling errors are amplified in under-resourced setting due to the potential infinite possible misspelling and leads to a large number of out-of-vocabulary words. Additionally, under-resourced morphological rich languages have morphological variation, which causes orthographic errors while using character level MT. A shared task was organised by Li et al. [129]; to deal with orthographic variation, grammatical

error and informal languages from the noisy social media text. Data cleaning was used along with suitable corpora to handle spelling errors. Belinkov and Bisk [130] investigated noise in NMT, focusing on kinds of orthographic errors. Parallel corpora were cleaned before submitting to NMT to reduce the spelling and typographical errors.

NMT with word embedding lookup ignores the orthographic representation of the words such as the presence of stems, prefixes, suffixes and another kind of affixes. To overcome these drawbacks, character-based word embedding was proposed by Kim et al. [131]. Character-based NMT [132–135] were developed to cover the disadvantages of the languages which do not have explicit word segmentation. This enhances the relationship between the orthography of a word and its meaning in the translation system. For spelling mistake data for under-resourced languages, the quality of word-based translation drops severely, because every non-canonical form of the word cannot be represented. Character-level model overcomes the spelling and typological error without much effort.

True-casing and Capitalization, Normalization, Tokenization and Detokenization

Although NMT can be trained end-to-end translations, many NMT systems are still language-specific and require language-dependent preprocessing, such as used in Statistical Machine Translation, Moses [16] a toolkit for SMT which has preprocessing tools for most languages which based on hand-crafted rules. In fact, these are mainly available for European languages. For Asian languages which do not use space between words, a segmenter is required for each language independently before feeding into NMT to indicate a word segment. This becomes a problem when we train Multilingual NMT [28].

A solution for the open vocabulary problems in NMT is to break up the rare words into subword units [136, 137] which has been shown to deal with multiple script languages ambiguities [138, 139]. A simple and language-independent tokenizer was introduced for NMT and Multilingual NMT by Kudo and Richardson [140]; it is based on two subword segmentation algorithms, byte-pair encoding (BPE) [80] and a unigram language model [141]. This system also normalise semantically equivalent Unicode character into canonical forms. Subword segmentation and true-casing model will be rebuilt whenever the training data changes. The preprocessing tools introduced by OpenNMT normalises characters and separates punctuation from words, and it can be used for any languages and any orthography [142].

Character-level NMT systems work at the character level to grasp orthographic similarity between the languages. They were developed to overcome the issue of limited parallel corpora and resolve the out-of-vocabulary problem for the under-resourced languages. For Hindi–Bhojpuri, where Bhojpuri is closely related to Hindi, Bhojpuri is considered as an under-resourced language, and it has an overlap of word with high-resource language Hindi due to the adoption of works from a common properties of two languages [143]. To solve the out-of-vocabulary problem the transduction of Hindi word to Bhojpuri words was adapted from NMT models by training on Hindi–Bhojpuri cognate pairs. It was a two-level system: first, the Hindi–Bhojpuri system was developed to translate the sentence; then the out-of-vocabulary words were transduced.

Transliteration (Cognate)

Transliteration emerged to deal with proper nouns and technical terms that are translated with preserved pronunciation. Transliteration can also be used to improve machine translation between closely related languages, which uses different scripts since closely related languages language have orthographic and phonological similarities between them.

Machine Translation often occurs between closely related languages or through a pivot language (like English) [144]. Translation between closely related languages or dialects is either a simple transliteration from one language to another language or a post-processing step. Transliterating cognates has been shown to improve MT results since closely related languages share linguistic features. To translate from English to Finnish and Estonian, where the words have similar orthography Grönroos et al. [145] used Cognate Morfessor, a multilingual variant of Morfessor which learns to model cognates pairs based on the unweighted Levenshtein distance [88]. The ideas are to improve the consistency of morphological segmentation of words that have similar orthography, which shows improvement in the translation quality for the resource-poor Estonian language.

Cherry and Suzuki [146] use transliteration as a method to handle out-of-vocabulary (OOV) problems. To remove the script barrier, Bhat et al. [147] created machine transliteration models for the common orthographic representation of Hindi and Urdu text. The authors have transliterated text in both directions between Devanagari script (used to write the Hindi language) and Perso-Arabic script (used to write the Urdu language). The authors have demonstrated that a dependency parser trained on augmented resources performs better than individual resources. The authors have shown that there was a significant improvement in BLEU (Bilingual Evaluation Understudy) [148] score and have shown that the problem of data sparsity is reduced.

Recent work by Kunchukuttan et al. [149] has explored orthographic similarity for transliteration. In their work, they have used related languages which share similar writing systems and phonetic properties such as Indo-Aryan languages. They have shown that multilingual transliteration leveraging

similar orthography outperforms bilingual transliteration in different scenarios. Phonetic transcription is a method for writing a language in the other scripts keeping the phonemic units intact. It is extensively used in speech processing research, text-to-speech, and speech database construction—phonetic transcription to a common script has shown to improve the results of machine translation [103]. The authors focus on the multilingual translation of languages which uses different scripts and studies the effect of different orthographies to common script with multilingual NMT. Multiway NMT system was created for Czech and Polish with Czech IPA transcription and Polish transcription to a 3-way parallel text together to take advantage of the phonology of the closely related languages [81]. Orthographic correspondence rules were used as a replacement list for translation between closely related Czech-Polish with added back-translated corpus [81]. Dialect translation was studied by Baniata et al. [150]. To translate Arabic dialects to Modern Standard Arabic, they used multitask learning which shares one decoder for standard Arabic, while every source has a separate encoder. This is due to the non-standard orthography in the Arabic dialects. The experiments showed that for the under-resourced Arabic dialects, it improved the results.

Machine Translation of named entities is a significant issue due to linguistic and algorithmic challenges found in between languages. The quality of MT of named entities, including the technical terms, was improved with the help of developing lexicons using orthographic information. The lexicon integration to NMT was studied for the Japanese and Chinese MT [151]. They deal with the orthographic variation of named entities of Japanese using large scale lexicons. For English-to-Japanese, English-to-Bulgarian, and English-to-Romanian Ugawa et al. [152] proposed a model that encodes the input word based on its NE tag at each time step. This helps to improve the BLEU score for machine translation results.

Code-Switching

A significant part of corpora for under-resourced languages comes from movie subtitles and technical documents, which makes it even more prone to code-mixing. Most of these corpora are movie speeches [153] transcribed to text, and they differ from those in other written genres: the vocabulary is informal, non-linguistics sounds like *ah*, and mixes of scripts in case of English and native languages [154–159]. Data augmentation [160, 161] and changing the foreign to native words using dictionaries or other methods have been studied. Removing the code-mixing word from the corpus on both sides was studied by Chakravarthi et al. [103, 127] for English–Dravidian languages. Song et al. [162] studied the data augmentation method, making code-switched training data by replacing source phrases with their target translation. Character-based NMT [133–135] can naturally handle intra-sentence codeswitching as a result of the many-to-one translation task.

Orthographic Information in Unsupervised Machine Translation

Building parallel corpora for the under-resourced languages is time-consuming and expensive. As a result parallel corpora for the under-resourced languages are limited or unavailable for some of the languages. With limited parallel corpora, supervised SMT and NMT cannot achieve the desired quality translations. However, monolingual corpora can be collected from various sources on the Internet, and are much easier to obtain than parallel corpora. Recent research has created a machine translation system using only monolingual corpora [163–165] by the unsupervised method to remove the dependency of sentence aligned parallel corpora. These systems are based on both SMT [166, 167] and NMT [168]. One such task is bilingual lexicon induction.

Bilingual lexicon induction is a task of creating word translation from monolingual corpora in two languages [169, 170]. One way to induce the bilingual lexicon induction is using orthographic similarity. Based on the assumptions that words that are spelled similarly are sometimes good translation and maybe cognates as they have similar orthography due to historical reasons. A generative model for inducing a bilingual lexicon from monolingual corpora by exploiting orthographic and contextual similarities of words in two different languages was proposed by Haghighi et al. [171]. Many methods, based on edit-distance and orthographic similarity are proposed for using linguist feature for word alignments supervised and unsupervised methods [172–174]. Riley and Gildea [175] proposed method to utilise the orthographic information in word-embedding based bilingual lexicon induction. The authors used the alphabets of two languages to extend the word embedding and modifying the similarity score functions of previous word-embedding methods to include the orthographic similarity measure. Bilingual lexicons are shown to improve machine translation in both RBMT [170] and CBMT [163, 176, 177].

In work by Bloodgood and Strauss [178], the authors translated lexicon induction for a heavily code-switched text of historically unwritten colloquial words via loanwords using expert knowledge with language information. Their method is to take word pronunciation (IPA) from a donor language and convert them into the borrowing language. This shows improvements in BLEU score for induction of Moroccan Darija-English translation lexicon bridging via French loan words.

Discussion

From our comprehensive survey, we can see that orthographic information improves translation quality in all types of machine translation from rule-based to completely unsupervised systems like bilingual lexicon induction. For RBMT, translation between closely related languages is simplified to transliteration due to the cognates. Statistical machine translation deals with data sparsity problem using orthographic information. Since statistical machine translation has been studied for a long time, most of the orthographic properties are studies for different types of languages. Even recent neural machine translation and other methods still use preprocessing tools such as truecasers, tokenizers, and detokenizers that are developed for statistical machine translation. Recent neural machine translation is completely end-to-end, however, it suffers from data sparsity when dealing with morphologically rich languages or under-resourced languages. These issues are dealt by utilising orthographic information in neural machine translation. One such method which improves the translation is a transliteration of cognates. Code-switching is another issue with under-resourced languages due to the data collected from voluntary annotator, web crawling or other such methods. However, dealing with code-switching based on orthography or using character-based neural machine translation has been shown to improve the results significantly.

From this, we conclude that orthographic information is much utilised while translating between closely related languages or using multilingual neural machine translation with closely related languages. While exciting advances have been made in machine translation in recent years, there is still an exciting direction for exploration from leveraging linguistic information to it, such as orthographic information. One such area is unsupervised machine translation or bilingual lexicon induction. Recent works show that word vector, along with orthographic information, performs better for aligning the bilingual lexicons in completely unsupervised or semi-supervised approaches. We believe that our survey will help to catalogue future research papers and better understand the orthographic information to improve machine translation results.

Conclusion

In this work, we presented a review of the current state-of-the-art in machine translation utilising orthographic information, covering rule-based machine translation, statistical machine translation, neural machine translation and unsupervised machine translation. As a part of this survey, we introduced different machine translations methods and have shown how orthography played a role in machine translation results. These methods to utilise the orthographic information have already let to a significant improvement in machine translation results.

Acknowledgements This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight), SFI/12/RC/2289_P2 (Insight_2), & SFI/18/CRT/6223 (CRT-Centre for Research Training in Artficial Intelligence) co-funded by the European Regional Development Fund as well as by the EU H2020 programme under grant agreements 731015 (ELEXIS-European Lexical Infrastructure), 825182 (Prêt-à-LLOD), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

Funding Open Access funding provided by the IReL Consortium.

Declarations

Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Karakanta A, Dehdari J, van Genabith J. Neural machine translation for low-resource languages without parallel corpora. Machans. 2018;32(1):167–89. https://doi.org/10.1007/s10590-017-9203-5.
2. Lewis W, Munro R, Vogel S. Crisis MT: Developing a cookbook for MT in crisis situations. In: Proceedings of the sixth workshop on statistical machine translation. Association for computational linguistics, Edinburgh, Scotland; 2011. p. 501–511. https://www.aclweb.org/anthology/W11-2164.
3. Neubig G, Hu J. Rapid adaptation of neural machine translation to new languages. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for computational linguistics, Brussels, Belgium; 2018;. p. 75–880. https://doi.org/10.18653/v1/D18-1103. https://www.aclweb.org/anthology/D18-1103
4. Abercrombie G. A rule-based shallow-transfer machine translation system for Scots and English. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16), European Language Resources Association (ELRA),

Portorož, Slovenia, 2016. p. 578–584. https://www.aclweb.org/anthology/L16-1092
5. Allauzen C, Byrne B, Gispert A, Iglesias G, Riley M. Pushdown automata in statistical machine translation. Comput Linguistics. 2014;40(3):687–723. https://doi.org/10.1162/COLI_a_00197. https://www.aclweb.org/anthology/J14-3008
6. Centelles J, Costa-jussà MR. Chinese-to-Spanish rule-based machine translation system. In: Proceedings of the 3rd workshop on hybrid approaches to machine translation (HyTra), association for computational linguistics, Gothenburg, Sweden; 2014. p. 82–86. https://doi.org/10.3115/v1/W14-1015. https://www.aclweb.org/anthology/W14-1015
7. Charoenpornsawat P, Sornlertlamvanich V, Charoenporn T. Improving translation quality of rule-based machine translation. In: COLING-02: machine translation in Asia; 2002. https://www.aclweb.org/anthology/W02-1605
8. Hurskainen A, Tiedemann J. Rule-based machine translation from English to Finnish. In: Proceedings of the second conference on machine translation. Association for computational linguistics, Copenhagen, Denmark; 2017. p. 323–329. https://doi.org/10.18653/v1/W17-4731. https://www.aclweb.org/anthology/W17-4731
9. Kaji H. An efficient execution method for rule-based machine translation. In: Coling Budapest 1988 volume 2: international conference on computational linguistics; 1988. https://www.aclweb.org/anthology/C88-2167.
10. Susanto RH, Larasati SD, Tyers FM. Rule-based machine translation between Indonesian and Malaysian. In: Proceedings of the 3rd workshop on South and Southeast Asian natural language processing. The COLING 2012 Organizing Committee, Mumbai, India; 2012. p. 191–200. https://www.aclweb.org/anthology/W12-5017
11. Carl M. A model of competence for corpus-based machine translation. In: COLING 2000 volume 2: the 18th international conference on computational linguistics; 2000. https://www.aclweb.org/anthology/C00-2145
12. Dauphin E, Lux V. Corpus-based annotated test set for machine translation evaluation by an industrial user. In: COLING 1996 volume 2: the 16th international conference on computational linguistics; 1996. https://www.aclweb.org/anthology/C96-2188
13. Green S, Cer D, Manning C. An empirical comparison of features and tuning for phrase-based machine translation. In: Proceedings of the ninth workshop on statistical machine translation, association for computational linguistics, Baltimore, Maryland, USA; 2014. p. 466–476, https://doi.org/10.3115/v1/W14-3360. https://www.aclweb.org/anthology/W14-3360
14. Junczys-Dowmunt M, Grundkiewicz R. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Association for computational linguistics, Austin, Texas; 2016. p. 1546–1556,.https://doi.org/10.18653/v1/D16-1161. https://www.aclweb.org/anthology/D16-1161
15. Koehn P. Europarl: a parallel corpus for statistical machine translation. In: Conference proceedings: the tenth machine translation summit, AAMT; 2005.
16. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for computational linguistics; 2007. p. 177–180.
17. Kondrak G, Marcu D, Knight K. Cognates can improve statistical translation models. In: Companion volume of the proceedings of HLT-NAACL 2003—short papers; 2003. p. 46–48. https://www.aclweb.org/anthology/N03-2016
18. Setiawan H, Li H, Zhang M, Ooi BC. Phrase-based statistical machine translation: a level of detail approach. In: Dale R, Wong KF, Su J, Kwong OY, editors. Natural language processing-IJCNLP 2005. Berlin Heidelberg: Springer; 2005. p. 576–87.
19. Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. In: 3rd international conference on learning representations, ICLR; 2015.
20. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar; 2014. p. 1724–1734. https://doi.org/10.3115/v1/D14-1179. https://www.aclweb.org/anthology/D14-1179
21. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proceedings of the 27th international conference on neural information processing systems - volume 2. MIT Press, Cambridge, MA, USA, NIPS'14; 2014. p. 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173
22. Zhang J, Wang M, Liu Q, Zhou J. Incorporating word reordering knowledge into attention-based neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). Association for computational linguistics, Vancouver, Canada; 2017. p. 1524–1534. https://doi.org/10.18653/v1/P17-1140. https://www.aclweb.org/anthology/P17-1140
23. Kim Y, Petrov P, Petrushkov P, Khadivi S, Ney H. Pivot-based transfer learning for neural machine translation between non-English languages. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for computational linguistics, Hong Kong, China; 2019. p. 866–876. https://doi.org/10.18653/v1/D19-1080. https://www.aclweb.org/anthology/D19-1080
24. Wu H, Wang H. Pivot language approach for phrase-based statistical machine translation. In: Proceedings of the 45th annual meeting of the association of computational linguistics, Prague, Czech Republic; 2007. p. 856–863. https://www.aclweb.org/anthology/P07-1108
25. Wu H, Wang H. Revisiting pivot language approach for machine translation. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. Association for computational linguistics, Suntec, Singapore; 2009. p. 154–162. https://www.aclweb.org/anthology/P09-1018
26. Currey A, Heafield K. Zero-resource neural machine translation with monolingual pivot data. In: Proceedings of the 3rd workshop on neural generation and translation, Association for computational linguistics, Hong Kong; 2019. p. 99–107. https://doi.org/10.18653/v1/D19-5610. https://www.aclweb.org/anthology/D19-5610
27. Gu J, Wang Y, Cho K, Li VO. Improved zero-shot neural machine translation via ignoring spurious correlations. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 1258–1268. https://doi.org/10.18653/v1/P19-1121. https://www.aclweb.org/anthology/P19-1121
28. Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G, Hughes M, Dean J. Google's multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguistics. 2017;5:339–51. https://doi.org/10.1162/tacl_a_00065. https://www.aclweb.org/anthology/Q17-1024
29. Pham NQ, Niehues J, Ha TL, Waibel A. Improving zero-shot translation with language-independent constraints. In: Proceedings of the fourth conference on machine translation (volume

1: research papers). Association for computational linguistics, Florence, Italy; 2019. p. 13–23. https://doi.org/10.18653/v1/W19-5202. https://www.aclweb.org/anthology/W19-5202
30. Tan X, Chen J, He D, Xia Y, Qin T, Liu TY. Multilingual neural machine translation with language clustering. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China; 2019. p. 963–973. https://doi.org/10.18653/v1/D19-1089. https://www.aclweb.org/anthology/D19-1089.
31. Artetxe M, Labaka G, Agirre E. Bilingual lexicon induction through unsupervised machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 5002–5007. https://doi.org/10.18653/v1/P19-1494. https://www.aclweb.org/anthology/P19-1494
32. Artetxe M, Labaka G, Agirre E. An effective approach to unsupervised machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 194–203. https://doi.org/10.18653/v1/P19-1019. https://www.aclweb.org/anthology/P19-1019
33. Pourdamghani N, Aldarrab N, Ghazvininejad M, Knight K, May J. Translating translationese: a two-step approach to unsupervised machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 3057–3062, https://doi.org/10.18653/v1/P19-1293. https://www.aclweb.org/anthology/P19-1293
34. Abney S, Bird S. The Human Language Project: building a universal corpus of the world's languages. In: Proceedings of the 48th annual meeting of the association for computational linguistics; 2010. p. 88–97. http://www.aclweb.org/anthology/P10-1010
35. Hauksdóttir A. An innovative world language centre : challenges for the use of language technology. In: Proceedings of the ninth international conference on language resources and evaluation (LREC-2014). European Language Resources Association (ELRA); 2014. http://www.aclweb.org/anthology/L14-1618
36. Alegria I, Artola X, De Ilarraza AD, Sarasola K. Strategies to develop language technologies for less-resourced languages based on the case of Basque; 2011.
37. Krauwer S. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Proc SPECOM. 2003;2003:8–15.
38. Maxwell M, Hughes B. Frontiers in linguistic annotation for lower-density languages. In: Proceedings of the workshop on frontiers in linguistically annotated Corpora 2006. Association for computational linguistics; 2006. p. 29–37. http://www.aclweb.org/anthology/W06-0605
39. Jimerson R, Prud'hommeaux E (2018) ASR for documenting acutely under-resourced indigenous languages. In: Chair NCC, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T, editors. Proceedings of the eleventh international conference on language resources and evaluation (LREC). European Language Resources Association (ELRA), Japan, Miyazaki; 2018.
40. Fromkin V, Rodman R, Hyams N. An introduction to language. Boston: Cengage Learning; 2018.
41. Fischer A, Jágrová K, Stenger I, Avgustinova T, Klakow D, Marti R. Orthographic and morphological correspondences between related slavic languages as a base for modeling of mutual intelligibility. In: Proceedings of the tenth international conference on language resources and evaluation (LREC'16); 2016. p. 4202–4209.
42. Min Z, Haizhou L, Jian S. Direct orthographical mapping for machine transliteration. In: Proceedings of the 20th international conference on computational linguistics. Association for computational linguistics; 2004. p. 716.
43. Kunchukuttan A, Khapra M, Singh G, Bhattacharyya P. Leveraging orthographic similarity for multilingual neural transliteration. Trans Assoc Comput Linguistics 2018;6:303–16. https://doi.org/10.1162/tacl_a_00022. https://www.aclweb.org/anthology/Q18-1022
44. Farrús M, Costa-Jussa MR, Marino JB, Poch M, Hernández A, Henríquez C, Fonollosa JA. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the catalan-spanish language pair. Language Resour Eval. 2011;45(2):181–208.
45. Lita LV, Ittycheriah A, Roukos S, Kambhatla N. Truecasing. In: Proceedings of the 41st annual meeting on association for computational linguistics-volume 1. Association for computational linguistics; 2003. p. 152–159.
46. Schlippe T, Zhu C, Gebhardt J, Schultz T. Text normalization based on statistical machine translation and internet user support. In: Eleventh annual conference of the international speech communication association; 2010.
47. Leusch G, Ueffing N, Vilar D, Ney H. Preprocessing and normalization for automatic evaluation of machine translation. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Association for Computational Linguistics, Ann Arbor, Michigan; 2005. p. 17–24. https://www.aclweb.org/anthology/W05-0903
48. Guzmán F, Bouamor H, Baly R, Habash N. Machine translation evaluation for Arabic using morphologically-enriched embeddings. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers. The COLING 2016 Organizing Committee, Osaka, Japan; 2016. p. 1398–1408. https://www.aclweb.org/anthology/C16-1132
49. Kumaran A, Kellner T. A generic framework for machine transliteration. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2007. p. 721–722.
50. Ayeomoni MO. Code-switching and code-mixing: Style of language use in childhood in Yoruba speech community. Nordic J Afr Stud. 2006;15(1):90–9.
51. Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N. What is India speaking? Exploring the "Hinglish" invasion. Phys A Stat Mech Appl. 2016;449:375–89. https://doi.org/10.1016/j.physa.2016.01.015. http://www.sciencedirect.com/science/article/pii/S0378437116000236
52. Ranjan P, Raja B, Priyadharshini R, Balabantaray RC. A comparative study on code-mixed data of Indian social media vs formal text. In: 2nd international conference on contemporary computing and informatics (IC3I), IEEE; 2016. p. 608–611. https://ieeexplore.ieee.org/document/7918035
53. Yoder MM, Rijhwani S, Rosé CP, Levin L. Code-switching as a social act: the case of Arabic Wikipedia talk pages. ACL. 2017;2017:73.
54. Chanda A, Das D, Mazumdar C. Columbia-Jadavpur submission for emnlp 2016 code-switching workshop shared task: system description. EMNLP. 2016;2016:112.
55. Chan JYC, Cao H, Ching PC, Lee T. Automatic recognition of Cantonese-English code-mixing speech. Int J Comput Linguistics Chin Language Process. 2009;14(3). https://www.aclweb.org/anthology/O09-5003
56. Lagarda AL, Alabau V, Casacuberta F, Silva R, Díaz-de Liaño E. Statistical post-editing of a rule-based machine translation system. In: Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics, companion volume: short papers. Association for Computational Linguistics, Stroudsburg,

PA, USA, NAACL-Short '09; 2009. p. 217–220. http://dl.acm.org/citation.cfm?id=1620853.1620913
57. Slocum J, Bennett WS, Whiffin L, Norcross E. An evaluation of metal: the lrc machine translation system. In: Proceedings of the second conference on European chapter of the association for computational linguistics; 1985. p. 62–69.
58. Armentano-Oller C, Carrasco RC, Corbí-Bellot AM, Forcada ML, Ginestí-Rosell M, Ortiz-Rojas S, Pérez-Ortiz JA, Ramírez-Sánchez G, Sánchez-Martínez F, Scalco MA. Open-source portuguese-spanish machine translation. In: Mamede NJ, Oliveira C, Dias MC, Vieira R, Quaresma P, Nunes MGV, editors. Computational processing of the Portuguese language. Berlin, Heidelberg: Springer; 2006. p. 50–9.
59. Forcada ML, Ginestí-Rosell M, Nordfalk J, O'Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, Sánchez-Martínez F, Ramírez-Sánchez G, Tyers FM. Apertium: a free/open-source platform for rule-based machine translation. Mach Transl. 2011;25(2):127–44.
60. Garrido-Alenda A, Gilabert-Zarco P, Pérez-Ortiz JA, Pertusa-Ibáñez A, Ramírez-Sánchez G, Sánchez-Martínez F, Scalco MA, Forcada ML. Shallow parsing for portuguese–spanish machine translation. In: Tagging and shallow processing of Portuguese: workshop notes of TASHA'2003, Citeseer; 2003. p. 21.
61. Xu Q, Chen A, Li C. Detecting English-French cognates using orthographic edit distance. In: Proceedings of the Australasian Language Technology Association Workshop 2015, Parramatta, Australia, 2015. p. 145–149. https://www.aclweb.org/anthology/U15-1020.
62. Scannell KP. Machine translation for closely related language pairs. In: Proceedings of the workshop strategies for developing machine translation for minority languages, Citeseer; 2006. p. 103–109.
63. Ruth J, O'Regan J. Shallow-transfer rule-based machine translation for Czech to Polish. In: Proceedings of the second international workshop on free/open-source rule-based machine translation, Universitat Oberta de Catalunya; 2011. p. 69–76.
64. Tyers FM, Nordfalk J, et al. Shallow-transfer rule-based machine translation for swedish to danish. In: Proceedings of the first international workshop on free/open-source rule-based machine translation. Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos; 2009. p. 27–33.
65. Tantuğ AC, Adalı E. Machine translation between Turkic languages. In: Saraçlar M, Oflazer K. editors. Turkish natural language processing. Springer; 2018. p. 237–254.
66. Tantuğ AC, Adalı E, Oflazer K.A MT system from Turkmen to Turkish employing finite state and statistical methods. In: Machine translation summit XI, European Association for Machine Translation (EAMT); 2007. p. 459–465.
67. Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Comput Linguistics 1993;19(2):263–311. https://www.aclweb.org/anthology/J93-2003
68. Koehn P. Statistical machine translation. 1st ed. New York, NY: Cambridge University Press; 2010.
69. Waite A, Byrne B. The geometry of statistical machine translation. In: Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. Association for computational linguistics, Denver, Colorado; 2015. p. 376–386. https://doi.org/10.3115/v1/N15-1041. https://www.aclweb.org/anthology/N15-1041
70. Wang YY, Waibel A. Decoding algorithm in statistical machine translation. In: 35th annual meeting of the association for computational linguistics and 8th conference of the European Chapter of the Association for Computational Linguistics. Association for computational linguistics, Madrid, Spain; 1997. p. 366–372. https://doi.org/10.3115/976909.979664. https://www.aclweb.org/anthology/P97-1047
71. El Kholy A, Habash N. Orthographic and morphological processing for english–arabic statistical machine translation. Mach Trans. 2012;26(1–2):25–45. https://doi.org/10.1007/s10590-011-9110-0.
72. Costa-Jussa MR, Farrús M, Marino JB, Fonollosa JA. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. Comput Inf. 2012;31(2):245–70.
73. Bertoldi N, Zens R, Federico M, Shen W. Efficient speech translation through confusion network decoding. IEEE Trans Audio Speech Language Process. 2008;16(8):1696–705. https://doi.org/10.1109/TASL.2008.2002054.
74. Bertoldi N, Cettolo M, Federico M. Statistical machine translation of texts with misspelled words. In: Human language technologies: the 2010 annual conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California; 2010. p. 412–419. https://www.aclweb.org/anthology/N10-1064
75. Formiga L, Fonollosa JAR. Dealing with input noise in statistical machine translation. In: Proceedings of COLING 2012: posters, the COLING 2012 organizing committee, Mumbai, India; 2012. p. 319–328. https://www.aclweb.org/anthology/C12-2032
76. Brill E, Moore RC. An improved error model for noisy channel spelling correction. In: Proceedings of the 38th annual meeting of the association for computational linguistics, Hong Kong; 2000. p. 286–293. https://doi.org/10.3115/1075218.1075255. https://www.aclweb.org/anthology/P00-1037
77. Toutanova K, Moore R. Pronunciation modeling for improved spelling correction. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for computational linguistics, Philadelphia, Pennsylvania, USA; 2002. p. 144–151. https://doi.org/10.3115/1073083.1073109. https://www.aclweb.org/anthology/P02-1019
78. Nakov P. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In: Proceedings of the third workshop on statistical machine translation. Association for computational linguistics, Columbus, Ohio; 2008. p. 147–150. https://www.aclweb.org/anthology/W08-0320
79. Oudah M, Almahairi A, Habash N. The impact of preprocessing on Arabic-English statistical and neural machine translation. In: Proceedings of machine translation summit XVII volume 1: research track. European Association for Machine Translation, Dublin, Ireland; 2019. p. 214–221. https://www.aclweb.org/anthology/W19-6621
80. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers). Association for computational linguistics, Berlin, Germany; 2016. p. 86–96. https://doi.org/10.18653/v1/P16-1009. https://www.aclweb.org/anthology/P16-1009
81. Chen Y, Avgustinova T. Machine translation from an intercomprehension perspective. In: Proceedings of the fourth conference on machine translation (volume 3: shared task papers, day 2). Association for computational linguistics, Florence, Italy; 2019. p. 192–196. https://doi.org/10.18653/v1/W19-5425. https://www.aclweb.org/anthology/W19-5425
82. Scannell K. Statistical models for text normalization and machine translation. In: Proceedings of the first Celtic language technology workshop. Association for computational linguistics and Dublin City University, Dublin, Ireland; 2014. p. 33–40. https://doi.org/10.3115/v1/W14-4605. https://www.aclweb.org/anthology/W14-4605

83. Schneider G, Pettersson E, Percillier M. Comparing rule-based and SMT-based spelling normalisation for English historical texts. In: Proceedings of the NoDaLiDa 2017 workshop on processing historical language, Linköping University Electronic Press, Gothenburg; 2017. p. 40–46. https://www.aclweb.org/anthology/W17-0508
84. Hämäläinen M, Säily T, Rueter J, Tiedemann J, Mäkelä E. Normalizing early English letters to present-day English spelling. In: Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature. Association for computational linguistics, Santa Fe, New Mexico; 2018. p. 87–96. https://www.aclweb.org/anthology/W18-4510
85. Honnet PE, Popescu-Belis A, Musat C, Baeriswyl M. Machine translation of low-resource spoken dialects: strategies for normalizing swiss German. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan; 2018. https://www.aclweb.org/anthology/L18-1597
86. Napoles C, Callison-Burch C. Systematically adapting machine translation for grammatical error correction. In: Proceedings of the 12th workshop on innovative use of NLP for building educational applications. Association for computational linguistics, Copenhagen, Denmark; 2017. p. 345–356, https://doi.org/10.18653/v1/W17-5039. https://www.aclweb.org/anthology/W17-5039
87. Nakov P, Tiedemann J. Combining word-level and character-level models for machine translation between closely-related languages. In: Proceedings of the 50th annual meeting of the association for computational linguistics: short papers-volume 2. Association for computational linguistics; 2012. p. 301–305.
88. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Doklady 1966;10(8):707–710, Doklady Akad Nauk SSSR 1965;163(4):845–848.
89. Melamed ID. Bitext maps and alignment via pattern recognition. Comput Linguistics 1999;25(1):107–130. https://www.aclweb.org/anthology/J99-1003
90. Ciobanu AM, Dinu LP. Automatic detection of cognates using orthographic alignment. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers). Association for computational linguistics, Baltimore, Maryland; 2014. p. 99–105, https://doi.org/10.3115/v1/P14-2017. https://www.aclweb.org/anthology/P14-2017
91. Mulloni A, Pekar V. Automatic detection of orthographics cues for cognate recognition. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy, 2006. http://www.lrec-conf.org/proceedings/lrec2006/pdf/676_pdf.pdf
92. Simard M, Foster GF, Isabelle P. Using cognates to align sentences in bilingual corpora. In: Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-volume 2. IBM Press; 1993. p. 1071–1082.
93. Simard M, Foster GF, Isabelle P. Using cognates to align sentences in bilingual corpora. In: Proceedings of the 1993 conference of the Centre for Advanced Studies on collaborative research: distributed computing - volume 2. IBM Press, CASCON '93; 1993. p. 1071–1082.
94. Church KW. Char_align: a program for aligning parallel texts at the character level. In: 31st annual meeting of the association for computational linguistics, Columbus, Ohio, USA; 1993. p. 1–8. https://doi.org/10.3115/981574.981575. https://www.aclweb.org/anthology/P93-1001
95. Bemova A, Oliva K, Panevova J. Some problems of machine translation between closely related languages. In: Coling Budapest 1988 Volume 1: international conference on computational linguistics; 1988. http://www.aclweb.org/anthology/C88-1010
96. Hajic J. Machine translation of very close languages. In: Sixth applied natural language processing conference. Association for computational linguistics, Seattle, Washington, USA; 2000. p. 7–12. https://doi.org/10.3115/974147.974149. https://www.aclweb.org/anthology/A00-1002
97. Nakov P, Ng HT. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Association for computational linguistics; 2009. p. 1358–1367. http://www.aclweb.org/anthology/D09-1141
98. Popović M, Ljubešić N. Exploring cross-language statistical machine translation for closely related South Slavic languages. In: Proceedings of the EMNLP'2014 workshop on language technology for closely related languages and language variants. Association for computational linguistics; 2014. p. 76–84. https://doi.org/10.3115/v1/W14-4210. http://www.aclweb.org/anthology/W14-4210
99. Popović M, Arcan M, Klubička F. Language related issues for machine translation between closely related South Slavic languages. In: Proceedings of the third workshop on NLP for similar languages, varieties and dialects (VarDial3). The COLING 2016 Organizing Committee; 2016. p. 43–52. http://www.aclweb.org/anthology/W16-4806
100. Beinborn L, Zesch T, Gurevych I. Cognate production using character-based machine translation. In: Proceedings of the sixth international joint conference on natural language processing; 2013. p. 883–891.
101. Menacer MA, Langlois D, Jouvet D, Fohr D, Mella O, Smaïli K. Machine translation on a parallel code-switched corpus. In: Canadian conference on artificial intelligence, Springer; 2019. p. 426–432.
102. Fadaee M, Monz C. Back-translation sampling by targeting difficult words in neural machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for computational linguistics, Brussels, Belgium; 2018. p. 436–446. https://doi.org/10.18653/v1/D18-1040. https://www.aclweb.org/anthology/D18-1040
103. Chakravarthi BR, Arcan M, McCrae JP. Improving wordnets for under-resourced languages using machine translation. In: Proceedings of the 9th global WordNet conference, The Global WordNet Conference 2018 Committee; 2018. http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_16
104. Dhar M, Kumar V, Shrivastava M. Enabling code-mixed translation: parallel corpus creation and MT augmentation approach. In: Proceedings of the first workshop on linguistic resources for natural language processing. Association for computational linguistics, Santa Fe, New Mexico, USA; 2018. p. 131–140. https://www.aclweb.org/anthology/W18-3817
105. Rijhwani S, Sequiera R, Choudhury MC, Bali K. Translating codemixed tweets: a language detection based system. In: 3rd workshop on Indian language data resource and evaluation-WILDRE-3; 2016. p. 81–82.
106. Niu X, Denkowski M, Carpuat M. Bi-directional neural machine translation with synthetic parallel data. In: Proceedings of the 2nd workshop on neural machine translation and generation. Association for computational linguistics, Melbourne, Australia; 2018. p. 84–91. https://doi.org/10.18653/v1/W18-2710.https://www.aclweb.org/anthology/W18-2710
107. Riyadh RR, Kondrak G. Joint approach to deromanization of code-mixed texts. In: Proceedings of the sixth workshop on NLP for similar languages, varieties and dialects; 2019. p. 26–34.

108. Cohn T, Lapata M. Machine translation by triangulation: making effective use of multi-parallel corpora. In: Proceedings of the 45th annual meeting of the association of computational linguistics, Prague, Czech Republic; 2007. p. 728–735. https://www.aclweb.org/anthology/P07-1092
109. Utiyama M, Isahara H. A comparison of pivot methods for phrase-based statistical machine translation. In: Human Language Technologies 2007: the conference of the North American Chapter of the Association for computational linguistics; proceedings of the main conference. Association for computational linguistics, Rochester, New York; 2007. p. 484–491. https://www.aclweb.org/anthology/N07-1061
110. Edunov S, Ott M, Auli M, Grangier D. Understanding back-translation at scale. In: Proceedings of the 2018 conference on empirical methods in natural language processing, Association for computational linguistics, Brussels, Belgium; 2018. p. 489–500. https://doi.org/10.18653/v1/D18-1045. https://www.aclweb.org/anthology/D18-1045
111. Ahmadnia B, Serrano J, Haffari G. Persian-Spanish low-resource statistical machine translation through English as pivot language. In: Proceedings of the international conference recent advances in natural language processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria; 2017. p. 24–30. https://doi.org/10.26615/978-954-452-049-6_004
112. Poncelas A, Popović M, Shterionov D, Maillette de Buy Wenniger G, Way A. Combining PBSMT and NMT back-translated data for efficient NMT. In: Natural language processing in a deep learning world, INCOMA Ltd., Varna, Bulgaria; 2019. p. 922–931, https://doi.org/10.26615/978-954-452-056-4_107. https://www.aclweb.org/anthology/R19-1107
113. Tiedemann J, Cap F, Kanerva J, Ginter F, Stymne S, Östling R, Weller-Di Marco M. Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. In: Proceedings of the first conference on machine translation: volume 2, shared task papers. Association for computational linguistics, Berlin, Germany; 2016. p. 391–398. https://doi.org/10.18653/v1/W16-2326. https://www.aclweb.org/anthology/W16-2326
114. Graça M, Kim Y, Schamper J, Khadivi S, Ney H. Generalizing back-translation in neural machine translation. In: Proceedings of the fourth conference on machine translation (volume 1: research papers). Association for computational linguistics, Florence, Italy; 2019. p. 45–52,.https://doi.org/10.18653/v1/W19-5205. https://www.aclweb.org/anthology/W19-5205
115. Hoang VCD, Koehn P, Haffari G, Cohn T. Iterative back-translation for neural machine translation. In: Proceedings of the 2nd workshop on neural machine translation and generation, Association for computational linguistics, Melbourne, Australia; 2018. p. 18–24. https://doi.org/10.18653/v1/W18-2703. https://www.aclweb.org/anthology/W18-2703.
116. Prabhumoye S, Tsvetkov Y, Salakhutdinov R, Black AW. Style transfer through back-translation. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). Association for computational linguistics, Melbourne, Australia; 2018. p. 866–876. https://doi.org/10.18653/v1/P18-1080. https://www.aclweb.org/anthology/P18-1080
117. Kunchukuttan A, Shah M, Prakash P, Bhattacharyya P. Utilizing lexical similarity between related, low-resource languages for pivot-based smt. arXiv preprint arXiv:170207203; 2017.
118. Saunders D, Stahlberg F, de Gispert A, Byrne B. Domain adaptive inference for neural machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 222–228. https://doi.org/10.18653/v1/P19-1022. https://www.aclweb.org/anthology/P19-1022
119. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. Curran Associates Inc., USA, NIPS'17; 2017. p. 6000–6010. http://dl.acm.org/citation.cfm?id=3295222.3295349
120. Wang Q, Li B, Xiao T, Zhu J, Li C, Wong DF, Chao LS. Learning deep transformer models for machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, 2019. p. 1810–1822. https://doi.org/10.18653/v1/P19-1176. https://www.aclweb.org/anthology/P19-1176
121. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar; 2014. p. 1724–1734. https://doi.org/10.3115/v1/D14-1179. https://www.aclweb.org/anthology/D14-1179
122. Luong T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for computational linguistics, Lisbon, Portugal; 2015. p. 1412–1421. https://doi.org/10.18653/v1/D15-1166. https://www.aclweb.org/anthology/D15-1166
123. Sen S, Gupta KK, Ekbal A, Bhattacharyya P. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 3083–3089. https://doi.org/10.18653/v1/P19-1297. https://www.aclweb.org/anthology/P19-1297
124. Wang Y, Zhou L, Zhang J, Zhai F, Xu J, Zong C. A compact and language-sensitive multilingual translation method. In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy; 2019. p. 1213–1223. https://doi.org/10.18653/v1/P19-1117. https://www.aclweb.org/anthology/P19-1117
125. Ha T, Niehues J, Waibel AH. Toward multilingual neural machine translation with universal encoder and decoder. In: Proceedings of the international workshop on spoken language translation; 2016. http://workshop2016.iwslt.org/downloads/IWSLT_2016_paper_5.pdf
126. Chakravarthi BR, Arcan M, McCrae JP. Comparison of different orthographies for machine translation of under-resourced Dravidian languages. In: 2nd conference on language, data and knowledge (LDK 2019), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany. Open access series in informatics (OASIcs); 2019;70. p. 6:1–6:14, https://doi.org/10.4230/OASIcs.LDK.2019.6. http://drops.dagstuhl.de/opus/volltexte/2019/10370
127. Chakravarthi BR, Arcan M, McCrae JP. Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In: Proceedings of the second workshop on multilingualism at the intersection of knowledge bases and machine translation; 2019. p. 1–7.
128. Chakravarthi BR, Priyadharshini R, Stearns B, Jayapal A, S S, Arcan M, Zarrouk M, McCrae JP. Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In: Proceedings of the 2nd workshop on technologies for MT of low resource languages. European Association for Machine Translation, Dublin, Ireland; 2019. p. 56–63. https://www.aclweb.org/anthology/W19-6809
129. Li X, Michel P, Anastasopoulos A, Belinkov Y, Durrani N, Firat O, Koehn P, Neubig G, Pino J, Sajjad H. Findings of the first shared task on machine translation robustness. In: Proceedings of the fourth conference on machine translation (volume 2: shared

task papers, day 1), Association for computational linguistics, Florence, Italy; 2019. p. 91–102, https://doi.org/10.18653/v1/W19-5303. https://www.aclweb.org/anthology/W19-5303
130. Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation. In: International conference on learning representations; 2018. https://openreview.net/forum?id=BJ8vJebC-
131. Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. In: Proceedings of the thirteenth AAAI conference on artificial intelligence. AAAI Press, AAAI; 2016:16. p. 2741–9.
132. Cherry C, Foster G, Bapna A, Firat O, Macherey W. Revisiting character-based neural machine translation with capacity and compression. In: Proceedings of the 2018 conference on empirical methods in natural language processing, association for computational linguistics, Brussels, Belgium; 2018. p. 4295–4305. https://doi.org/10.18653/v1/D18-1461. https://www.aclweb.org/anthology/D18-1461
133. Costa-jussà MR, Fonollosa JAR. Character-based neural machine translation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers). Association for computational linguistics, Berlin, Germany; 2016. p. 357–361. https://doi.org/10.18653/v1/P16-2058. https://www.aclweb.org/anthology/P16-2058.
134. Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation. Trans Assoc Comput Linguistics. 2017;5:365–78. https://doi.org/10.1162/tacl_a_00067. https://www.aclweb.org/anthology/Q17-1026
135. Yang Z, Chen W, Wang F, Xu B. A character-aware encoder for neural machine translation. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, The COLING 2016 Organizing Committee, Osaka, Japan; 2016. p. 3063–3070. https://www.aclweb.org/anthology/C16-1288
136. Chitnis R, DeNero J. Variable-length word encodings for neural translation models. In: Proceedings of the 2015 conference on empirical methods in natural language processing, association for computational linguistics, Lisbon, Portugal; 2015. p. 2088–2093. https://doi.org/10.18653/v1/D15-1249. https://www.aclweb.org/anthology/D15-1249
137. Ding S, Renduchintala A, Duh K. A call for prudent choice of subword merge operations in neural machine translation. In: Proceedings of machine translation summit XVII volume 1: research track, European Association for Machine Translation, Dublin, Ireland; 2019. p. 204–213. URL https://www.aclweb.org/anthology/W19-6620
138. Schuster M, Nakajima K. Japanese and Korean voice search. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2012. p. 5149–5152. https://doi.org/10.1109/ICASSP.2012.6289079
139. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144; 2016.
140. Kudo T, Richardson J. Sentence piece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations. Association for Computational Linguistics, Brussels, Belgium; 2018. p. 66–71. https://doi.org/10.18653/v1/D18-2012. https://www.aclweb.org/anthology/D18-2012
141. Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia; 2018. p. 66–75. https://doi.org/10.18653/v1/P18-1007. https://www.aclweb.org/anthology/P18-1007
142. Klein G, Kim Y, Deng Y, Senellart J, Rush AM. OpenNMT: open-source toolkit for neural machine translation. CoRR arXiv:abs/1701.02810; 2017.
143. Jha S, Sudhakar A, Singh AK. Learning cross-lingual phonological and orthagraphic adaptations: a case study in improving neural machine translation between low-resource languages. J Language Model. 2019;7(2):101–42.
144. Bhattacharyya P, Khapra MM, Kunchukuttan A. Statistical machine translation between related languages. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: tutorial abstracts, association for computational linguistics, San Diego, California; 2016. p. 17–20, https://doi.org/10.18653/v1/N16-4006. URL https://www.aclweb.org/anthology/N16-4006
145. Grönroos SA, Virpioja S, Kurimo M. Cognate-aware morphological segmentation for multilingual neural translation. In: Proceedings of the third conference on machine translation: shared task papers, association for computational linguistics, Belgium, Brussels; 2018. p. 386–393. https://doi.org/10.18653/v1/W18-6410. https://www.aclweb.org/anthology/W18-6410
146. Cherry C, Suzuki H. Discriminative substring decoding for transliteration. In: Proceedings of the 2009 conference on empirical methods in natural language processing, association for computational linguistics; 2009. p. 1066–1075. http://www.aclweb.org/anthology/D09-1111
147. Bhat RA, Bhat IA, Jain N, Sharma DM. A house united: bridging the script and lexical barrier between Hindi and Urdu. In: COLING 2016, 26th international conference on computational linguistics. Proceedings of the conference: technical papers, December 11–16, 2016, Osaka, Japan; 2016. p. 397–408. http://aclweb.org/anthology/C/C16/C16-1039.pdf
148. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for computational linguistics, Philadelphia, Pennsylvania, USA; 2002. p. 311–318. https://doi.org/10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040
149. Kunchukuttan A, Khapra M, Singh G, Bhattacharyya P. Leveraging orthographic similarity for multilingual neural transliteration. Trans Assoc Comput Linguistics 2018;6:303–316. http://aclweb.org/anthology/Q18-1022
150. Baniata LH, Park S. Park SB. A neural machine translation model for Arabic dialects that utilizes multitask learning (MTL). Comput Intell Neurosci. 2018.
151. Halpern J. Very large-scale lexical resources to enhance Chinese and Japanese machine translation. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan; 2018. https://www.aclweb.org/anthology/L18-1137
152. Ugawa A, Tamura A, Ninomiya T, Takamura H, Okumura M. Neural machine translation incorporating named entity. In: Proceedings of the 27th international conference on computational linguistics. Association for computational linguistics, Santa Fe, New Mexico, USA; 2018. p. 3240–3250. https://www.aclweb.org/anthology/C18-1274
153. Birch A, Haddow B, Tito I, Barone AVM, Bawden R, Sánchez-Martínez F, Forcada ML, Esplà-Gomis M, Sánchez-Cartagena V, Pérez-Ortiz JA, Aziz W, Secker A, van der Kreeft P. Global under-resourced media translation (GoURMET). In: Proceedings of machine translation summit XVII volume 2: translator, project

and user tracks. European Association for Machine Translation, Dublin, Ireland; 2019. p. 122–122. https://www.aclweb.org/anthology/W19-6723
154. Chakravarthi BR, Jose N, Suryawanshi S, Sherly E, McCrae JP (2020) A sentiment analysis dataset for code-mixed Malayalam-English. In: Proceedings of the 1st joint workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL). European Language Resources Association (ELRA). France: Marseille; 2020.
155. Chakravarthi BR, Muralidaran V, Priyadharshini R, McCrae JP (2020b) Corpus creation for sentiment analysis in code-mixed Tamil-English text. In: Proceedings of the 1st joint workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL). European Language Resources Association (ELRA). France: Marseille; 2020.
156. Jose N, Chakravarthi BR, Suryawanshi S, Sherly E, McCrae JP. A survey of current datasets for code-switching research. In: 2020 6th international conference on advanced computing and communication systems (ICACCS); 2020.
157. Priyadharshini R, Chakravarthi BR, Vegupatti M, McCrae JP. Named entity recognition for code-mixed Indian corpus using meta embedding. In: 2020 6th international conference on advanced computing and communication systems (ICACCS); 2020.
158. Ranjan P, Raja B, Priyadharshini R, Balabantaray RC. A comparative study on code-mixed data of Indian social media vs formal text. In: 2016 2nd international conference on contemporary computing and informatics (IC3I); 2016. p. 608–611. https://doi.org/10.1109/IC3I.2016.7918035
159. Tiedemann J. Synchronizing translated movie subtitles. In: Proceedings of the sixth international conference on language resources and evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco; 2008. http://www.lrec-conf.org/proceedings/lrec2008/pdf/484_paper.pdf
160. Fadaee M, Bisazza A, Monz C. Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), Vancouver, Canada; 2017. p. 567–573. https://doi.org/10.18653/v1/P17-2090. https://www.aclweb.org/anthology/P17-2090
161. Li Z, Specia L. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In: Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019). Association for computational linguistics, Hong Kong, China; 2019. p. 328–336, https://doi.org/10.18653/v1/D19-5543. https://www.aclweb.org/anthology/D19-5543
162. Song K, Zhang Y, Yu H, Luo W, Wang K, Zhang M. Code-switching for enhancing NMT with pre-specified translation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Association for computational linguistics, Minneapolis, Minnesota; 2019. p. 449–459. https://doi.org/10.18653/v1/N19-1044. https://www.aclweb.org/anthology/N19-1044
163. Dou Q, Vaswani A, Knight K. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for computational linguistics, Doha, Qatar; 2014. p. 557–565. https://doi.org/10.3115/v1/D14-1061. https://www.aclweb.org/anthology/D14-1061
164. Koehn P, Knight K. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In: Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence. AAAI Press; 2000. p. 711–715.
165. Ravi S, Knight K. Deciphering foreign language. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for computational linguistics, Portland, Oregon, USA; 2011. p. 12–21. https://www.aclweb.org/anthology/P11-1002
166. Artetxe M, Labaka G, Agirre E. Unsupervised statistical machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for computational linguistics, Brussels, Belgium; 2018. p. 3632–3642, https://doi.org/10.18653/v1/D18-1399. https://www.aclweb.org/anthology/D18-1399
167. Klementiev A, Irvine A, Callison-Burch C, Yarowsky D. Toward statistical machine translation without parallel corpora. In: Proceedings of the 13th conference of the European chapter of the association for computational linguistics. Association for computational linguistics, Avignon, France; 2012. p. 130–140. https://www.aclweb.org/anthology/E12-1014
168. Artetxe M, Labaka G, Agirre E, Cho K. Unsupervised neural machine translation. In: Proceedings of the sixth international conference on learning representations; 2018.
169. Rosner M, Sultana K. Automatic methods for the extension of a bilingual dictionary using comparable corpora. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland; 2014. p. 3790–3797. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1169_Paper.pdf
170. Turcato D. Automatically creating bilingual lexicons for machine translation from bilingual text. In: 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, volume 2, association for computational linguistics, Montreal, Quebec, Canada; 1998. p. 1299–1306, https://doi.org/10.3115/980691.980781. https://www.aclweb.org/anthology/P98-2212
171. Haghighi A, Liang P, Berg-Kirkpatrick T, Klein D. Learning bilingual lexicons from monolingual corpora. In: Proceedings of ACL-08: HLT, association for computational linguistics, Columbus, Ohio; 2008. p. 771–779. https://www.aclweb.org/anthology/P08-1088
172. Berg-Kirkpatrick T, Bouchard-Côté A, DeNero J, Klein D. Painless unsupervised learning with features. In: Human language technologies: the 2010 annual conference of the North American Chapter of the Association for computational linguistics, Los Angeles, California; 2010. p. 582–590. https://www.aclweb.org/anthology/N10-1083
173. Dyer C, Clark JH, Lavie A, Smith NA. Unsupervised word alignment with arbitrary features. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Association for computational linguistics, Portland, Oregon, USA; 2011. p. 409–419. https://www.aclweb.org/anthology/P11-1042
174. Hauer B, Nicolai G, Kondrak G. Bootstrapping unsupervised bilingual lexicon induction. In: Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics: volume 2, short papers, Valencia, Spain; 2017. p. 619–624. https://www.aclweb.org/anthology/E17-2098
175. Riley P, Gildea D. Orthographic features for bilingual lexicon induction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers). Association for computational linguistics, Melbourne, Australia; 2018. p. 390–394. https://doi.org/10.18653/v1/P18-2062. https://www.aclweb.org/anthology/P18-2062
176. Chu C, Nakazawa T, Kurohashi S. Improving statistical machine translation accuracy using bilingual lexicon extraction with

paraphrases. In: Proceedings of the 28th Pacific Asia conference on language, information and computing, Department of Linguistics, Chulalongkorn University, Phuket, Thailand; 2014. p. 262–271. URL https://www.aclweb.org/anthology/Y14-1032

177. Dou Q, Knight K. Dependency-based decipherment for resource-limited machine translation. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for computational linguistics, Seattle, Washington, USA; 2013. p. 1668–1676. https://www.aclweb.org/anthology/D13-1173

178. Bloodgood M, Strauss B. Acquisition of translation lexicons for historically unwritten languages via bridging loanwords. In: Proceedings of the 10th workshop on building and using comparable Corpora, association for computational linguistics; 2017. p. 21–25. https://doi.org/10.18653/v1/W17-2504. http://aclweb.org/anthology/W17-2504

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.