A Survey on Document-level Machine Translation: Methods and Evaluation

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari
Faculty of Information Technology, Monash University, Clayton VIC, Australia
{firstname.lastname}@monash.edu

Abstract

Machine translation (MT) is an important task in natural language processing (NLP) as it automates the translation process and reduces the reliance on human translators. With the advent of neural networks, the translation quality surpasses that of the translations obtained using statistical techniques. Up until three years ago, all neural translation models translated sentences independently, without incorporating any extra-sentential information. The aim of this paper is to highlight the major works that have been undertaken in the space of document-level machine translation before and after the neural revolution so that researchers can recognise where we started from and which direction we are heading in. When talking about the literature in statistical machine translation (SMT), we focus on works which have tried to improve the translation of specific discourse phenomena, while in neural machine translation (NMT), we focus on works which use the wider context explicitly. In addition to this, we also cover the evaluation strategies that have been introduced to account for the improvements in this domain.

1 Introduction

Machine translation (MT) is the process of automating translation between natural languages with the aid of computers. Translation, in itself, is a difficult task even for humans as it requires a thorough understanding of the source text and a good knowledge of the target language, hence requiring the human translators to have high degree of proficiency in both languages. Due to the dearth of professional translators and the rapid need of availability of multilingual digital content, for example, on the Internet, MT has grown immensely over the past few decades for the purposes of international communication.

Up until a few years ago, MT was mostly formalised through statistical techniques, hence very aptly named statistical machine translation (SMT), which involved meticulously crafting features to extract implicit information from corpora of bilingual sentence-pairs [10]. These hand-engineered features were an intrinsic part of SMT and were one of the reasons behind its inflexibility. MT has come a long way from then to the state-of-the-art neural machine translation (NMT) systems [107, 3, 113] employed commercially today, which are based on neural network black-box models requiring little to no feature engineering. The results obtained by MT systems have seen rapid improvements in the past few years, and have added to their popularity among the general public [70, 53] and the research community [121, 38, 12].

Inspite of its success, MT has been based on strong independence and locality assumptions, that is either translating word-by-word or phrase-by-phrase (as done by SMT) or translating sentences in isolation (as done by NMT). Text, on the contrary, does not consist of isolated, unrelated elements, but of collocated and structured group of sentences bound together by complex linguistic elements, referred to as the discourse [42]. Ignoring the inter-relations among these discourse elements, results in translations which may be perfect at the sentence-level but lack crucial properties of the text.
hindering understanding. One way to address this issue is to exploit the underlying discourse
structure of a text by utilising the information in the wider-sentential context. This is not a novel
idea in itself and has been advocated by MT pioneers for decades [5, 94], but was mostly ignored
in the era of SMT due to computational efficiency and tractability concerns by the MT community.
Recently, with the increase in computational power available to us and the wide-scale application
of neural networks to machine translation, we are finally in a position to forego the independence
constraints that have impeded the progress in MT since long.

The aim of this survey is to highlight the major works that have been undertaken in the space of
document-level machine translation before and after the neural revolution (Sections 3 and 4 respec-
tively). By document-level MT, we mean works which utilise inter-sentential context information
comprising discourse aspects of a document or surrounding sentences in the document. In addition
to this, we also cover the evaluation strategies introduced to account for improvements in this do-
main (Section[5]) and conclude by presenting avenues for future research (Section 6). Before moving
on with the main agenda, we briefly describe the basics of statistical and neural MT models and
their evaluation in the following section.

2 Machine Translation: Foundations and Evaluation

Machine translation has been around for a long time and various approaches have been proposed to
make it on-par with human translation. The approach to MT which is strongly correlated with the
current state-of-the-art NMT models and worth mentioning here is statistical machine translation
(SMT). SMT models the probability of a sentence translation in one language given a source sentence
in another language. This probability is determined automatically by training a statistical model
using a parallel corpus containing source and target translation-pairs. The advantages of SMT
over its predecessors were that it was data-driven and language independent and was considered a
state-of-the-art technique up until the advent of neural-based approaches.

Mathematically, the goal of SMT (and NMT) is to find the most probable target sequence \( \hat{y} \)
given a source sentence, that is:

\[
\hat{y} = \arg \max_y P(y | x)
\]

Using Bayes’ rule, this conditional probability can be reformulated as follows:

\[
\hat{y} = \arg \max_y P(y)P(x | y)
\]

where \( P(y) \), aka the language model (LM) usually based on trigram probabilities and estimated
using monolingual corpora, assigns a higher probability to fluent, grammatical sentences and \( P(x | y) \), aka the translation model, assigns a higher probability to sentences that have corresponding
meaning. The translation model is parameterised using an alignment function which represents how
a source word is aligned to a target word [10]. The more often two words occur together in different
sentence-pairs, the more likely it is that they are aligned to each other and have equivalent meaning.
These word-based models were superseded by phrase-based models [63, 47] which used many-to-many
alignments between the source and target words stored in a phrase table [80].

While SMT was successfully deployed in many commercial systems, it did not work very well and
suffered from two major drawbacks. First, translation decisions were local as the translation was
performed phrase-by-phrase and long-distance dependencies were often ignored. Secondly and more
problematically, the entire MT pipeline became increasingly complex as many different components
had to be tuned separately, e.g., translation models, language models, reordering models, etc., which
made it difficult to combine them together and have a single end-to-end model. As a result, when the
AI winter was over and neural networks resurfaced as the new approach to solve natural language
processing (NLP) problems, it was seen as the next logical step to use them for machine translation
as well. NMT, even though quite recent (since 2014), has opened a new era in MT for both research
and commercial purposes.

NMT models, in general, are based on an encoder-decoder framework (Figure 1) where the
encoder reads the source sentence to compute a real-valued representation and the decoder generates
the target translation one word at a time given the previously computed representation. The initial
model by Sutskever et al. \cite{DBLP:journals/corr/SutskeverVL14} used a fixed representation of the source sentence to generate the

target sentence. It was quickly replaced by the attention-based encoder-decoder architecture which
generated a dynamic context representation \cite{bahdanau2014neural}. These models were mostly based on recurrent
neural networks (RNNs) \cite{schuster1997bidirectional} which use recurrent connections to exhibit temporal dynamic behavior
over time, and were thus considerably suitable for modelling sequential information. However,
the major drawback of such sequential computation was that it hindered parallelisation within
training examples and became a bottleneck when processing longer sentences. Most recently, a new
model architecture, the Transformer, was proposed which is based solely on attention mechanisms,
dispensing with the recurrence entirely. It has proved to achieve state-of-the-art results on several
language-pairs \cite{vaswani2017attention}.

To evaluate the quality of the generated translations, the most popular automatic evaluation
metric is BLEU (Bilingual Evaluation Understudy) \cite{papineni2002bleu} which has been a de-facto standard for
evaluating translation outputs since it was first proposed in 2002. The core idea is to aggregate the
count of words and phrases (n-grams) that overlap between machine and reference translations. The
BLEU metric ranges from 0 to 1 where 1 means an identical output with the reference. Although
BLEU correlates well with human judgment \cite{papineni2002bleu}, it relies on precision alone and does not take into
account recall – the proportion of the matched n-grams out of the total number of n-grams in the
reference translation. METEOR \cite{denkowski2014meteor,denkowski2011meteor} was proposed to address the shortcomings of BLEU. It scores
a translation output by performing a word-to-word alignment between the translation output and a
given reference translation. The alignments are produced via a sequence of word-mapping modules,
that is, if the words are exactly the same, same after they are stemmed using the Porter stemmer,
and if they are synonyms of each other. After obtaining the final alignment, METEOR uses
F\textsubscript{mean}, which is just the parameterised harmonic mean of unigram precision and recall \cite{rst}, where unigram
precision \( P \) is the ratio of the number of mapped unigrams to the total number of unigrams in the
output translation \( (m/t) \), while unigram recall \( R \) is the ratio of the number of mapped unigrams to
the total number of unigrams in the reference translation \( (m/r) \). METEOR has also demonstrated
to have a high level of correlation with human judgment, even outperforming that of BLEU \cite{denkowski2014meteor}. To
make the results of the aforementioned MT evaluation metrics more reliable, a statistical significance
test should be performed \cite{tetreault2009statistical}, which indicates whether the difference in translation quality of two or
more systems is due to a difference in true system quality. Although other MT evaluation metrics
have been proposed, we only mention the most popular BLEU and METEOR, as these are relevant
for the purposes of this survey.

Figure 1: A general overview of an encoder-decoder model.
3 Discourse in Statistical Machine Translation

Most MT models are built on strong independence assumptions whether it is based on locality assumptions within a sentence as done by phrase-based models or that outside the sentence as done by even the most advanced NMT models today. From a linguistics perspective, this assumption in practice is invalid, as any piece of text is much more than just a single sentence and making this assumption means ignoring the underlying discourse structure of the text and still hoping that the translation would not fall short. Discourse is defined as a group of sentences that are contiguous, structured and exhibit coherency [42]. Although the problem of machine translation itself has been around for decades, the works which have tried to address the problem of discourse in MT are still just brushing the surface with more research yet to be undertaken. In this and following sections, we will describe the research on discourse in SMT (Section 3) and NMT (Section 4) followed by a description of how to evaluate translation outputs of larger pieces of text (Section 5).

In terms of SMT, we will be mentioning research which has tried to incorporate different aspects of discourse in SMT, beginning with the document-level discourse structure and moving on to specific discourse phenomena like pronominal anaphora, lexical cohesion and consistency, coherence, and discourse connectives. An overview of the works for discourse in SMT is provided in Table 1. For the purposes of this survey, we will only be mentioning works which have considered inter-sentential context information.

3.1 Discourse and Document Structure

An initial work on discourse in MT [62] (predating SMT) used a discourse transfer model to re-order the clauses and sentences of an input text (in Japanese) to make it closer to the natural discourse structure of text in a target language (English) and thus cater to the cross-lingual discourse shift. Almost a decade later, Foster et al. [16] presented an SMT system for translating the Canadian Hansard corpus (parliamentary proceedings) in which they changed the language model to incorporate structural features at the sentence-level (year, source language, speaker name, title, and section) without being explicitly dependent on the content of the other sentences. Louis and Webber [57] proposed a structured model for translating Wikipedia biography articles using a cache to encourage the use of article sub-structure (based on topics) by using words conforming to the smaller topic segments in the article.

It is a challenge to include document structure when training an SMT model, but a more challenging problem is to incorporate this information at the decoding stage. This is because the decoding of phrase-based SMT models not only relies on the sentence-independence assumption but is realised as a search for the highest-scoring translation in the space of exponentially possible translations that could be generated by the model [47]. One possible solution, proposed by Hardmeier et al. in [30, 31], is to start from an initial translation generated from a baseline decoder like Moses [46] and make local changes to that translation via elementary operations (changing phrase translations or word-order, and resegmentation) and transform it into a better translation. This decoder, referred to as Docent, was followed up by Stymne et al. [106] who incorporated readability constraints (including the ones to promote lexical consistency) into Docent to produce simplified translations; however, this resulted in deteriorated performance based on automatic evaluation.

3.2 Cohesion

Cohesion is a surface property of the text and refers to the way textual units are linked together grammatically or lexically [27]. The first form, grammatical cohesion, is based on the logical and structural content, while the second, lexical cohesion, is based on the usage of semantically related words. Most research on discourse in SMT has focused on lexical cohesion while some has focused on grammatical cohesion in terms of pronominal anaphora.

Pronominal Anaphora  Pronominal anaphora is the use of a pronoun to refer to someone or something mentioned previously in a text. It is a challenging problem in MT due to the variation of
the usage and distribution of pronouns across languages and can only be dealt with access to inter-
sentential context, specifically if the antecedent is not present in the same sentence. For example, a
neutral pronoun in a source language (English) may have a gender-sensitive pronoun in the target
language (German), requiring access to the antecedent to resolve the gender.

Initial attempts to exploiting anaphora information for the improvement of SMT systems, by
using a word-dependency model to incorporate the output of a coreference resolution system in SMT
[29], and by using a two-pass approach that includes annotations from a coreference system in the
second pass [52], did not yield promising results. There have also been attempts to cross-lingual
pronoun prediction in [79] and [32] where the latter attempt to use anaphora links as latent variables
in a neural network classifier.

Luong and Popescu-Belis [58] proposed to use a pronoun-aware language model that determines
a target pronoun based on the number and gender of preceding nouns or pronouns. Their method
then re-rank the translation hypotheses using the new LM and demonstrated improvements over the
baseline for English→French shared task in DiscoMT 2015. They also developed a fully probabilistic
model [59] that combined an additional translation model for pronouns, based on morphological and
semantic features, with a Spanish→English SMT system to improve the translation of personal and
possessive pronouns in Spanish to English. Werlen and Popescu-Belis [76] presented a coreference-
aware decoder for SMT based on similarity of coreference links in the source (Spanish) and target
(English) texts. Their post-editing scheme resulted in significant improvements in the accuracy of
pronoun translation [77], while the BLEU scores were unchanged.

**Lexical Cohesion** Lexical cohesion has two forms: repetition and collocation. The former is
achieved through synonyms and hyponyms (sometimes also referred to as lexical consistency), while
the latter uses related words that generally co-occur. There are three lines of work that try to
incorporate lexical cohesion in SMT, by employing: (i) cache-based approaches, (ii) lexical chains,
and (iii) two-pass approaches.

In terms of the first line of work, Tiedemann [108] tried to promote lexical consistency in SMT
by using adaptive language and translation models that use an exponentially decaying cache to
carry over word preferences from one sentence to the next. Gong et al. [20] also used a cache-
based approach in which they employ three types of caches: (i) a dynamic cache (similar to [108])
built using bilingual phrase pairs from the best translation hypotheses of previous sentences, (ii) a static cache which stored relevant bilingual phrase pairs extracted from similar bilingual documents, and (iii) a topic cache which stored the relevant target-side topic words. Their approach yielded significant improvements over the baseline in terms of BLEU score.

Falling into the second line of work, Xiong et al. [123] proposed a model that looks for lexical cohesion devices in the translation outputs of their MT system and then rewards the model for their appropriate usage based on conditional likelihood and mutual information. They reported significant improvements for Chinese→English SMT in terms of BLEU. In [124], a framework was presented that attempted to incorporate lexical cohesion in the translations via lexical chains. The source document lexical chains were first identified and then projected to the target-side using maximum entropy classifiers. Then, a lexical cohesion based translation was generated from the target lexical chains by integrating their cohesion models into a hierarchical phrase-based SMT system. Instead of relying on lexical resources, the method proposed in [69] detected lexical chains in the source and tried to preserve the semantic similarity among the words in their corresponding lexical chains in the target via word embeddings. The model was integrated into Docent and it was found through manual evaluation that it had a tendency to produce consistent translations of words in the chain.

The last line of work is based on incorporating document contexts into an initial translation obtained from a baseline MT system. Xiao et al. [122] first identified ambiguous words in the source and then obtained a set of consistent translations for each word using the distribution of its translation over the target document, after which the phrase table is updated by removing inconsistent phrase-pairs and a second pass of decoding is performed. The semantic document language model in [30] rewarded the use of semantically related words (found based on latent semantic analysis) in the translation output thus promoting lexical cohesion. Garcia et al. [18] proposed a two-pass approach to improve the translations already obtained by a sentence-level model. After the initial translation is obtained, they detect incorrect translations in the target document based on inconsistencies in meaning, gender and number disagreement among words, and suggest possible corrections. Their method did not yield improvement based on automatic evaluation which they claim to be due to the local changes made by their model. Later on, they designed a document-level scoring feature for lexical consistency [19] by measuring the suitability of a word translation according to its context and its other possible translations in the document based on word embeddings. They also extended Docent to incorporate a new operation that guides the search process to yield consistent translations. Finally, in [17] Garcia et al. made use of bilingual word vector models as the semantic language model in Docent to enforce translation choices that are semantically similar to the context.

3.3 Coherence

As opposed to cohesion which is a surface property of the text, coherence refers to the underlying meaning relation between units of text and its continuity [42]. It is a stronger requirement for a piece of text to meet than is required by cohesion, and not only embodies cohesion, but other referential components like different parts of text referring to the same entities (entity-based coherence), and relational components like connections between utterances in a discourse via coherence relations [28, 96]. Hence, coherence governs whether a text is semantically meaningful overall and how easily a reader can follow it.

Coherence has been explored for monolingual text but not much for bilingual text, like the one we deal with in MT. For SMT, the research in coherence mostly deals with studies that try to extend previously proposed coherence models for monolingual text to translation outputs [98, 96]. Smith et al. [97] further extend these models by proposing a new method to learn the syntactic patterns in a text.

3.4 Discourse Connectives

Discourse connectives, also referred to as discourse markers or cue words, are the words that signal the existence of a specific discourse relation or discourse structure in the text. These are mostly domain-specific and may be implicit or explicit depending on the language. If implicit, these may be
missed by the MT system in the translation although a human translator may be able to introduce them explicitly in the translation [33]. There have been studies that have tried to assess the ambiguity of discourse connectives for MT and have reported that the mismatch between implicit and explicit discourse connectives across languages results in deteriorated translation quality [54, 55]. Even explicitly annotating the discourse markers in the source text has a limited effect on translation quality for Chinese→English as reported in [128] and [102].

Meyer et al. [73] proposed to use an automatic scheme which annotates words with discourse sense by gathering information from the different ways they are translated in their correct translations, also referred to as translation spotting [11]. The impact of using this methodology was pretty low in terms of BLEU score for English-French [71].

3.5 Conclusion

After going through related work for discourse in SMT, it must be clear that incorporating discourse in SMT is a hard problem due to the various components in the SMT pipeline and the reliance on well-crafted and intuitive hand-engineered features for the various discourse phenomena. Furthermore, SMT is not very good at handling sentence-level phenomena such as syntactic reordering and long-distance agreement. Even if one can improve the discourse characteristics of a MT system output (that frequently contains local grammatical mistakes) via a post-editing step, noise from local errors makes such improvements difficult to measure. These were the main reasons that for a long time the MT community was put off to pursue valuable research in this area, mostly resulting in studies which highlighted the importance of pursuing document-level MT but less hands-on work which actually attempted to do it.

4 Discourse in Neural Machine Translation

Up until two years ago, there was no work in NMT that tried to incorporate any type of discourse phenomena mentioned previously, but with most sentence-based NMT systems achieving state-of-the-art performance compared to their SMT counterparts, this area of research has finally started to gain the popularity it deserves. The main difference between the research on discourse in NMT and SMT, apart from the general building blocks, is that the works in NMT rarely try to model discourse phenomena explicitly. On the contrary, they use sentences in the context directly via different modelling techniques and show how they perform on automatic evaluation while sometimes measuring the performance on specific test sets. An overview of the works for document-level NMT is provided in Table 2.

4.1 Incorporating context via additional components

The works in this section can be divided into two types: (i) those which use an additional context encoder and attention, and (ii) those which extend the translation units with the context.

Among the first line of work, Jean et al. [36] augment the attentional RNN-based NMT architecture with an additional attentional component over the previous source sentence. The context vector generated from this source-context attention is then added as an auxiliary input to the decoder hidden state. Through automatic evaluation and cross-lingual pronoun prediction, they found that although their approach yielded moderate improvements for a smaller training corpus, there was no improvement when the training set was much larger. Furthermore, their method suffers from an obvious limitation: the additional encoder and attention component introduced a significant amount of parameters meaning that their method could only incorporate limited context.

Wang et al. [118] proposed the first context-dependent NMT model to yield significant improvements over a context-agnostic sentence-based NMT model in terms of automatic evaluation. They employ a two-level hierarchical RNN to summarise the information in three previous source sentences, where the first-level RNNS are run over individual sentences and the second-level RNN is run over the single output vectors produced from the first-level RNN over context sentences. The final summary vector is then used to initialise the decoder, or added as an auxiliary input to the decoder.
| Approach                                          | Context Type | Lang. Pair       | Reference |
|--------------------------------------------------|--------------|-----------------|-----------|
| context encoding                                 |              | past future     |           |
| concatenated inputs                              | s            | 1               | De→En    | [109]    |
|                                                  | s, t         | 3               | En→It    | 4        |
|                                                  | s, t         | 1               | En→Fr    | 7        |
|                                                  | s, t         | variable        | En→De    | 52       |
| augmented input                                  | s s          | all             | De→En/En | 59       |
|                                                  | s s          | all             | En→Fr, En→De | 60 |
| cache decoder                                    | s, t         | variable        | Zh→En    | 111, 49  |
| attention                                        | s, t         | 3               | Zh/En→En | 118      |
|                                                  | s, t, s, t   | all             | En→De    | 58       |
| encoder w/attention                              | s, t         | 1               | En→Fr    | 6        |
| decoder                                          | s            | 1               | En→Ru    | 110      |
|                                                  | s, t         | 1               | En→Fr/De | 36       |
|                                                  | s           | all             | Fr/De/En/Ru→En | 67 |
|                                                  | s            | 1               | Zh→En    | 113      |
|                                                  | s, t         | 1               | De/Zh/Ja→En | 126    |
| decoder, output                                  | s, t         | all             | Fr/De/En | 65       |
| encoder, decoder                                 | s            | 2               | Zh/Fr→En | 129      |
|                                                  | s            | 2               | Fr→En    | 119      |
| second-pass decoder                              | t t          | all             | Zh→En    | 125      |
|                                                  | s, t         | 3               | En→Ru    | 113      |
| context-dependent post-editing                   | t*           | 4               | En→Ru    | 114      |
| learning w/context regularisation                | s            | 1               | En→Ru    | 35       |
| learning w/oracles                               | s            | 1               | En→De    | 105      |

Table 2: Overview of works which successfully incorporate extra-sentential context information in NMT. The initials of $s$ and $t$ in the context type column denote whether the context was from the source or target-side and amount is the maximum amount of context used in the referenced work. * means that the work does not have a notion of past and future and instead uses groups of sentences for training and evaluation.

state directly or after passing through a gate. Their approach showed promising results when using source-side context even though they found that considering target-side history inversely harmed translation performance.

Bawden et al. [4] used multi-encoder NMT models to exploit context from the previous source sentence whereby the information from the context and current source sentence is combined using either concatenation, gating or hierarchical attention. Further, they introduced an approach which combines the multiple encoders and decoding of both the previous and current sentence. They highlight the importance of target-side context but report deteriorated BLEU scores when using it.

Voita et al. [116] changed the encoder in the state-of-the-art Transformer architecture [113] to a context-aware encoder which has two sets of encoders, a source encoder and a context encoder, and the first $L−1$ layers shared. The previous source sentence serves as an input to the context encoder and its output is attended to by the $L^{th}$ layer of the source encoder, and then combined with the source encoder output using a gate. The final output of the context-aware encoder is then fed into the decoder. Their experiments on English→Russian subtitles data and analysis on the effect of context information for translating pronouns revealed that their model implicitly learned anaphora resolution which is quite promising as the model did not use any specialised features. They also experimented with using the following sentence as context and found that it underperformed the baseline Transformer architecture.

Among the works which extend the translation units with context, Rios et al. [89] focused on the problem of word sense disambiguation (WSD) in neural machine translation. One of the methods they employ to address the issue is to input the lexical chains of semantically similar words in a document as features to the NMT model. The lexical chains are detected via learnt sense embeddings. Although this methodology did not yield substantial improvements over the baseline.
on a generic test set but there is some improvement in terms of accuracy over a targeted test set introduced in the same work. They also found evidence that even humans are unable to resolve some of the ambiguities without document-level context which was not catered for in the targeted test set they used.

Tiedemann and Scherrer [109] conducted a pilot study which extends the translation units in two ways: (i) only extend the source sentence to include a single previous sentence, and (ii) extend both source and target sentences to include previous sentence in the corresponding context, without changing the underlying RNN-based NMT model. They reported marginal improvements in terms of BLEU for German→English subtitle translation, but through further analysis and manual evaluation found output examples in which referential expressions across sentence boundaries could be handled properly.

Agrawal et al. [1] extend the idea of concatenating translation units with Transformer [113] as the base model. For the source, they experimented with up to three previous and one next sentence, while for the target, they used up to two previous sentences as the context, that is, they generate the previous and current target sentence together. They also used an RNN-based version for comparison with their models and found that, when using RNNs, concatenation underperforms multi-encoder models like the ones used by [7]. They attribute this to the RNN’s inherent problem of not being able to accommodate long-range dependencies in a sequence. For the Transformer, they found that the next source sentence does help in improving NMT performance while using a larger number of previous target sentences deteriorates performance due to error propagation. They conclude that the Transformer’s ability to capture long-range dependencies via self-attention enables a simple technique like concatenation of context sentences to outperform its counterpart and multi-encoder approaches with RNNs.

Most recently, Scherrer et al. [92] have investigated the performance of concatenation-based context-aware NMT models in terms of different aspects of the discourse. They considered two popular datasets, the OpenSubtitles2016 [56] corpus and a subset of the corpus made available for the WMT 2019 news translation task. The experimental configurations for the concatenation setups are inspired from those used by [1] and [40]. To test the general performance of the document-level systems, they evaluated the systems with consistent (natural order of context sentences) and artificially scrambled context (random or no context). They found that using the scrambled context deteriorates performance for the Subtitles data but not as much for the WMT data and attribute this to the difference in the length and the number of sentences in the context (when using fixed number of tokens in the context).

4.2 Incorporating context via a two-pass approach

The works in this section can be divided into five types: (i) those which augment the source sentence with document-level token, (ii) those which employ a cache to store context information, (iii) those which use an additional context encoder and attention, (iv) those which use just an attention model over the context, and (v) those which introduce a context-aware decoder.

The recent work by Macé and Servan [60] falls under the first category. They account for the global source context information by adding a document tag as an additional token at the beginning of the source sentence and replace it with a document-level embedding when training the model. The document-level embedding is just the average of the word embeddings learnt while training the sentence-level model. Furthermore, the word embeddings are fixed while training the document-level model to maintain the relation between the word and the document embeddings. This minor change in the encoder input yields promising results for both translation directions of the English-French language-pair even though it does not yield significant improvements for English→German for two of the three test sets.

Among the second line of work, there have been two approaches that use cache to store relevant information from a document and then use this external memory to improve the translation quality [111] [49]. The first of these approaches by Tu et al. [111] uses a continuous cache to store recent hidden representations from the bilingual context, that is the key is designed to help match the

[111] [49]
query (current context vector produced via attention) to the source-side context, while the value is designed to help find the relevant target-side information to generate the next target word. The final context vector from the cache is then combined with the decoder hidden state via a gating mechanism. The cache has a finite length and is updated after generating a complete translation sentence. Their experiments on multi-domain Chinese→English datasets showed the effectiveness of their approach with negligible impact on the computational cost. The second approach by Kuang et al. [49] uses dynamic and topic caches (similar to the ones in [20]) to store target words from the preceding sentence translations and a set of target-side topical words semantically related to the source document, respectively. As opposed to the cache in [111], their dynamic cache follows a first-in, first-out scheme and is updated after generating each target word. At each decoding step, the target words in the final cache are scored and a gating mechanism is used to combine the score from the cache and the one produced by the NMT model. Their experimental results on the NIST Chinese→English translation task revealed that the cache-based neural model achieves consistent and significant improvements in terms of translation quality.

Falling in the third line of work which uses an additional context encoder and attention, Maruf and Haffari [65] present a document-level neural MT model that successfully captures global source and target document context via coarse attention over the sentences in the source and target documents. Their model augments the vanilla RNN-based sentence-level NMT model with external memories to incorporate documental interdependencies on both source and target sides. They use a two-level RNN to encode only the source sentences in the document before applying the attention and do not perform any additional encoding for the target sentences due to risk of error propagation. They also propose an iterative decoding algorithm based on block coordinate descent and show statistically significant improvements in the translation quality over the context-agnostic baseline for three language-pairs.

Zhang et al. [129] use a context-aware encoder in the Transformer. However, instead of training their model from scratch, like [116], they use pre-trained embeddings from the sentence-based Transformer as input to their context encoder. In the second stage of training, they only learn the document-level parameters and do not fine-tune the sentence-level parameters of their model similar to [111]. They experimented with NIST Chinese→English and IWSLT French→English translation tasks and reported significant gains over the baseline in terms of BLEU score. Around the same time, Kuang and Xiong [48] proposed an inter-sentence gate model to control the amount of information coming from the previous sentence while generating the translation of the current one. Their model contains two attention-based context vectors, one from the current sentence and another from the preceding one. They integrate an inter-sentence gate to combine the information from the two context vectors when updating the decoder hidden state. They show through experiments on Chinese→English language-pair that the gate is effective in capturing cross-sentence dependencies and lexical cohesion devices like repetition.

A recent work by Yamagishi and Komachi [126] investigates the use of source and target-side contexts by comparing models which encode these contexts via separate context encoders and the ones which utilise the states obtained from the pre-trained baseline system. The latter is referred to as weight sharing and has also been utilised by [65] [75] [68]. This work reports that the models using weight sharing almost always outperform the ones using separate context encoders for a variety of language-pairs and hypothesise this to be due to better regularisation. Further, they also reach the same conclusion as [65] [75] [68], that is, using target-side context is as important as the source-side context.

Dialogue translation is another practical aspect of document translation but is underexplored in the literature. Maruf et al. [67] investigated the challenges associated with translating multilingual multi-speaker conversations by exploring the simpler task of bilingual multispeaker conversation MT. They extracted Europarl v7 and OpenSubtitles2016 to obtain an introductory dataset for the task. They used source and target-side histories in both languages as context in their model where the base architecture comprised two separate sentence-level NMT models, one for each translation direction. The source-side history was encoded using separate Turn-RNNs and then a single source-context representation vector was computed using one of five ways. The target-context representation vector was computed using a language-specific attentional component. The source and target-side histories
were then incorporated in the base model’s decoder separately or simultaneously to improve NMT performance. Their experiments on both public and real-world customer-service chat data [64] demonstrated the significance of leveraging the bilingual conversation history in such scenarios, in terms of BLEU and manual evaluation.

A few approaches use only an attentional component over the context to improve NMT performance. Inspired from [127], Miculicich et al. [75] use three previous sentences as context by employing a hierarchical attention network (HAN) having two levels of abstraction: the word-level abstraction allows to focus on words in previous sentences, and the sentence-level abstraction allows access to relevant sentences in the context for each query word. They combine the contextual information with that from the current sentence using a gate. The context is used during encoding or decoding a word, and is taken from previous source sentences or previously decoded target sentences. Their experiments on Chinese→English and Spanish→English datasets demonstrated significant improvements in terms of BLEU. They further evaluated their model based on noun and pronoun translation and lexical cohesion and coherence, but did not report whether the gains achieved by their model were statistically significant with respect to the baseline.

Maruf et al. [68] presented a novel and scalable top-down approach to hierarchical attention for context-aware NMT, which uses sparse attention to selectively focus on relevant sentences in the document context and then attend to key words in those sentences. The document-level context representation, produced from the hierarchical attention module, is integrated into the encoder or decoder of the Transformer architecture depending on whether the context is monolingual or bilingual. They performed experiments and evaluation on three English→German datasets in both offline and online document MT settings and showed that their approach surpasses context-agnostic and recent context-aware baselines in most cases. Their qualitative analysis indicated that the sparsity at sentence-level allowed their model to identify key sentences in the document context and the sparsity at word-level allowed it to focus on key words in those sentences allowing for better interpretation of their document-level NMT models.

Similar to the two-pass approaches in SMT, Xiong et al. [125] used a two-pass decoder approach to encourage coherence in NMT. In the first pass, they generate locally coherent preliminary translations for each sentence using the Transformer architecture. In the second step, their decoder refines the initial translations with the aid of a reward teacher [9] which promotes coherent translations by minimising the similarity between a sequence encoded in its forward and reverse direction. Their model improved the translation quality in terms of sentence-level and document-level BLEU and METEOR scores where the document-level scores were measured by concatenating sentences in one document into one long sentence and then using the traditional metrics.

Voita et al. [115] proposed a context-aware decoder to refine the translations obtained via the base context-agnostic model similar to [65] [125]. However, while most previous works use the same data to train the model in both stages, they used a larger amount of data to train the sentence-level model and a smaller subset of the parallel data with the context sentences to train their context-aware model. They modify the decoder in the Transformer architecture by allowing the multi-head attention sub-layer to also attend to the previous source context sentences in addition to the current source sentence. They also add an additional multi-head attention sub-layer which attends to the previous target context sentences and the current target. It should be mentioned that they do not mask the future target tokens when performing the multi-head attention with the context unlike previous work [65] [68]. Although, their model performs quite well on targeted test sets that they introduced, it yields comparable performance with the sentence-level model in terms of BLEU. They go on to propose a context-aware model which performs automatic post-editing [114] on a sequence of sentence-level translations and corrects the inconsistencies between the individual translations in context of each other. The main novelty in this work is that the model is trained using only monolingual document-level data in the target language and thus learns to map inconsistent group of sentences to their consistent counterparts. They reported significant improvements for their model in terms of BLEU, targeted contrastive evaluation of several discourse phenomenon [115] and human evaluation.

The works we have mentioned so far have focused on using contextual information to improve performance of NMT in particular. Wang et al. [119] present a more general framework for the
context-aware setting, where a model is given both a source and context containing relevant information and is required to produce the corresponding output. They use document-level MT as one of the applications for their model. They explore three different ways of combining the source and context information in the decoder, that is, concatenating the source and context encoder outputs, adding an extra context attention sub-layer in the decoder before the usual source attention sub-layer, or interleaving the attention sub-layers by replacing the source attention sub-layer by the context attention sub-layer in the middle layers of the decoder. Since their model encodes the source and context separately, they also present a data augmentation technique in which they randomly remove the context information so that the model learns to generate the target given only the source or randomly ask the model to predict the context given the source. The latter lets the encoder learn which parts of the context are relevant to the source sentence. They also employ a focused context attention to encourage better utilisation of the long and noisy context. They showed that their approach yields improvements over both the Transformer and a context-dependent NMT model [129] for the French→English translation task on the TED Talks corpus.

4.3 Promoting positive use of context during training

All the works mentioned up to this point have proposed neural architectures which employ structural modifications in the base sentence-level NMT model to incorporate context information. But it is possible that even with an extended NMT model, not all additional context information is useful and some of it must be ignored for better performance. Zheng et al. [130] introduced a general framework which utilises discriminators to encourage the NMT model to ignore the irrelevant information in the external context. Even though, they did not test their approach with document-level context (or any kind of sequential external context), but the idea could very well be extended to that. Jean and Cho [35] looked at the problem from a learning perspective and designed a regularisation term to encourage an NMT model to exploit the additional context in a useful way. This regularisation term is applied at the token, sentence and corpus levels and is based on pair-wise ranking loss, that is, it helps to assign a higher log-probability to a translation paired with the correct context than to the translation paired with an incorrect one. Employing their proposed approach to train a context-aware model [116] showed that the model becomes more sensitive to the additional context and outperforms the context-agnostic Transformer baseline in terms of empirical evaluation (BLEU scores).

Most recently, Stojanovski and Fraser [105] introduced a curriculum learning approach [8] that leverages oracle information to promote anaphora resolution while training the context-aware NMT model. They propose to initially have the gold-standard target pronouns along with the previous context and source sentence to bias the model to pay attention to the context related to the oracle pronoun. They then gradually remove the oracle pronouns from the data to bias the model to pay more attention to only the context when running into ambiguous pronouns in the source sentence leading to superior anaphora resolution. Their experiments showed that with a higher learning rate, their scheme is unable to beat a context-aware NMT model fine-tuned with only the context in terms of pronoun translation and overall translation quality. However, with a lower learning rate and 25% initial oracle samples, their approach is effective but still lags behind a context-aware NMT model trained with a higher learning rate. Their approach can be extended to other discourse phenomena, provided useful oracles are easily available.

4.4 Shared Tasks in WMT19 and WNGT19

Given the significant amount of work in document-level NMT in the past two years, the Fourth Conference on Machine Translation (WMT19) [6] and the Third Workshop on Neural Generation and Translation (WNGT 2019) [34] introduced document-level translation of news and sports articles respectively, as one of the shared tasks. This opened up remarkable novelties in this domain subsuming approaches for document-level training that utilise wider document context and also document-level evaluation. To aid with this task, WMT19 produced new versions of Europarl, news-commentary, and the Rapid corpus with the document boundaries intact. They also released new versions of
monolingual Newscrawl corpus containing document boundaries for English, German, and Czech. Following suit of WMT19, WNGT 2019 manually translated a portion of the RotoWire dataset, which contains basketball-related articles, to German. Further, they allowed any parallel and monolingual data made available by WMT19 English-German news task and the full RotoWire English dataset.

To have a reliable comparison of document-level MT systems and human performance, two human evaluation setups were introduced in WMT19: Document Rating + Document Context (DR+DC) and Segment Rating + Document Context (SR+DC). In these evaluation setups, human assessors were provided with the whole document context and asked to rate the MT generated individual segments in original sequential order or the MT generated document as a whole comprising the most recently rated segments. WNGT 2019, however, still relied on the BLEU score for evaluation.

Document-level MT systems in WMT19 Among the 153 submissions on news translation task at WMT19, only a few utilised document-level context. The Microsoft Translator introduced three systems, two of which focused on large-scale document-level NMT with 12-layer Transformer-Big systems. For the first system, they combined real document-parallel data with synthetic document-parallel data (produced by putting in document boundaries at random) and created document-level sequences of up to 1000 subword units to train deep transformer models. In addition, they used back-translated documents and further fine-tuning techniques. Their second document-level system comprised a BERT-style encoder trained on monolingual English documents and sharing its parameters with the translation model. They also experimented with second-pass decoding and ensembling techniques to combine the sentence-level systems with the document-level ones. Based on human evaluation, it was found that their document-level systems were preferred over the sentence-level ones.

Popel et al. proposed document-level NMT systems for English-Czech implemented in the Marian and T2T frameworks. Apart from the real document-level data, they also used back-translation to create synthetic data. They extracted context-augmented data from the real and synthetic data by extracting sequences of consecutive sentences with up to 1000 characters along with sentence separators. First, they trained sentence-level baselines using the same setting as [82], followed by a fine-tuning step on the context-augmented data. They experimented with different document-level decoding strategies for Marian and T2T. For Marian, they translated up to three consecutive sentences at once. The final translation hypothesis is selected to be the middle sentence, the shorter sentence of two-sentence context or a single sentence for no context. For T2T, each document was split into overlapping multi-sentence segments consisting pre-context (used to improve the translation of main content), main content and post-context (similar to pre-context). The final translation hypothesis is selected to be the “middle” sentence in a sequence (corresponding to the main content). They were unable to achieve significant improvements over the sentence-level baseline with any framework.

The Cambridge University Engineering Department’s system relied on document-level language models to improve the sentence-level NMT system. They modified the Transformer architecture for document-level language modelling by introducing separate attention layers for inter- and intra-sentential context. The LM could be trained independently of the translation model and act as a post-editor for the sentence-level translations when the context was available. They reported minor improvements in BLEU over strong baselines for their approach.

Stojanovski and Fraser introduced a context-aware NMT model which explicitly modelled the local context (previous sentence), but also took advantage of the larger context (previous ten sentences). The main architecture is similar to what they proposed earlier in and directly includes the previous sentence. For the larger context, they created a simple document-level representation by averaging word embeddings which are added to the source embeddings in the same manner as positional encodings in the Transformer. Their experiments showed that most gains over the baseline come from the addition of the implicit larger context.

España-Bonet et al. enriched the document-level data by adding coreference tags in the
source sentences where the tags were obtained by running a mention-ranking model CoreNLP \cite{61} over the source documents. Since only a few (or no) words in each sentence were annotated, they were unable to obtain significant improvements with their approach and were only able to report some gains when using ensembling.

**Document-level MT systems in WNGT 2019** WNGT 2019 was not as popular as WMT19 in terms of the number of submissions, having received only four for the document translation task. The teams from Edinburgh and Microsoft fine-tuned a pre-trained sentence-level system with in-domain sentence-level data but did not utilise any type of context information. Particularly, Microsoft \cite{74} got significant improvements when fine-tuning with the RotoWire parallel and back-translated English RotoWire dataset, perhaps because of the large model size.

Maruf and Haffari \cite{66} re-used their previously proposed document-level NMT model \cite{68} for the shared task. For the first stage of training, they used sentence-level data from Europarl, Common Crawl, News Commentary v14, the Rapid corpus and the Rotowire parallel data, while for the second stage, they used document-level data from the latter three corpora. They did not perform any further fine-tuning on the Rotowire parallel dataset. For decoding, they ensembled three independent runs of all models using two ensemble decoding strategies by averaging probability distributions at the softmax level with and without averaging the context representations obtained from the different runs. For both translation directions, their hierarchical attention approach surpassed the WNGT baseline by at least 4.48 BLEU score even without in-domain fine-tuning showing that their approach is indeed beneficial even in a large-data setting.

Naver Labs Europe’s system \cite{91} relied heavily on transfer learning from document-level MT (high-resource) to document-level NLG (low-resource). They first trained a sentence-level MT model with all WMT19 data, RotoWire parallel data, and back-translated Newscrawl data. A domain-adapted document-level MT system was then trained via two levels of fine-tuning: (i) fine-tuning the best sentence-level model (according to perplexity on validation set) on all document-level data, (ii) fine-tuning the best document-level model on Rotowire parallel plus back-translated monolingual Rotowire dataset. They ranked first in all tracks of the document-translation task outperforming the baseline by at least 11.76 BLEU score.

The MT results of the submissions on DGT task show that pre-training with back-translated data and re-training document-level MT models on document-level in-domain data leads to drastic improvements. It is also shown that adding structured-data for this specific task does not lead to improvements over the baselines \cite{34}.

### 4.5 Conclusion

It should be noted from the aforementioned works that document-level NMT has flourished significantly in the past three years, mostly due to the improvements in translation quality that we were unable to see with SMT techniques. A recent study by Kim et al. \cite{44} makes a case against the usefulness of document-level context as opposed to the aforementioned works. Firstly, they found that filtering the redundant or irrelevant words in the context does not harm the translation quality measured via BLEU in comparison to when using full context sentences. The document-level NMT models that they experimented with were also unable to outperform a baseline trained on a larger corpus and thus they attributed the improvements gained by the current document-level NMT models to better regularisation. They also point out that the document-level context is seldom utilised for improving discourse aspects in the translation like coreference and lexical choice, as opposed to what has been demonstrated by numerous other works \cite{116, 75, 115}. Further, the improvements are mostly general in terms of adequacy and fluency which, in their view, is not an interpretable way of utilising the document-level context.

In light of this work and the numerous others that we have cited in this survey, we believe that one study is not sufficient to invoke skepticism of the success achieved by document-level NMT. Having said that, to sustain the progress in this domain, it is also necessary that the use of document-level NMT models should be motivated by more in-depth and painstaking analysis.
5 Evaluation

MT outputs are almost always evaluated using metrics like BLEU and METEOR which use n-gram overlap between the translation and reference to judge translation quality. However, these metrics do not look for specific discourse phenomena in the translation, and thus may fail when it comes to evaluating the quality of longer pieces of generated text. A recent study by Läubli et al. [50] contrasts the evaluation of individual sentences and entire documents with the help of human raters. They found that the human raters prefer human translations over machine-generated ones when assessing adequacy and fluency of translations. Hence, as translation quality improves, there is a dire need for document-level evaluation since errors related to discourse phenomena remain invisible in a sentence-level evaluation.

There has been some work in terms of proposing new evaluation metrics for specific discourse phenomena, which may seem promising but there is no consensus among the MT community about their usage. Most of these perform evaluation based on a reference and do not take the context into account. There are also those that suggest using evaluation test sets or better yet combining them with semi-automatic evaluation schemes [24]. More recently, Stojanovski and Fraser [103] propose to use oracle experiments for evaluating the effect of pronoun resolution and coherence in MT. The works introducing document-level MT evaluation approaches are outlined in Table 3. For a more elaborate review of studies that evaluate NMT output specifically in terms of analysing limitations of the translation of contextual phenomena, we direct the reader to [53].

5.1 Automatic Evaluation for Specific Discourse Phenomena

There have been a few works which have proposed reference-based automatic evaluation metrics for evaluating specific discourse phenomena. For pronoun translation, the first metric proposed by Hardmeier and Federico [29] measures the precision and recall of pronouns directly. Firstly, word alignments are produced between the source and translation output, and the source and the reference translation. For each pronoun in the source, a clipped count is computed, which is defined as the number of times the pronoun occurs in the translation output limited by the number of times it occurs in the reference translation. The final metric is then the precision, recall or F-score based on these clipped counts. Werlen et al. [77] proposed a metric that estimates the accuracy of pronoun translation (APT), that is for each source pronoun, it counts whether its translation can be considered correct. It first identifies triples of pronouns: (source pronoun, reference pronoun, candidate pronoun) based on word alignments which are improved through heuristics. Next, the translation of a source pronoun in the MT output and the reference are compared and the number of identical, equivalent, or different/incompatible translations in the output and reference, as well as cases where candidate translation is absent, reference translation is absent or both, are counted. Each of these cases is assigned a weight between 0 and 1 to determine the level of correctness of MT output given the reference. The weights and the counts are then used to compute the final score. Most recently, Jwalapuram et al. [43] proposed a specialised evaluation measure for pronoun evaluation which is trained to distinguish a good translation from a bad one based on pairwise evaluations between two candidate translations (with or without past context). The measure performs the evaluation irrespective of the source language and is shown to be highly correlated with human judgments. They also present a targeted pronoun test suite that covers multiple source languages and various target pronouns in English. Both their test set and evaluation measure are based on actual MT system outputs.

For lexical cohesion, Wong and Kit [120] extended the sentence-level evaluation metrics like BLEU [81], METEOR [4] and TER (Translation Edit Rate) [100] to incorporate a feature that scores lexical cohesion. To compute the new score, they identify lexical cohesion devices via WordNet [15] clustering and repetition via stemming, and then combine this score with the sentence-level one through a weighted average. They claimed that this new scoring feature increases the correlation of BLEU and TER with human judgments, but does not have any effect on the correlation of METEOR. Along similar lines, Gong et al. [21] augmented a cohesion score, based on simplified lexical chain, and a gist consistency score, based on topic model, with document-level BLEU or METEOR (concatenat-
Table 3: Overview of works which introduce techniques to evaluate discourse phenomena in MT. It should be mentioned that many of these do not come with inter-sentence context information.

For discourse connectives, Hajlaoui and Popescu-Belis [26] proposed new automatic and semi-automatic metrics referred to as ACT (Accuracy of Connective Translation) [72]. For each connective in the source, ACT counts one point if the translations are the same and zero otherwise based on a dictionary of possible translations and word alignments. The insertion of connectives is counted manually. The final score is the total number of counts divided by the number of source connectives.

Guillou and Hardmeier [24] studied the performance of automatic metrics for pronouns [29, 77], on the PROTEST test suite [23] of English→French translations and explored the extent to which automatic evaluation based on reference translations can provide useful information about an MT system’s ability to handle pronouns. They found that such automatic evaluation can capture some linguistic patterns better than others and recommend emphasising high precision in the automatic metrics and referring negative cases to human evaluators.
5.2 Evaluation Test Sets

As discussed by [22, 110, 50], a tie of human and machine when both are evaluated on the basis of isolated translation segments cannot serve as an indication of human parity. The current document-level evaluations, for instance the DR+DC evaluation setup in WMT19, are not reliable due to their low statistical power (small sample size) of the rated documents [22]. Furthermore, even the best translation systems may fall behind when it comes to handling discourse phenomena [117]. These highlighted shortcomings can explain the motivation of using targeted test suites which can draw better conclusions of whether or not a machine achieves human parity, in addition to aiding a more in-depth analysis of various aspects of the translation.

A few of the discourse-targeted test sets are contrastive, that is each instance contains a correct translation and a few incorrect ones. Models are then assessed on their ability to rank the correct translation of a sentence in the test set higher than the incorrect translation. Sennrich et al. [93] were the first to introduce a large contrastive test set to evaluate five types of grammatical phenomena known to be challenging for English→German translation. Even though contextual information may be beneficial for such phenomena, this particular test suite did not contain any, due to which we will not be presenting details of this and any similar test suites any further. Inspired by examples from OpenSubtitles2016 [56], Bawden et al. [7] hand-crafted two contrastive test sets for evaluating anaphoric pronouns, coherence and cohesion in English→French translation. All test examples were designed so as to have the particular phenomena in the current English sentence ambiguous such that its translation into French relied on the previous context sentence. Hence, these test sets required a model to leverage the previous source and target sentences in order to improve the said phenomena. Müller et al. [78] presented a contrastive test suite to evaluate the accuracy with which NMT models translate the English pronoun it to its German counterparts es, sie and er while having access to a variable amount of previous source sentences as context. Voita et al. [115] created English→Russian test sets focused on deixis, ellipsis and lexical cohesion as they found that 80% of the inconsistencies in Russian translations were due to these three phenomena. They showed that their context-aware NMT model performed substantially well on these test sets, even though it did not show significant gains over the context-agnostic baseline (Transformer) in terms of general translation quality measured via BLEU. Rios et al. [89, 88] also introduced a test set for word sense disambiguation but it does not contain document-level context. However, they also provide the corpora from which the test set was extracted along with sentence IDs to recover the document context.

WMT19 [6] also went beyond the rating evaluation by providing complimentary test suites. Vojtechová et al. [117] introduce the SAO (Supreme Audit Office of the Czech Republic) WMT19 Test Suite to evaluate the domain-independent performance and document-level coherence in terms of lexical choice when applying MT models trained for the news domain on audit reports. They provide a tri-parallel test set consisting of English, Czech and German documents in addition to documents in Polish and Slovak. Their evaluation experiment on the systems submitted for WMT19 news task showed that although the MT systems could perform quite well for audit reports in terms of automatic MT evaluation, a thorough in-domain knowledge was required for assessing aspects like semantics and domain-specific terminology which was further restricted by using a single reference translation. They conclude that the most practically reliable solution for translation would be an interactive system supporting a domain expert who manually amends the machine terminological choices.

Rysová et al. [90] introduced an English→Czech test suite to assess the translation of discourse phenomena by the WMT19 systems. They manually evaluated the translations and identified translation errors for three document-level discourse phenomena: topic-focus articulation (information structure), discourse connectives, and alternative lexicalisation of connectives. The evaluation results showed that MT systems in WMT19 performed well and almost reached the reference quality on this test suite. They reported document-level MT systems to have the most errors for the category of topic-focus articulation while showing satisfactory performance on the other two phenomena.

The test suites introduced by [21, 86] and [85] were also used to evaluate document-level MT systems at WMT19 but these consist of sentence-pairs only and do not rely on any extra-sentential
context information. We conclude that an evaluation using test suites is feasible but has a restricted scope since it is designed for specific language-pairs and has limited guarantees.

6 Conclusions and Future Directions

In this survey, we have presented a comprehensive overview of the research that has attempted to incorporate extra-sentential context to enhance SMT and NMT systems. We started off by providing an introduction to the statistical and neural MT models along with their evaluation framework. After presenting the foundations, we delved deep into the literature highlighting works in SMT (Section 3) and NMT (Section 4) that have used some form of extraneous context information not accommodated in the sentence-based SMT and NMT models. We wrapped up this survey by identifying the evaluation strategies that have been introduced to measure different aspects of a context-based machine translation.

Despite the progress that document-level MT has seen due to the end-to-end learning framework provided by neural models, there is still a lot that needs to be done not only in terms of better modelling of the context but also context-dependent evaluation strategies. Let us now mention a few of the possible future research directions.

Document-aligned Datasets While there are many popular datasets for MT, most of them consist of aligned sentence-pairs without any metadata. Hence, the first problem that we and other researchers working on the problem of document-level machine translation encounter is to curate datasets for this purpose. Furthermore, it is not necessary that the discourse phenomena we aim to observe actually exist in the current public datasets. This problem further exacerbates when one tries to translate dialogues since datasets, like subtitles, lack speaker annotations. It is high time that the MT community starts investing its efforts in creating such resources (as initiated by WMT19) so that the research process can be standardised with respect to the datasets used.

Explicit Linguistic Annotation If the process of obtaining linguistic annotations could be automated and we could obtain annotations of, for instance, entities in the discourse, it could directly impact the translation of their mentions thus improving lexical cohesion. The translation could also be conditioned on the evolution of entities as they are introduced in the source and target text [37]. We believe annotation of discourse phenomena, for example, coreference or discourse markers, could be beneficial in generating better quality translation outputs more faithful to the source text.

Document-level MT Evaluation From the previous section, it is evident that there is no consensus among the MT community when it comes to the evaluation of document-level MT. Reference-based automatic evaluation metrics, like BLEU and METEOR, which look at the overlap of MT output with a reference, are insensitive to the underlying discourse structure of the text [50]. These are still being used to evaluate MT outputs as they have been the de-facto standard in the community for more than a decade. The proposed document-level automatic metrics (detailed in Section 5) have their own flaws and are not widely accepted. A middle ground should be found between automatic and manual evaluation for MT that could make the process of manual evaluation cheaper and would still be better than the current automatic metrics at evaluating discourse phenomena. Evaluation test sets only resolve a part of the problem as they are mostly hand-engineered for specific language-pairs. Moreover, comparison to a single reference translation is also not a good way to evaluate translation output as it has its own shortcomings. To actually progress in document-level MT, we not only need models that address it but also evaluation schemes that have the ability to correctly gauge their performance.

We hope that this survey presents as a useful resource to highlight the different aspects of the literature in document-level MT and makes it easier for researchers in the area, both old and new, to take stock of where we stand today and identify our accomplishments in this domain in a more standardised way.
References

[1] Ruchit Agrawal, Marco Turchi, and Matteo Negri. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Célia Rico, André Martins, Joachim Van den Bogaert, and Mikel L. Forcada, editors, Proceedings of the 21st Annual Conference of the European Association for Machine Translation, pages 11–20, Alacant, Spain, 2018. European Association for Machine Translation.

[2] Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. Linguistic evaluation of German-English machine translation using a test suite. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 445–454, Florence, Italy, August 2019. Association for Computational Linguistics.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of the Third International Conference on Learning Representations, 2015.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[5] Yehoshua Bar-Hillel. The present status of automatic translation of languages. Advances in Computers, 1:91–163, 1960.

[6] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics.

[7] Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1304–1313, New Orleans, Louisiana, USA, 2018. Association for Computational Linguistics.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, pages 41–48, New York, NY, USA, 2009. ACM.

[9] Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. Discourse-aware neural rewards for coherent text generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 173–184, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[10] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311, 1993.

[11] Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. Dialogue & Discourse, 4:65–86, 2013.

[12] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In International Conference on Learning Representations, 2019.
[13] Jeffrey L. Elman. Finding structure in time. Cognitive Science, 14(2):179–211, 1990.

[14] Cristina España-Bonet and Dana Ruiter. UdS-DFKI participation at WMT 2019: Low-resource (en-gu) and coreference-aware (en-de) systems. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 183–190, Florence, Italy, August 2019. Association for Computational Linguistics.

[15] Christiane Fellbaum, editor. WordNet: An electronic lexical database. MIT Press, Cambridge, Massachusetts, USA, 1998.

[16] George Foster, Pierre Isabelle, and Roland Kuhn. Translating structured documents. In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas, Denver, Colorado, USA, 2010.

[17] Eva Martínez García, Carles Creus, Cristina España-Bonet, and Lluís Màrquez. Using word embeddings to enforce document-level lexical consistency in machine translation. The Prague Bulletin of Mathematical Linguistics, 108:85–96, 2017.

[18] Eva Martínez García, Cristina España-Bonet, and Lluís Màrquez. Document-level machine translation as a re-translation process. Procesamiento del Lenguaje Natural, 53:103–110, 2014.

[19] Eva Martínez García, Cristina España-Bonet, and Lluís Màrquez. Document-level machine translation with word vector models. In Proceedings of the 18th Conference of the European Association for Machine Translation, pages 59–66, Antalya, Turkey, May 2015.

[20] Zhengxian Gong, Min Zhang, and Guodong Zhou. Cache-based document-level statistical machine translation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 909–919, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics.

[21] Zhengxian Gong, Min Zhang, and Guodong Zhou. Document-level machine translation evaluation with gist consistency and text cohesion. In Proceedings of the Second Workshop on Discourse in Machine Translation, pages 33–40, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[22] Yvette Graham, Barry Haddow, and Philipp Koehn. Translationese in machine translation evaluation. CoRR, abs/1906.09833, 2019.

[23] Liane Guillou and Christian Hardmeier. PROTEST: A test suite for evaluating pronouns in machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 636–643, Portorož, Slovenia, 5 2016. European Language Resources Association.

[24] Liane Guillou and Christian Hardmeier. Automatic reference-based evaluation of pronoun translation misses the point. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4797–4802, Brussels, Belgium, 2018. Association for Computational Linguistics.

[25] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 687–698, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[26] Najeh Hajlaoui and Andrei Popescu-Belis. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2, pages 236–247, Samos, Greece, 2013. Springer Berlin Heidelberg.
[27] Michael Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

[28] Christian Hardmeier. *Discourse in Statistical Machine Translation*. PhD thesis, Uppsala University, Sweden, 2014.

[29] Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the Seventh International Workshop on Spoken Language Translation*, pages 283–289, Paris, France, 2010.

[30] Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[31] Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[32] Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[33] B. Hatim and I. Mason. *Discourse and the Translator*. Longman, London, 1990.

[34] Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhiro Sudoh. Findings of the third workshop on neural generation and translation. In *Proceedings of the Third Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong, November 2019. Association for Computational Linguistics.

[35] Sébastien Jean and Kyunghyun Cho. Context-aware learning for neural machine translation. *CoRR*, abs/1903.04715, 2019.

[36] Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135, 2017.

[37] Yangfeng Ji, Chenhao Tan, Sebastian Martstch, Yejin Choi, and Noah A. Smith. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[38] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[39] Shafiq Joty, Francisco Guzmán, Lluís Márquez, and Preslav Nakov. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722, December 2017.

[40] Marcin Junczys-Dowmunt. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy, August 2019. Association for Computational Linguistics.

[41] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation
in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[42] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2nd edition, 2009.

[43] Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing*, pages 2957–2966, Hong Kong, China, November 2019. Association for Computational Linguistics.

[44] Yunsu Kim, Duc Thanh Tran, and Hermann Ney. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China, November 2019. Association for Computational Linguistics.

[45] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[46] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. Asia-Pacific Association for Machine Translation.

[47] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54, Edmonton, Canada, 2003. Association for Computational Linguistics.

[48] Shaohui Kuang and Deyi Xiong. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 607–617, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[49] Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

[50] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, 2018. Association for Computational Linguistics.

[51] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[52] Ronan Le Nagard and Philipp Koehn. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[53] Gideon Lewis-Kraus. The great A.I. awakening. *The New York Times Magazine*, 12 2016.
[54] Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. Assessing the discourse factors that influence the quality of machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 283–288, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[55] Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, pages 577–587, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[56] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 923–929, Portoroz, Slovenia, May 2016. European Language Resources Association.

[57] Annie Louis and Bonnie Webber. Structured and unstructured cache models for SMT domain adaptation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 155–163, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[58] Ngoc-Quang Luong and Andrei Popescu-Belis. A contextual language model to improve machine translation of pronouns by re-ranking translation hypotheses. In Proceedings of the 19th Conference of the European Association for Machine Translation, pages 292–304, Riga, Latvia, May 2016.

[59] Ngoc-Quang Luong and Andrei Popescu-Belis. Machine translation of Spanish personal and possessive pronouns using anaphora probabilities. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 631–636, Valencia, Spain, April 2017. Association for Computational Linguistics.

[60] Valentin Macé and Christophe Servan. Using whole document context in neural machine translation. In Proceedings of the 16th International Workshop on Spoken Language Translation, Hong Kong, China, 10 2019.

[61] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[62] Daniel Marcu, Lynn Carlson, and Maki Watanabe. The automatic translation of discourse structures. In Proceedings of the First North American Chapter of the Association for Computational Linguistics Conference, pages 9–17, Seattle, Washington, USA, 2000. Association for Computational Linguistics.

[63] Daniel Marcu and William Wong. A phrase-based, joint probability model for statistical machine translation. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pages 133–139. Association for Computational Linguistics, 2002.

[64] Sameen Maruf. Document-wide Neural Machine Translation. PhD thesis, Monash University, VIC, Australia, 2019.

[65] Sameen Maruf and Gholamreza Haffari. Document context neural machine translation with memory networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1275–1284, Melbourne, Australia, July 2018. Association for Computational Linguistics.
[66] Sameen Maruf and Gholamreza Haffari. Monash University’s submissions to the WNGT 2019 document translation task. In *Proceedings of the Third Workshop on Neural Generation and Translation*, pages 256–261, Hong Kong, China, November 2019. Association for Computational Linguistics.

[67] Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Contextual neural model for translating bilingual multi-speaker conversations. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 101–112, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[68] Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. Selective attention for context-aware neural machine 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)*, pages 3092–3102, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[69] Laura Mascarell. Lexical chains meet word embeddings in document-level statistical machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 99–109, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[70] Cade Metz. An infusion of AI makes google translate more powerful than ever. *Wired*, 09 2016.

[71] Thomas Meyer and Andrei Popescu-Belis. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France, 2012. Association for Computational Linguistics.

[72] Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, San Diego, California, USA, 2012.

[73] Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Proceedings of the SIGDIAL 2011 Conference*, pages 194–203, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[74] Lesly Miculicich, Marc Marone, and Hany Hassan. Selecting, planning, and rewriting: A modular approach for data-to-document generation and translation. In *Proceedings of the Third Workshop on Neural Generation and Translation*, pages 289–296, Hong Kong, November 2019. Association for Computational Linguistics.

[75] Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, 2018. Association for Computational Linguistics.

[76] Lesly Miculicich Werlen and Andrei Popescu-Belis. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the Second Workshop on Coreference Resolution Beyond OntoNotes*, pages 30–40, Valencia, Spain, April 2017. Association for Computational Linguistics.

[77] Lesly Miculicich Werlen and Andrei Popescu-Belis. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 61–72, Brussels, Belgium, October 2018. Association for Computational Linguistics.

M. Novák and Z. Žabokrtský. Cross-lingual coreference resolution of pronouns. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, pages 14–24, Dublin, Ireland, 2014. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. Computational Linguistics, 30(4):417–449, December 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

Martin Popel. CUNI transformer neural MT system for WMT18. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 482–487, Belgium, Brussels, October 2018. Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. English-czech systems in wmt19: Document-level transformer. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 342–348, Florence, Italy, August 2019. Association for Computational Linguistics.

Andrei Popescu-Belis. Context in neural machine translation: A review of models and evaluations. CoRR, abs/1901.09115, 2019.

Maja Popović. Evaluating conjunction disambiguation on english-to-german and french-to-german wmt 2019 translation hypotheses. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 464–469, Florence, Italy, August 2019. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 470–480, Florence, Italy, August 2019. Association for Computational Linguistics.

C. J. Van Rijsbergen. Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

Annette Rios, Mathias Müller, and Rico Sennrich. The word sense disambiguation test suite at WMT18. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 588–596, Belgium, Brussels, October 2018. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. Improving word sense disambiguation in neural machine translation with sense embeddings. In Proceedings of the Second Conference on Machine Translation, pages 11–19, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Katerína Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. A test suite and manual evaluation of document-level NMT at WMT19. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 455–463, Florence, Italy, August 2019. Association for Computational Linguistics.
[91] Fahimeh Saleh, Alexandre Berard, Ioan Calapodescu, and Laurent Besacier. Naver labs Europe’s systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the Third Workshop on Neural Generation and Translation*, pages 273–279, Hong Kong, November 2019. Association for Computational Linguistics.

[92] Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China, November 2019. Association for Computational Linguistics.

[93] Rico Sennrich. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April 2017. Association for Computational Linguistics.

[94] Rico Sennrich. Why the time is ripe for discourse in machine translation? Presented at the Second Workshop on Neural Machine Translation and Generation, 2018.

[95] Karin Sim Smith. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[96] Karin Sim Smith. *Coherence in Machine Translation*. PhD thesis, University of Sheffield, United Kingdom, 2018.

[97] Karin Sim Smith, Wilker Aziz, and Lucia Specia. The trouble with machine translation coherence. In *Proceedings of the 19th Conference of the European Association for Machine Translation*, pages 178–189, Riga, Latvia, 2016.

[98] Karin Sim Smith and Lucia Specia. Examining lexical coherence in a multilingual setting. In Katrin Menzel, Ekaterina Lapshinova-Koltunski, and Kerstin Kunz, editors, *New perspectives on cohesion and coherence*, pages 131–150. Language Science Press, Berlin, 2017.

[99] Karin Sim Smith and Lucia Specia. Assessing crosslingual discourse relations in machine translation. *CoRR*, abs/1810.03148, 2018.

[100] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231, 2006.

[101] Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. Cued@wmt19:ewc&lms. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 364–373, Florence, Italy, August 2019. Association for Computational Linguistics.

[102] David Steele and Lucia Specia. Predicting and using implicit discourse elements in Chinese-English translation. In *Proceedings of the 19th Conference of the European Association for Machine Translation*, pages 305–317, Riga, Latvia, 2016.

[103] Dario Stojanovski and Alexander Fraser. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[104] Dario Stojanovski and Alexander Fraser. Combining local and document-level context: The LMU munich neural machine translation system at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 400–406, Florence, Italy, August 2019. Association for Computational Linguistics.
[105] Dario Stojanovski and Alexander Fraser. Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland, 19–23 August 2019. European Association for Machine Translation.

[106] Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 375–386, Oslo, Norway, May 2013. Linköping University Electronic Press, Sweden.

[107] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[108] Jörg Tiedemann. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[109] Jörg Tiedemann and Yves Scherrer. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[110] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[111] Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420, 2018.

[112] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA, March 2018. Association for Machine Translation in the Americas.

[113] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[114] Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair for neural machine translation. 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)*, pages 877–886, Hong Kong, China, November 2019. Association for Computational Linguistics.

[115] Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019. Association for Computational Linguistics.
Elena Voïta, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1264–1274, Melbourne, Australia, 2018. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. SAO WMT19 test suite: Machine translation of audit reports. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 481–493, Florence, Italy, August 2019. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting cross-sentence context for neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2826–2831, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. Improving conditioning in context-aware sequence to sequence models. CoRR, abs/1911.09728, 2019.

Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1060–1068, Jeju Island, Korea, 2012. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144, 2016.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. Document-level consistency verification in machine translation. In Proceedings of the 13th Machine Translation Summit, pages 131–138, Xiamen, China, 01 2011. Asia-Pacific Association for Machine Translation.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü, and Qun Liu. Modeling lexical cohesion for document-level machine translation. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI ’13, pages 2183–2189, Beijing, China, 2013. AAAI Press.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. Lexical chain based cohesion models for document-level statistical machine translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1563–1573, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 2019. Association for the Advancement of Artificial Intelligence.

Hayahide Yamagishi and Mamoru Komachi. Improving context-aware neural machine translation with target-side context. CoRR, abs/1909.00531, 2019.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, San Diego, California, USA, 2016. Association for Computational Linguistics.
[128] Frances Yung, Kevin Duh, and Yuji Matsumoto. Crosslingual annotation and analysis of implicit discourse connectives for machine translation. In Proceedings of the Second Workshop on Discourse in Machine Translation, pages 142–152, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[129] Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the transformer translation model with document-level context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 533–542, Brussels, Belgium, 2018. Association for Computational Linguistics.

[130] Zaixiang Zheng, Shujian Huang, Zewei Sun, Rongxiang Weng, Xin-Yu Dai, and Jiajun Chen. Learning to discriminate noises for incorporating external information in neural machine translation. CoRR, abs/1810.10317, 2018.