Document level contexts for neural machine translation

Sadaf Abdul Rauf, François Yvon

Collection Notes et Documents

DOCUMENT LEVEL CONTEXTS FOR NEURAL MACHINE TRANSLATION

Sadaf ABDUL-RAUF and François YVON

DEC 2020

#2020-01

Document level contexts for neural machine translation

Sadaf Abdul Rauf François Yvon

Revision 2, august 2022

Résumé

Ce rapport étudie les méthodes visant à intégrer un contexte discursif étendu en traduction automatique, en se focalisant sur les méthodes de traduction neuronales. Les systèmes de traduction automatique traduisent en général chaque phrase indépendemment de ses voisines, ce qui entraine des erreurs systématiques qui résultent d'un contexte discursif trop étroit. Nous présentons en premier lieu les phénomènes linguistiques qui justifient la prise en compte d'un contexte élargi, en illustrant cet exposé par des exemples typiques d'erreurs de traduction et en évoquant diverses architectures de traduction statistique s'intéressant à ces phénomènes. L'avènement des méthodes neuronales a permis de revisiter ces problèmes, donnant lieu à diverses variantes des architectures neuronales que nous passons ensuite en revue. Nous discutons également des métriques utilisées pour mesurer l'effet de ces contextes élargis sur la qualité des traductions obtenues. Nous concluons en résumant les acquis des travaux actuels et les principales directions de recherche.

Abstract

This report examines methods for integrating an extended discourse context in machine translation, focusing on neural translation methods. Machine translation systems generally translate each sentence independently of its neighbours, which leads to systematic errors resulting from a limited discourse context. We first present the linguistic phenomena that justify taking into account a wider context, illustrating this presentation with typical examples of translation errors and evoking various statistical translation architectures interested in these phenomena. The advent of neural methods has made it possible to revisit these problems, giving rise to multiple variants of the neural translation architectures that we review. We also discuss the metrics used to measure the effect of these extended contexts on the quality

of the resulting translations. We conclude with a summary of current work and the main research directions.

Contents

Progress in machine translation (MT) in recent years has been genuine, to the point where some have claimed MT to reach human parity [Hassan et al., 2018, Popel et al., 2020]. However, most of the effort to date has focused on systems performing translation *on a per sentence basis*, meaning that each sentence translation is performed out of its discursive context. When it comes to translation with contextual information, the area is still under exploration and recent studies have shown even when sentence-level MT evaluations could fail to distinguish human from machine translation, performing evaluations with a discourse context was less favourable for neural machine translation (NMT) [Läubli et al., 2018, Lopes et al., 2020].

The Conference on Machine Translation (WMT) had accordingly started to consider inter-sentential translations in their annual shared task [Guillou et al., 2016]. Neural architectures have shown promising avenues for quality improvement by incorporating contextual information. This has sparked numerous studies in various directions of discourse disambiguation, including word sense disambiguation [Rios Gonzales et al., 2017a, Marvin and Koehn, 2018, Rios et al., 2018, Tang et al., 2018] and pronoun anaphora [Hardmeier et al., 2015, Wong et al., 2020].

Most state-of-the-art Neural Machine Translation (NMT) models [Sutskever et al., 2014, Bahdanau et al., 2014, Vaswani et al., 2017a] use independent sentence pairs for training as well as decoding units, irrespective of the previous sentence or document level context. By doing so valuable contextual information is not considered during translation and hence the translations tend to lack cohesion and coherence, especially in translation consistency at topic and document level [Hardmeier et al., 2013a, Zheng et al., 2020]. A context-aware or document level NMT framework is able to model the *local* context of each sentence, with the awareness of the *global* context of the document.

Correctly handling discourse-related phenomena requires taking the full document context into account and extending the scope of the translation model beyond the sentence level [Hardmeier and Federico, 2010a, Wang et al., 2017b, Bawden et al., 2018b, Lopes et al., 2020]. A context-aware translation model has the capacity to leverage contextual features while translating entire texts, which is especially important when it is critical to achieve a correct translation. This requires the ability to model long-range dependencies between words, phrases, or sentences, which are typically studied in linguistics under the topics of discourse and pragmatics. For a translation system, the capacity to model the context may notably improve certain translation decisions, e.g. a better or most consistent lexical choice [Kuang et al., 2018], or a better translation of anaphoric pronouns [Voita et al., 2018, Bawden et al., 2019]. Such issues are attracted a growing interest from the research communities, as witnessed by the survey papers [Maruf et al., 2019b, Lopes et al., 2020, Ma et al., 2021].

In this document, we present a detailed and comprehensive overview of current works aimed to include a larger context in NMT. We start with a brief history of discourse in machine translation, followed by a thorough introduction of the main discourse phenomena that are typically targeted in MT (Section 1). We feel that a detailed presentation of the discourse phenomenon is important to fully understand the challenges that MT systems have to address. We then give a short overview of NMT, focusing on topics that are especially relevant to describe contextual models: the attention component and the

decoding procedure. A detailed survey of the research works undertaken to incorporate document-level context into translation is presented in Section 3. As pointed out by many authors, automatic metrics such as BLEU [Papineni et al., 2002] and Meteor [Banerjee and Lavie, 2005] fail to properly reward such attempts on document-level phenomena, which calls for better way to evaluate the difficulties of document-level translation. We thus thoroughly discuss in Section 4 specific evaluation schemes that are used to measure improvements in these matters. A last section (5) lists a set of useful resources for document-level MT, before we conclude and discuss further prospects (Section 6).

1 Discourse issues in Machine Translation

1.1 Context in MT: a brief retrospective

Correctly translating with contextual information has long been a topic of primary interest in MT research and has been addressed at each stage of the development of MT technologies, using the resources and modeling capabilities known at the time. For instance, anaphora resolution was a topic of considerable interest for Rule-Based Machine Translation (RBMT) systems, and was typically addressed using transfer-based systems and rules directly implemented into the MT system. Some of these early efforts have been documented a special issue of the Machine Translation journal [Mitkov, 1999]. However, these studies could not help much in Statistical Machine Translation (SMT), as the problems encountered in RBMT outputs were very different from the issues observed in SMT outputs [Le Nagard and Koehn, 2010].

Incorporating discourse-level in SMT remained a difficult problem, as most systems performed translation at the sentence level, meaning that each sentence was processed independently of the other sentences in the same document (both in training and inference)). With respect to SMT, the main focus has been on the following issues: the translation of pronouns [Le Nagard and Koehn, 2010, Hardmeier and Federico, 2010a, Hardmeier et al., 2013b, 2015, Guillou, 2016], of discourse connectives [Meyer, 2011, Meyer and Poláková, 2013, Meyer et al., 2015], the correct disambiguation of word-senses [Carpuat and Wu, 2007, Rios Gonzales et al., 2017a] and verb tenses [Gong et al., 2012], the enforcement of lexical consistency [Carpuat, 2009, Gong et al., 2011], and document-level topic adaptation [Su et al., 2012, Hasler et al., 2014]. While the former problems (pronouns, connective, word senses) can usually be solved with a relatively local context, the latter ones typically require a document-level view. Another important distinction in this respect is between attempts that modified the translation model (phrase-table), and those that targeted more specifically the language model component. This is for instance the case with topic models [Blei et al., 2003] that adapt (cross-lingually) the target language model (LM), thereby enforcing some intra-document consistency. The work of Hardmeier et al. [2013a] and Stymne et al. [2013] goes one step further and makes amendments to the decoding process: following an initialcontext-independent translation of each sentence, a local search procedure improves the overall translation.

When modeling discourse, SMT systems rarely model the discourse phenomena explicitly. This is in sharp contrast with NMT systems, where context sentences can be

modelled in various ways [Maruf et al., 2019b, Kim et al., 2019, Popescu-Belis, 2019, Li et al., 2020]. These, for instance, include *concatenation*, that uses previous sentence as context as done by Tiedemann and Scherrer [2017], *multi-source models* using multiple past sentences with additional encoder by Zhang et al. [2018], *cache-based approaches* that use all previous sentences i.e. in source and target [Tu et al., 2018], *doc-star* transformer based on Star architecture by Guo et al. [2019], Lopes et al. [2020], Flat Transformer by Ma et al. [2020] and many others that we review in Section 3.

In this introductory section, we aim to present and discuss discourse phenomena in detail and to relate them to the relevant works from the MT literature. We will discuss anaphora, deixes, ellipses, discourse connectives, lexical consistency and word sense disambiguation.

1.2 Anaphora Resolution

We start with a thorough account of anaphora, since it has received the most attention in MT research. Anaphora resolution, a.k.a. co-reference resolution[1] refers to the process of establishing connections between references of the same entity.

The most common example of anaphora is the problem of pronoun resolution, which aims to associate a pronominal expression to an nominal entity – the *antecedent* – occurring earlier in the discourse. For instance, in the following example, the pronoun *them* refers to the noun *Catholics* in the first sentence:

> The *Catholics* described the situation as "safe" and "protecting." This made *them* "relaxed and peaceful."

The antecedent can also occur after the *cataphoric* pronoun, as in "*If she is in town, Mary will join us for dinner*". An important constraint is the fact that pronouns must agree with their antecedent; depending on the language, agreement can concern various grammatical features such as number, gender, case, etc.

Some pronouns also accept other types of antecedents: *event reference* pronouns, such as *it* in English,[2] can the refer to an event, which can be expressed by a verb, verb phrase (VP), a clause, or any longer passage of text, as in "*He lost his job. It came as a total surprise*". For cases of non-pleonastic pronouns one may encounter *intra-sentential anaphora* (pronoun and refering entity occurring in the same sentence) or *inter-sentential anaphora*, where they are in different sentences.

Coreference resolution has been a widely researched problem in natural language processing (see eg. [Poesio et al., 2016] for a detailed account), but its integration into Machine Translation has been lacking, mainly due to the requirement to process several consecutive sentences, whereas the traditional focus of MT has been the processing of individual sentences (in training and in testing).

[1] Anaphora and co-reference resolution are identical for the case of pronouns, but they differ in other contexts.

[2] *Pleonastic it* does not even refer to anything, but is required syntactically (e.g. "*I have an umbrella. It is raining*").

A typical translation issue that might imply some sort of anaphora resolution is the translation of pronouns between languages having a grammatical gender (as in most European languages) and those which are gender-neutral, like English.[3] Take, for instance, the translation of the English pronoun *it* into French: if the reference has masculine grammatical gender in French, then its translation is the masculine pronoun (ramme.g., *il* in French), while if it refers to a feminine entity then the translation should also be feminine (e.g. *elle* in French). There are of course more options to consider as *it* could as well be pleonastic and be translated with the masculine or refer to an event and be turn into a neutral French pronoun such as *ce, ça,cela*. This is illustrated in Table 1, which displays examples taken from [Le Nagard and Koehn, 2010]. In this examples, the translation of *it* depends upon the previous sentence and it is essential to connect it to the entity *window* in the previous sentence to get the correct translation.

The window is open. **It** is blue.	La fenêtre est ouverte. **Elle** est bleue.	OK
the window is open. **It** is black.	La fenêtre est ouverte. **Il** est noir.	KO
The oven is open. **It** is new.	Le four est ouvert. **Elle** est neuve.	KO
The door is open.**It** is new.	La porte est ouverte. **Elle** est neuve.	OK

Table 1: Typical pronoun translation errors due to lack of proper anaphora resolution.

Elaborating further, the issue of mapping pronouns across languages is made difficult by many factors such as differences in formality, number, case, gender, etc., due to language specific restriction on pronoun usage and placement. Indeed, translating *it* in French, with only three possible options (*il*, *elle*, and *cela*) seems rather simple compared to the translation into German, where the possible choices are *er*, *sie*, *es*, *ihn*, *ihr*, *ihm* to also account for the case differences. This is a difficult choice, as choosing the correct pronoun may require discourse and linguistic information as well as world knowledge.

Translating from pro-drop languages (such as Spanish and Czech) is even more challenging [Loáiciga et al., 2017]. In such cases, the person and number are expressed by morphological inflections of the main verb, so there is no overt pronoun in the source side input. This means that pronoun generation in the target language may require combining multiple information, eg. the person and number would be cued by the verb and the gender by previous sentences.

With most MT systems translating on a sentence-per-sentence basis, the inter-sentential anaphoric pronouns will be translated without knowledge of their antecedent, which means that pronoun-antecedent agreement cannot be guaranteed. Also, the cases of pronouns having several translation options are most likely to be wrongly translated by an MT system that is not aware of these constraints.

While there was some limited work in this direction for rule-based systems [Mitkov et al., 1995], it has become one of the most worked-upon discourse phenomena for SMT and NMT, also fostering the design of dedicated shared tasks and evaluation metrics.

[3]Similar problems arise when the grammatical gender of a given concept differs between languages - "sun" is "die (fem) Sohne" in German, but "le (masc) soleil" in French.

For pronoun translation Hardmeier and Federico [2010b] first proposed a metric based on precision and recall, followed by Hajlaoui and Popescu-Belis [2013] proposing a reference based evaluation metric (APT). Several shared tasks have been organised on the subject [Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017]. Guillou [2016] presents a semi-automatic pronoun evaluation test suite (PROTEST for English to French) consisting of manually selected anaphoric, cataphoric, event, textual, pleonastic, speaker and addressee reference pronoun examples. It remains an active research topic for NMT, e.g. Bawden et al. [2018a], Müller et al. [2018], Jwalapuram et al. [2019] present contrastive test pairs for pronoun evaluation, cohesion and coreference. We discuss these in Section 4.

1.3 Deixes

Deixes are referential expressions (that may be words or grammatical phrases), whose interpretation in an utterance depends on extralinguistic circumstances. Such expressions may depend on factors such as the time of speech, the identity of the speaker and/or listener and their spatial location, body language and facial expressions etc. Deictics are sometimes compared to anaphoras but anaphoras are interpreted based on the linguistic terms that precede them unlike context of speaker or situation [Huls et al., 1995]. Typical examples include:

- *Personal deixes:* "*I, we, you*" involve first and second person pronouns. Again due to agreement constraint, keeping track of the gender / number of the speaker / addressee in a conversion matters when translating to languages where these categories are marked;
- *Temporal deixis:* "*he lives in Amsterdam*"; include sense of time and use both the tense system and temporal adverbs or expressions. Likewise, keeping track of the temporal organization is critical to transfer the right tense mark into the target language;
- *Place deixis:* "*there*" or "*here*"
- *Spatial deixis:* "*this, that, which*"; they typically include pointing gestures;
- *Discourse deixis:* "*that's a good question*" makes reference to past of future utterances in the same discourse;

Based on an analysis of a subtitle corpus, [Voita et al., 2019a] mentions deixis as one important problem for NMT and designs a dedicated test to evaluate the progress in solving these issues. They also refined a context-agnostic NMT model by fine-tuning with parallel documents. Their evaluation of diexis related issues showed significant improvements through contrastive evaluation.

Voita et al. [2019b] works on personal, place and discourse deixis for English to Russian translation, differentiating between informal and formal "you" (Latin "tu" and "vos") in translations into Russian. This work used sequence-to-sequence models trained on

monolingual documents to correct contextual inconsistencies of sentence-level translations. Using this two-pass decoding strategy, they achieve substantial improvements in diectic translation and report deixis scores to be less sensitive to the amount of training data and noise. The found that most errors in their annotated corpus were related to personal deixis, specifically gender marking in the Russian translation, and to distinctions between informal and formal "*you*". For deixis, their model achieves the final quality quite quickly; for the other issues considered in this work (lexical consistency and ellipsis) , it needs a large number of training steps to converge.

1.4 Ellipsis

Ellipsis is the omission of one or more words from a clause that are nevertheless understood in the context of the sentence. Verb phrase (VP) Ellipsis constitute the most prevalent type of ellipsis in languages and are also the hardest discourse phenomenon to capture in machine translation [Voita et al., 2019a]. For instance, in the following example (a), the meaning of the second part of the sentence is understood as "ICC reversed the decision too" by echoing the VP *reversed* that occurs in the first part of the sentence, even though this VP is omitted in the second part [Krovetz, 1998].

a The decision was *reversed* by the FBI, and the ICC did too.

b You might *do it*, but I won't $\emptyset_{\text{do it}}$.

Handling ellipsis can help to considerably improve consistency in translation. Ellipses become a problem when the source and target languages do not have the same types of ellipses and/or when the syntax is effected due to elision [Hamza, 2019, Voita et al., 2019a]. A thorough analysis of ellipsis phenomena and their impact on translation errors is given in [Hamza, 2019] and the first example in Table 2 shows this phenomenon with an elliptic source sentence and the corresponding translation from a human translator and the MT system Systran. This is the case of the example showing a *to*-triggered ellipse where the segment "make myself agreeable to young McClure" is omitted. This example obviously did not pose a problem for the human translator who identified the presence of an antecedent and restored it by the clitic pronoun *y*. By contrast, the MT simply repeated the original statement word to word, making the translation unacceptable in this context. The second example in Table 2 displays two types of elliptic errors encountered in English to Russian translation which include wrong morphological forms and VP ellipsis. Morphological error occurs when a noun phrase is incorrectly marked as subject and VP ellipsis error is when the corresponding VP ellipsis is not there in the Russian translation.

An early attempt to address such problems is presented in JETR [Yoshii, 1987], a rule-based Japanese to English translation system that handled ungrammatical sentences and used chain of result states for context analysis to resolve ellipses and pronoun references. Recently, Voita et al. [2019b,a] have built NMT models handling ellipsis and show performance improvements through human and constrastive evaluations.

Source:	**Human Translation:**
	Soyez gentil avec le jeune McClure.
	Je n'y manquerai pas
Make yourself agreeable to young MacClure.	**Machine Translation (Systran):**
I won't fail to Ø.	Faites-vous plaisir au jeune macclure.
	Je ne manquerai pas.

Ellipses in human and machine translation (borrowed from [Hamza, 2019]).

(a)			
EN	You call her your friend but have you been to her home ? Her work ?		
RU	Ты называешь её своей подругой, но ты был у неё дома? Её **работа**?	OK	
			wrong morphological form
EN	You call her your friend but have you been to her home ? Her work ?		
RU	Ту называыеш' ыеыо своыеы подругоы, но ты был у неые дома? Ыеыо **работа**?	KO	
(b)			
EN	Veronica, thank you, but you **saw** what happened. We all **did**.		
RU	Вероника, спасибо, но ты **видела**, что произошло. Мы все **хотели**.	OK	
			wrong VP ellipsis
EN	Veronica, thank you, but you **saw** what happened. We all **did**.		
RU	Вероника, спасибо, но ты **видела**, что произошло. Мы все **хотели**.	KO	

Table 2: Ellipsis translation discrepancies reported by [Voita et al., 2019a] (a) wrong morphological form, incorrectly marking the noun phrase as a subject. (b) correct meaning is "*see*", but MT produces хотели ("*want*").

1.5 Discourse connectives

Discourse connectives(DCs) are the cohesive markers that join clauses in texts and indicate discourse relations between adjacent spans. These include words such as: "*although*", "*while*", "*however*", "*since*", "*for example*", "*in addition*", etc. [Meyer and Webber, 2013]. If the translated sentence uses a wrong connective, the resulting translation may be ambiguous or fully incomprehensible. Discourse relations may also be present implicitly

(inferred from the context), in the source, yet needed or useful in the target [Meyer and Webber, 2013].

DCs are highly ambiguous in their usage. Focusing on three DCs ("*since*" and "*while*" in English and "*alors que*" in French), [Meyer, 2011] note that "*while*" may convey temporality or contrast, or both at the same time; likewise, "*since*" can have a causal and or temporal meaning or even both. They further discuss possible disambiguation strategies, using a parallel corpus to collect disambiguated instances.

Dcs are not only ambiguous, but their use vary across languages, makes them difficult for human translators, and even more so for MT. For instance, [Meyer and Poláková, 2013, Meyer and Webber, 2013] report that explicit discourse connectives in source language are not always translated to comparable words or phrases in the target language; in their study, human translators did not translate explicit discourse connectives in about 18% of all the casesfor English to French and German translation. This is illustrated in the first example in Table 3, where the causal "*as*" is not explicitly translated by human translators, for French causality is implicitely conveyed by the relative clause "qui lui était inconnue " and for German since a translation equivalent existed, it is realized by means of a preposition, "*wegen*" ("*because of*"). The second example shows wrong MT due to incorrect rendering as "*since*" is used in causal sense in the English sentence, while the translation "*depuis que*" is in temporal sense.

	Human Translations(FR,DE) [Meyer and Webber, 2013]
EN:	The man with the striking bald head was still needing a chauffeur **as** the town was still unknown to him.
FR:	L'homme, dont le crâne chauve attirait l'attention, se laissa conduire **dans la ville qui** lui était encore étrangère.
DE:	Der Mann mit der markanten Glatze liess sich **wegen** der ihm noch fremden Stadt chauffieren.
	Machine Translation [Meyer, 2011]
EN:	Finally, and in conclusion, Mr President, with the expiry of the ECSC Treaty, the regulations will have to be reviewed **since** I think that the aid system will have to continue beyond 2002.
FR:	Enfin, et en conclusion, Monsieur le président, à l'expiration du traité CECA, la réglementation devra être revue **depuis que** je pense que le système d'aides devra continuer au-delà de 2002.

Table 3: Examples of translation of discourse connectives.

[Meyer et al., 2012] tries to improve SMT with automatically annotated source language discourse connective labels and factored translation models [Koehn and Hoang, 2007].

Birch et al. [2007] and Meyer and Popescu-Belis [2012] used sense labels for the automatic disambiguation of discourse connectives for English to French translation. Using BLEU, they report a slight improvement in the translation of connectives. Meyer and Poláková [2013] later presented a SMT system with manually annotated DC's for English to Czech translation. They show slight improvements in their discourse aware system through automatic scoring, error analysis and human evaluation. Yung et al. [2015] and Li et al. [2017] describe the conversion of implicit DCs in Chinese language to explicit DCs in English through a cross-lingual annotated and aligned corpus for a SMT system which showed significant improvement through manual evaluation.

1.6 Lexical consistency

Consistency of a text is achieved via the conjunction of multiple cohesion-building devices, such as repetitions, collocations, tense or pronoun use, etc. The most studied lexical cohesion devices are *reiterations* (repetition of words), possibly associated to some morphological variation, *collocations* (regular co-occurrence of words, lemmas or terms), and more generally the succession of semantically related words. All these cues are overt lexical indicators of cohesion and are key for lexical consistency. When grouped together, these occurrences form *lexical chains* [Morris and Hirst, 1991]. Guillou [2013] present an analysis of human authored translations and conclude that human translators use lexical consistency to support the important parts in a text. She reports proper nouns to have a high lexical consistency of translation across genres. For example, the noun " *Dracula*" in a particular story has more importance than the common noun "*driver*": every occurrence of the former is translated as "*Dracula*" in French, while latter is translated as either "*(le) chauffeur*" or "*(le) conducteur*" as illustrated in Table 4.

EN	... and the **driver** said in excellent German
FR	... **Le conducteur** me dit alors, en excellent allemand
EN	Then the **driver** cracked his whip
FR	Puis **le chauffeur** fit claquer son fouet

Table 4: EN-FR translation examples demonstrating lexical consistency [Guillou, 2013]

In machine translation, lexical consistency needs to be reproduced in the target text, requiring to enforce some sort of global constraints spaning over long stretches of texts: a paragraph, a section, or even a complete document. One such constraint would notably ensure that repeated words in the source text are translated consistently in the target. Carpuat [2009] claims that "the one translation per discourse constraint" in SMT can potentially improve translation quality. Carpuat and Simard [2012] further analyse SMT translated documents and report that the MT output tends to have more incorrect repetition than human translation, especially when the MT model is trained on small corpora. They report consistency, even without document knowledge and attribute translation in-

consistencies to inherent SMT translation errors. Several follow-up sudies have reported improvements by incorporating lexcial chains in SMT: Xiong et al. [2013] use lexical chain based cohesion models for Chinese to English document-level SMT and report improvement in lexical cohesion. Their models reward hypothesis with chain words and use chain word translation probabilities. Gong et al. [2015] also report improvements in document-level SMT via lexical chains and propose an evaluation metric using cohesion to evaluate text cohesion of SMT models using lexical chains and gist consistency at topic level. Mascarell [2017] used lexical chains in their SMT systems using word embeddings to evaluate semantic similarity and integrate a document-level context in the MT decoder. More recently, this issue is documented in [Voita et al., 2019b,a] who notably study the reiteration of named entities in their context-aware NMT systems (see above Section 1.3).

1.7 Word Sense Disambiguation

Word-sense disambiguation (WSD) deals with the determination of a correct meaning or sense of a word in a given sentencial context. In other words, it can be defined as the process of figuring out the correct meaning of a word with multiple potential senses based on the analysis of its context. *Homographs* are another form of WSD, which have the same surface form with different meanings. For instance, in the sentences below (borrowed from [Marvin and Koehn, 2018]), the word "*like*" represents four different senses. When ambiguous words are not translated with the correct sense, the resulting translations may be incomprehensible or even misleading.

1. **similar:** Her English, *like* that of most people here, is flawless.
2. **speech:** We were *like*, what do we do?
3. **enjoy:** Of the youngers, I really *like* the work of Leo Arill.
4. **request:** I would *like* to be a part of them, but I cannot.

An important question is how much *WSD* affects the discourse in source and target language in MT? Translation quality has an inverse relationship to the number of senses of a word. Carpuat and Wu [2007] showed that performance of word-level translation decreases as the number of senses for each word increases. For their Chinese-English MT, adding *WSD* to a baseline SMT system performing phrasal multi-word disambiguation proved to be useful. Another illustration is in Table 5, where we display a comparison of sentence-level versus document-level translation for Arabic-English MT taken from [Zhang and Ittycheriah, 2015]. Arabic sentences are written in roman scripts, where *mrsy* represents *Morsi* in English. We see that in the sentence-level translation *mrsy* is replaced with *Thank*. On the contrary, in document-level translation with access to a local context, the correct translation *Morsi* is used.

Embeddings are now frequently used to incorporate word sense: Rios Gonzales et al. [2017b] report improvement in NMT performance by (a) using sense embeddings either

AR	Alrys AlmSry AlmEzwl mHmd **mrsy**	
	ysf nfsh binh rys Aljmhwryp	
EN	The deposed Egyptian president Mohamed **Morsi**	
	describes himself as the president of the republic	
		sentence level translation
AR	**mrsy** ytHdY AlqADy fy mHAkmth	
	bthmp Alhrwb mn Alsjn	
EN	**Thank** you defy the judge in his trial on	
	charges of escaping from prison	
AR	Alrys AlmSry AlmEzwl mHmd mrsy	
	ysf nfsh binh rys Aljmhwryp	
EN	The deposed Egyptian president Mohamed Morsi	
	describes himself as the president of the republic	
		document level translation
AR	mrsy ytHdY AlqADy fy mHAkmth	
	bthmp Alhrwb mn Alsjn	
EN	Morsi defies the judge in his trial on	
	charges of escaping from prison	

Table 5: Comparison of sentence versus document level translation. (from [Zhang and Ittycheriah, 2015]

as additional input to the encoder or (b) by extracting structured lexical chains from the training data for English to German and German to French language directions. Liu et al. [2018a] built their NMT system using context-aware embeddings for English-French, English-Chinese and English-German and showed improvements in BLEU scores with respect to a baseline model. Recently, Zouhar et al. [2020][4] studied *markables* and their impact in document level translations for Czech and English for the news, lease and audit domains in WMT20 submissions. They defined *markables* as the expressions bearing most of the documents meaning, fulfilling one of the three cases: (1) term was translated into two or more different ways within one document. (2) term was translated into two or more different ways across several translations. (3) two or more terms were translated to a specific expression in one document but have different meanings. The *markables* were identified by annotators for each domain. They reported that MT systems make specific errors in markables, which no human translator would do.

[4] https://github.com/ELITR/wmt20-elitr-testsuite

Discouse phenomena	Language pair(s)	Reference	MT MT
Deixes	EN→RU	[Voita et al., 2019b], [Voita et al., 2019a]	NMT
Anaphora Resolution	{CS, FR, DE, ES} → EN	[Mitkov et al., 1995]	SMT
	EN→FR	[Le Nagard and Koehn, 2010]	SMT
	EN→DE	[Guillou, 2016]	SMT
	EN→FR	[Bawden et al., 2018a]	NMT
	ZH→EN, ES→EN	[Miculicich et al., 2018]	NMT
	EN→DE	[Maruf et al., 2019a],[Junczys-Dowmunt, 2019]	NMT
	EN→FR, EN→DE	[Lopes et al., 2020]	NMT
Ellipsis	JP→EN	[Yoshii, 1987]	RBMT
	EN→RU	[Voita et al., 2019b],[Voita et al., 2019a]	NMT
Discourse Connectives	NL→EN, DE→EN	[Birch et al., 2007]	SMT
	EN→FR	[Meyer, 2011],[Meyer et al., 2012], [Meyer and Popescu-Belis, 2012]	SMT
	EN→FR, EN→CZ	[Meyer et al., 2011]	SMT
	EN→CZ	[Meyer and Poláková, 2013]	SMT
	EN→FR, EN→DE	[Meyer and Poláková, 2013]	SMT
	ZH→EN	[Yung et al., 2015],[Li et al., 2017]	SMT
Lexical Consistency	EN→FR	[Guillou, 2013]	SMT
	ZH→EN	[Xiong et al., 2013]	SMT
	ZH↔EN	[Gong et al., 2015]	SMT
	EN→RU	[Voita et al., 2019b], [Voita et al., 2019a]	NMT
Word Sense Disambiguation	ZH→EN	[Carpuat and Wu, 2007]	SMT
	AR→EN	[Zhang and Ittycheriah, 2015]	SMT
	DE→EN	[Mascarell, 2017]	SMT
	EN→DE, DE→FR	[Rios Gonzales et al., 2017b]	NMT
	EN→FR	[Marvin and Koehn, 2018]	NMT
	EN→FR, EN→ZH, EN→DE	[Liu et al., 2018a]	NMT
	DE→EN	[Tang et al., 2018]	NMT
	EN→CZ	[Zouhar et al., 2020]	NMT

Table 6: Overview of discourse phenomena in SMT and NMT

1.8 Conclusion

Incorporating discourse in MT is a hard problem, often specifically due to the sentence level approach with which MT has traditionally been modelled. This results in incoherent translations as the translation is unable to take into account the local sentence and the global document context. We presented in this section an overview of the main discourse phenomenon that MT research has focused on and the way SMT and NMT have tried to model these phenomena. A tabulated summary is given in Table 6. We have paid special attention to first explain the phenomenon linguistically and then give a brief account on the research works which tried to address them.

The increase in computational power has enabled researchers to exploit neural networks of ever increasing complexity and to build models relaxing these independence assumptions. In fact, apart from their general building blocks, SMT and NMT differ in discourse modelling in the sense that SMT studies were mostly focused on modeling

discourse phenomena explicitly, whereas NMT uses context sentences directly using different modelling techniques, for example using source side context [Wang et al., 2017a, Zhang et al., 2018, Voita et al., 2018, Yang et al., 2019] and incorporating target side context along with [Tu et al., 2018, Kuang et al., 2018, Xiong et al., 2019, Voita et al., 2019a, Zheng et al., 2020]. The next sections explore these methods and architecture in more detail.

2 Neural Machine Translation

In this section, we briefly introduce the main concepts of neural machine translation (NMT), with an emphasis on concepts whose understanding is important for accurately describing methods that handle context. We focus on the concept of attention, and on decoding algorithms, which are specifically targeted by developments in document-level MT. A much more complete presentation of NMT is in [Koehn, 2020, Stahlberg, 2020].

NMT, since its advent, has been through a rapid development process, with the introduction of new paradigms and new approaches, achieving new milestones and ultimately attaining far better accuracy levels than the prevailing statistical machine translation (SMT) approaches. The works of Kalchbrenner and Blunsom [2013], Cho et al. [2014a], Bahdanau et al. [2014] could be marked as the first landmarks towards successfully training end-to-end neural MT systems under an encoder-decoder framework. They departed from earlier attempts [Bengio et al., 2003, Zamora-Martinez et al., 2010, Le et al., 2012, Schwenk, 2012, Cho et al., 2014b, Devlin et al., 2014] at integrating neural components in the MT pipeline, which did not consider the end-to-end training for MT. Among the early adopters, Systran [Crego et al., 2016] and Google [Wu et al., 2016] quickly deployed their own NMT systems, reporting large improvements over then existing state-of-the art SMT models.

This section first reviews the general principles of recent NMT encoder-decoder architectures, with a special emphasis on the Transformer model, and its various extensions aiming at coping with large contexts. We then expose the details of the main decoding algorithm, based on approximative search.

2.1 Some general principles

NMT differs in many ways from the previous generation of models embodied in the IBM models of Brown et al. [1990, 1993], then in their successive evolutions from the phrase-based statistical models of Och and Ney [2002], Koehn et al. [2003], Koehn [2010] to the hierarchical models of Chiang [2005], culminating in the Moses system [Koehn et al., 2007].

However, the statistical general principle of these architectures remains the same: that of producing in the target language the best possible translation $\mathbf{e}$ of the input source sentence $\mathbf{f}$, according to a probabilistic decision rule:

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\mathbf{e}|\mathbf{f}), \tag{1}$$

where the parameters ($\boldsymbol{\theta}$) are estimated from sentence-aligned parallel corpora.

Learning such a distribution is unrealistic if one consider complete sentences as the random variable in the models. Statistical MT models resort to an approximation taking the following form:[5]

$$p(\mathbf{e}|\mathbf{f}) = \sum_{\sigma} \prod_{t} p_{\boldsymbol{\theta}}(\mathbf{e}_{\sigma,t}|\mathbf{f}_{\sigma,t}) \approx \max_{\sigma} \prod_{t} p_{\boldsymbol{\theta}}(\mathbf{e}_{\sigma,t}|\mathbf{f}_{\sigma,t})$$

where σ represents all the possible ways of reordering $\mathbf{f}$ and segmenting it into segments that are synchronous with those of $\mathbf{e}$, and $\mathbf{f}_{\sigma,t}$ (resp. $\mathbf{e}_{\sigma,t}$) represents the variable-length segments (or *phrases*[6]) that are the basic building blocks of the model. Training aims at estimating parameters $\theta_{e,f}$ which express the translation probabilities of segment e into segment f, as well as a number of auxiliary parameters (to evaluate the distortion, the probability of target sequences) which are useful to solve (albeit in an approximate way) the program defined by equation (1). A prerequisite is then to align word by word (or segment by segment) the sentences that make up the parallel corpus used to learn these patterns.

The revolution introduced by neural methods is essentially to make the following factorisation of the conditional probability distribution tractable:

$$p_{\boldsymbol{\theta}}(\mathbf{e}|\mathbf{f}) = \prod_{t} p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}).$$

This simply states that the probability of each target word is generated conditioned on the current prefix of the target sentence ($\mathbf{e}_{<t}$) and on the entire source sentence ($\mathbf{f}$). The manipulation of such distributions is made possible by transforming discrete contexts ($\mathbf{e}_{<t}, \mathbf{f}$) and words $\mathbf{e}_t$ in the vocabulary into continuous representation spaces. This means that each word $\mathbf{f}_t$ (and accordingly each context or each sub string) is associated with a large numerical vector which carries all the useful information about that word or context.

Such a factorization suggests that the resolution of the program (1) can be solved by generating the words from left to right according to the following greedy procedure:

$$p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{f}) = \operatorname{argmax} \prod_{t} p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}). \tag{2}$$

Further information on decoding algorithms for neural translation is given in section 2.2.

The estimation of $\boldsymbol{\theta}$ is performed by maximizing the conditional log-likelihood (or cross-entropy) $\sum_t \log p_\theta(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f})$ accumulated over a large set of sentences, resulting in a complex optimization program that is solved in an approximate manner by generic large numerical optimization algorithms.

The main neural architectures are mainly distinguished according to the way the encoding of the conditioning context is performed: in the *recurrent* neural architectures, on which the first NMT systems are based, this encoding is performed by a recurrent

[5] This presentation is deliberately simplified (), see a detailed account in [Koehn, 2010].

[6] To avoid the confusion with linguistic phrases, we use here the more neutral term of "segments".

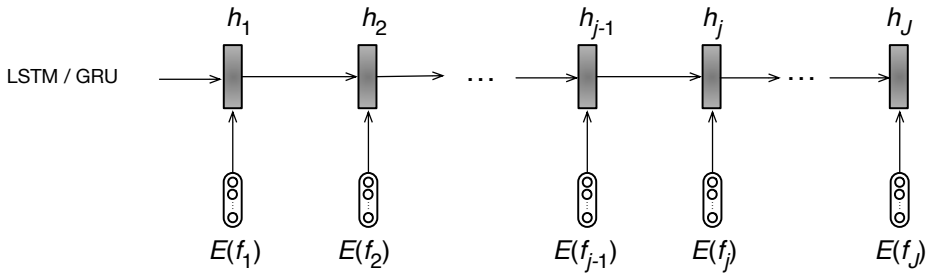

Figure 1: Encoding of the source sentence. $E(\mathbf{f}_j)$ represents the embedding of token $\mathbf{f}_j$.

neural network (RNN) [Cho et al., 2014c], that was complemented by an *attention model* in [Bahdanau et al., 2014]. More recent architectures get rid of the recurrent component and rely on more general sophisticated versions of the attention component [Vaswani et al., 2017a]. Learning such systems becomes equivalent to that of learning a classifier that should produce the most likely class (the next word) given a context encoding a rich information, potentially at a long distance.

2.1.1 Recurrent encoder-decoder architectures

In the recurrent architectures initially proposed by Cho et al. [2014c,a] the context $(\mathbf{e}_{<t}, \mathbf{f})$ is seen as a sequence consisting of a series of source words juxtaposed with a sequence of target words. Each source word $\mathbf{f}_j$ is associated, after a more or less complex encoding (monodirectional or bidirectional, single or multi-layered), with a multidimensional vector h_j, the source token end symbol being associated with a distinguished state h_J (see Figure 1).

The decoder combines h_J with the current target prefix to compute s_t, a representation of the prediction context for word $\mathbf{e}_t$, which is then predicted according to $p_{\boldsymbol{\theta}}(\mathbf{e}_t|s_t)$. This simplistic approach, which summarizes the entire source sentence in a single vector h_J, was quickly augmented by Bahdanau et al. [2014], whose architecture takes into account a richer source context. This position-dependent context, denoted c_j, is computed at each step as a convex linear combination of the vectors h_j according to:

$$c_j = \sum_i \alpha_{ji} h_i, \text{ with } \alpha_{ji} = \text{softmax}(h_i^T W_a s_{j-1}).$$

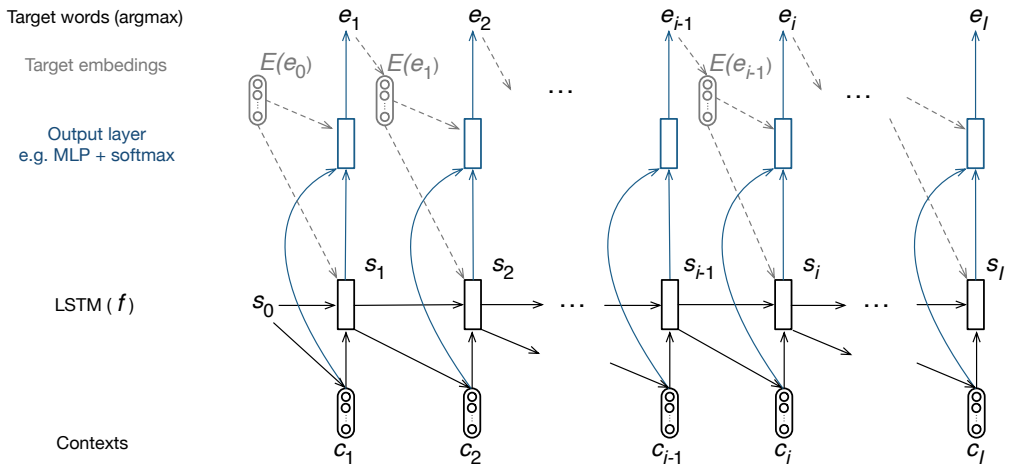

Figure 2: Decoding of the source sentence.

The coefficients α_{ji} measure a normalized affinity between the current target context summarized in s_{j-1} and each of the source words $\mathbf{f}_i$ based on its representation h_j. The set of vectors $[\boldsymbol{\alpha}_i], i = 1\dot{I}$ forms a stochastic matrix that quantifies, for each computation step of the translation process, the local importance of each source word in the decision. Decoding then relies on $p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}) = p_{\boldsymbol{\theta}}(\mathbf{e}_t|s_t, c_t)$ (see Figure 2).

If the attention matrix can be viewed as a (probabilistic) alignment matrix between target and source words, many subsequent works (for example [Cohn et al., 2016, Koehn and Knowles, 2017, Ghader and Monz, 2017] for RNN models, or [Li et al., 2019, Ding et al., 2019] for *Transform* models) have shown that in the absence of additional constraints on the structure of this matrix, the values it contains differ quite strongly from the values produced by classical probabilistic aligners such as IBM models [Brown et al., 1993, Och and Ney, 2003]. Based on these observations, many attempts have been made to either supervise the attention component by reference alignments, or to constrain the attention matrix to look more like an alignment matrix (by limiting the fertility, enforcing distortion constraints, etc).

2.1.2 Attention-based architectures: Transformers

Recurrent architectures pose a major problem for learning: the calculation of the loss function that serves as the basis for estimating the parameters must be performed sequen-

tially, since its value upon processing word $\mathbf{e}_t$ recursively depends on the representations calculated for previous words: s_t depends on s_{t-1}, which depends on s_{t-2} as well as on all the past translation decisions. Recurrent architectures also implicitly rely on the assumptions that recent words are more important when computing representations than more distant ones, an issue that is only partly mitigated with more complex recurrent cells such as LSTMs or GRUs.

The Transformer architecture of Vaswani et al. [2017b] aims to overcome these problem, and replaces the recurrent components (in the source and target) by generalized attentional modules, while keeping the source-target cross-attention component of RNN-based architectures. This change makes all previous positions $t-1, t-2....1$ equally important in the selection of the current word (and likewise for source representations), it also enable to process all the target tokens in parallel during training, causing vast increase in training time. Decoding must continue to be carried out from left to right, since each prefix word already generated conditions the generation of future words.

In a nutshell, the Transformer architecture transforms a *structured context* (a sequence of tokens, but this can also be a tree or a graph) into a single numerical vector. The core operation in this transformation is the iterative computation of each individual token's representation based on their similarity to other tokens in the context. Denoting $H^{l-1} = [h_1^l, \ldots, h_T^l]$ the $(T \times d)$ matrix representing a context of length T at the input of layer l, the representation $\tilde{h}_i^{kl}$ for token i is computed by attention head kl as:

$$\tilde{h}_i^{kl} = \text{softmax}(\frac{h_i^{l-1} Q^{kl} [H^{l-1} K^{kl}]^T}{\sqrt{d_k}}) H^{l-1} V^{kl} \tag{3}$$

,

with Q^{kl}, K^{kl}, V^{kl} parameter matrices associated to these head and layer, d is the model dimension, and d_k is the size of each of the K heads ($d_k = d/K$). In this model, H_0 contains the lexical embeddings. The output of these k computations are then concatenated and passed through a feedforward (FFN) layer with ReLU activation; each of these steps also includes a summation with H^{l-1} and a layer normalization. In equations:

$$\begin{aligned}
\tilde{H}^{kl} &= \text{Attn}(H^{l-1}Q, H^{l-1}K^{kl}, H^{l-1}V^{kl}) \\
H^l &= \text{LayerNorm}(H^{l-1} + [\tilde{H}^{1l}, \ldots, \tilde{H}^{Kl}]) \\
H^l &= \text{LayerNorm}(H^l + \text{FFN}(H^l)),
\end{aligned}$$

where Attn denotes the computation expressed by equation (3).

When used for MT, self-attention mechanism is computed on the source side; on the target side an additional *cross-attention* module combines at each layer the decoder-side representations with the encoder representations output by the top-layer H^K.

2.1.3 Larger Transformers

Since their introduction, transformers have been studied and enhanced in many ways. For our concerns, an important line of research has focused on the extension of the context window. While the initial model only considered a sentencial context, the need

to enlarge the context window has quickly emerged, for instance to make the output of a language generating transformer more consistent, or, in MT, to model supra sentential phenomena, or more generally discourse related phenomena.

Computationally, the attention computation (equation (3)) has complexity $O(T^2)$, is performed k times for each of the l layers; for gradient computation, it is also needed to store the entire attention matrix of size (T^2). These are the main obstacles towards enlarging the transformer context, that we briefly review below. A complete survey of the current landscape of large transformer models is presented in [Tay et al., 2020].

Improving the time complexity Improving the time complexity requires speeding up the dot product operations involved in the attention. There are of course generic methods to speed up computation (eg. use half precision float representations), that we do not discuss any further here.

One specific way to proceed is to reduce the number of neighbours to a fixed size. In Liu et al. [2018b], this is achieved by restricting the attention computation to blocks of limited size. This means that the representation of a token only recombines the representation of tokens *within the same block*, thereby localizing these representations. It also creates boundary effects at the frontier between neighbor blocks. An alternative, *memory compressed attention*, explored in the same work, uses convolutions to compress the representations of neighbor blocks and to reduce the number of neighbors, while preserving access to a global context.

Boundary effects can also be avoided by considering neighbors in a (sliding) window of S words, which means that only the near-diagonal terms of the attention matrix will be computed. If context is localized in the lower layers, it still remains global at the upper layers since the influence of more distant words diffuses within the network. A further trick is to "dilate" these local contexts to speed up the diffusion in the network. To preserve the overall performance, a critical aspect is to make sure that a restricted number of positions still keep a global view over the full input, meaning that they attend to (and are attended to) by all positions. These positions can be described as performing local summaries that are propagated through the entire network. Having one such position every $\sqrt{T}$ block of length $\sqrt{T}$ ensures an overall $O(T\sqrt{T})$ complexity for the attention computation.

Such methods are notably used in the *Sparse Transformer* of Child et al. [2019], in the *Longformer* of Beltagy et al. [2020], and also employed in the GPT-3 architecture of Brown et al. [2020]. Introduced in [Ye et al., 2019], the *Binary-Tree Transformer* describes an alternative way to combine local and global attention patterns by explicitly organizing the propagation of attention in a (binary) tree structure. Each token's representation recombines local neighbors, as well as distant tokens whose representations are first condensed in *span-based* representations organized in a tree. This means that each token only needs to compute similarity scores with $O(\log(T))$ other nodes. In this approach, the root span representation remains the only place that integrates the full context in its representation.

The paper by Zaheer et al. [2020] finally introduces the *Long Bird*, which somehow

generalizes these ideas by introducing neighbor graphs specifying, for each position i, the sets of other positions that are used in the computation of $\tilde{h}_i^{kl}$. In addition to the local context, and to global tokens, the authors propose including a random component by adding random neighbors. By choosing specific random graphs, these authors show that random neighbors help speed up the "diffusion" of information across positions in the context, and that they do not compromise on the theoretical or the empirical performance of these sparser networks.

Another way to speed up this computation is to use approximations: in the *Reformer* of Kitaev et al. [2020], amongst other tricks, *locally-sensitive hashing* is used to identify the most significant terms in the attention (corresponding to the most similar neighbours), thereby also yielding sparse attention matrices. The *Linformer* approach of Wang et al. [2020] rests on the observation that the computation performed by attention heads can be approximated by the product of two low-ranks matrices. Furthermore, these low rank matrices can be obtained by introducing two random matrices for each head, one to project the $H^{l-1}V^{kl}$ term ($T \times d_k$) into a ($S \times d_k$) matrix (through the multiplication by a $S \times T$ matrix), the other to project $H^{l-1}K^{kl}$ also a ($S \times d_k$) matrix. As a result, the term within the softmax will output a $T \times S$ matrix (instead of $T \times t$). By choosing $T \gg S$, we get a complexity reduction from quadratic to linear, at almost zero cost in terms of performance. As in other papers cited in these section, the authors show that parameter sharing (here sharing the projection matrices across layers) can also help speed up the computation without harming performance (too much).

Saving memory The other computational bottleneck of large transformers is related to memory usage. All the attention values computed during the forward pass need to be stored as they are needed to compute the gradients during the backward pass. A common trick which is used in many implementations, is to resort to *gradient checkpointing*, a strategy which only stores a restricted number of attention matrices, from which the other can be recomputed when needed. This method enables the processing of larger batches or larger contexts, at the cost of an increased computation time. The proposal of Kitaev et al. [2020] is, in this respect, more effective: it replaces the standard computation performed in each layer in such a way that it becomes revertible. This means that only the upper layer of attention values needs to be stored after the forward pass, as it is sufficient to recursively retrieve all the lower level attention values.

Attempts to enlarge the transformer context at a reduced computational cost have flourished in the past year, enabling the development of models capable of textual handling contexts made of thousands or tens of thousands tokens. While we have probably not entirely covered this lively field (see [Tay et al., 2020] for a more thorough account), these extensions of the Transformer models matter for document-level MT, since they open the way for architectures with a document-wide attention. Experiments with large contexts in NMT are reported in [Ye et al., 2019] and further discussed in Section 3.1.1.

2.2 Decoding in NMT

2.2.1 Greedy search

Decoding, the process of generating the target sentence generally proceeds from left to right, reproducing the natural order of writing (in Indo-European languages). The *greedy* approach produces at each time step t the most likely next word $\mathbf{e}_t$ (or more precisely next token[7]) in the target vocabulary V_e, conditioned on the current target prefix $\mathbf{e}_{<t}$ and on the complete source sentence, according to:

$$\mathbf{e}_t^* = \underset{\mathbf{e}_t \in V_e}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\mathbf{e}_t|\mathbf{e}_{<t}, \mathbf{f}).$$

In the greedy version, decoding stops as soon as the system generates the end of sentence symbol </s>. This method rests too confidently on premature decisions and can lead to search errors, which are situations where the decoder fails to find $\arg\max_{\mathbf{e}\in V_e^\star} p_\theta(\mathbf{e}|\mathbf{f})$.

2.2.2 Beam search

Beam search (*beam search*) is a heuristic search method which extends the greedy method by keeping a set of active prefixes $B_t = \{\mathbf{e}_{<t,k}, k = 1 \dots B\}$. At each time step, the possible successors of these B prefixes are evaluated and a new B_{t+1} set is derived. There are two main families of approaches to develop B_{t+1}:

- to keep any prefix whose cumulative probability is not too far from the best prefix $\mathbf{e}_{<t,1}$, keeping any hypothesis $\mathbf{e}_{<t,1}$ such that $p_\theta(\mathbf{e}_{<t,k}|\mathbf{f}) > (1-\alpha)p_\theta(\mathbf{e}_{<t,1}|\mathbf{f})$, with $\alpha \in [0, 1]$ a parameter that controls the width of the beam. This leads to a variable number of active prefixes;
- to keep the best B prefixes, which has the merit of keeping the number of active prefixes constant. This variant is also known as *histogram pruning* .

In the beam search procedure, the decoder stops as soon as:

- an active prefix corresponds to a complete sentence (ending in </s>) which cannot be developed any further;
- all other active prefixes score lower than the full sentence and therefore will not be able to surpass it in the future.

The complexity of this algorithm is $O(IBV_e)$, since at each time step the possible V_e continuations of the B prefixes are calculated.

[7] To be able to deal with open vocabularies, MT systems rely on a division of words into sub-lexical units calculated on purely statistical bases [Sennrich et al., 2016]. These units have the immense advantage of allowing the homogeneous processing of various languages, but lead to the disappearance of the notion of words which is nevertheless central to access traditional linguistic resources and models.

2.2.3 Search errors

Early implementations of beam search have faced an apparent paradox, according to which increasing the size of B, thus of the search space, led to degraded results. Various explanations have been put forward [Murray and Chiang, 2018, Stahlberg and Byrne, 2019], the most convincing of which is based on the following observations: (a) in the absence of information on the length of the source, the criterion for stopping decoding is the generation of an end-of-sentence symbol (</s>); (b) the probability of sequences is a product, and short sequences are generally more likely than long sequences. Increasing the search space leads to the inclusion in the beam of sequences that are too short, but which will nevertheless prove to be the most likely for the decoder. To illustrate this with an extreme case, denote k the rank of the hypothesis which generates </s> at the first time step and thus results in an empty translation; when $B \geq k$, this hypothesis will enters the beam at $t = 1$ and its score will never change, while all other competing hypotheses will see their score decrease when more words are generated, often leading to this hypothesis to be the preferred one in the end. Various remedies (length normalization, trainable word penalty, etc) are put forward in the cited publications, to which we refer the reader.

2.3 Summary

Encoder-decoder NMT architectures have quickly converged towards the Transformer model, which lies at the core of most, if not all, modern NMT systems. Transformers have the ability to compute useful representations from complex, possibly heterogeneous, possibly very long, input sequences. Such architectures made it possible to combine local and non local contexts in a principled manner, as we discuss in the next section. Also note that their optimization can be made very effective through parallelization, and that they support multiple auto-regressive and non-autoregressive decoding algorithms. This latter property is a facilitating factor for the injection of complex constraints during decoding.

3 Document Level Neural Machine Translation

Context-aware neural machine translation aims to relax the assumption that sentences should be translated independently from their neighbour. For each sentence, the translation is thus conditioned on the current source as well as on other source and/or target sentence(s) from the same document; this additional information will be referred to as the *context*. Formally, given a document D containing K sentence pairs $\{(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}), (\mathbf{f}^{(2)}, \mathbf{e}^{(2)}), \ldots, (\mathbf{f}^{(K)}, \mathbf{e}^{(K)})\}$, the probability of translating $\mathbf{f}^{(k)}$ into $\mathbf{e}^{(k)}$ is given by:

$$p(\mathbf{e}^{(k)}|\mathbf{f}^{(k)}) = \prod_{t=1}^{l_k} p(\mathbf{e}_t^{(k)}|\mathbf{e}_{<t}^{(k)}, F, E^{(<k)}) \tag{4}$$

where $F := [\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)}]$ represents the source sentences and $E^{(<k)} := [\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(k-1)}]$ the previously generated target sentences.

Most current document-level NMT models can be broadly classified into two main categories [Chen et al., 2020]: *context-aware models*, discussed in section 3.1, and *post-processing models*, that are studied in section 3.2. The context-aware models use the contextual information during the translation process, whereas the post-processing approaches introduce an additional module that learns to refine the translations produced by a context-agnostic systems to be more discourse coherent [Voita et al., 2019b, Xiong et al., 2019]. To be complete, it is also worth mentioning the approach of Saunders et al. [2020], who propose to integrate a document level metrics (TER) in the training objective function, using the framework of Minimum Risk Training.

Early approaches to document-level NMT (2016-2019) explored several models and architectures to take the context into account, starting with the simple concatenation of a few preceding sentences [Tiedemann and Scherrer, 2017, Agrawal et al., 2018, Junczys-Dowmunt, 2019], or cache-based methods to store their representation [Tu et al., 2018, Kuang et al., 2018, Maruf and Haffari, 2018]. More recently, Miculicich et al. [2018], Yang et al. [2019], Maruf et al. [2019a] use a Transformer architecture to integrate contextual information with context-related modules.

Despite the additional computing cost, the observed improvements remained small when measured with standard metrics such as BLEU and were only significant when the evaluation was performed using artificial test sets targeting specific phenomena such as the translation of pronouns [Tiedemann and Scherrer, 2017, Bawden et al., 2018a, Voita et al., 2019b]. For the cases where metric scores were found to increase, the improvements were limited to specific datasets/testsets [Tu et al., 2018, Maruf and Haffari, 2018, Kuang et al., 2018, Zheng et al., 2020]. Year 2019 onwards, more studies started focusing on comparing the existing architectures and on analyzing and diagnosing the learnt representations, the effect of context length, etc. [Kim et al., 2019, Li et al., 2020, Chen et al., 2020].

3.1 Context-Aware Models

Traditional machine translation models rely on strong independence and locality assumptions: source word and phrases in SMT are (conditionally) independent; sentences are processed independently both in SMT and NMT. As discussed in Section 1, this assumption ignores many important types of dependencies in translations, notably those that are related to the discourse structure, and which typically span over several sentence(s).

Attempts to model such extended contexts in NMT can be broadly classified depending how they incorporate context: *single encoder approaches*, presented in Section 3.1.1, simply concatenate the context with the current sentence. More sophisticated approaches recourse to additional neural networks modules to encode the context: this is the case of multi-encoder approaches (Section 3.1.2), and of approaches that rely on an dedicated memory structure (Section 3.1.3). A high-level overview of these alternatives is in Table 7, where we sort systems based of the size and representation of the context.

Approach		Context			Lang. Pair	Reference
context encoding	integration in NMT	prev	next	size		
concatenated inputs		s		1	DE→EN	[Tiedemann and Scherrer, 2017]
		s, t	s	3	EN→IT	[Agrawal et al., 2018]
		s, t		1	EN→FR	[Bawden et al., 2018b]
		s, t	s	variable	EN↔DE	[Scherrer et al., 2019]
augmented input		s	s	all	DE→EN/FR	[Rios Gonzales et al., 2017b]
		s	s	all	EN↔FR, EN→DE	[Macé and Servan, 2019]
cache	decoder	s, t		variable	ZH→EN	[Tu et al., 2018], [Kuang et al., 2018]
encoder	decoder	s		3	ZH→EN	[Wang et al., 2017a]
attention	encoder, decoder	s, t		3	ZH/ES→En	[Miculicich et al., 2018]
		s, t	s, t	all	EN→DE	[Maruf et al., 2019a]
encoder w/attention	source context	s, t		1	EN→FR	[Bawden et al., 2018b]
	encoder	s		1	EN→RU	[Voita et al., 2018]
	decoder	s		1	EN→FR/DE	[Jean et al., 2017]
		s, t		all	FR/DE/ET/RU↔EN	[Maruf et al., 2018]
		s		1	ZH↔EN	[Kuang and Xiong, 2018]
		s, t		1	DE/ZH/JA↔EN	[Yamagishi and Komachi, 2019]
	decoder, output	s, t	s, t	all	FR/DE/ET→EN	[Maruf and Haffari, 2018]
	encoder, decoder	s		2	ZH/FR→EN	[Zhang et al., 2018]
		s		2	FR→EN	[Wang et al., 2019]
second-pass decoder		t	t	all	ZH→EN	[Xiong et al., 2019]
		s, t		3	EN→RU	[Voita et al., 2019a]
document context language model				2	ZH→EN	[Yu et al., 2020]
context-dependent post-editing		t*	t*	4	EN→RU	[Voita et al., 2019b]
learning w/context regularisation		s		1	EN→RU	[Jean and Cho, 2019]
learning w/oracles		s		1	EN→DE	[Stojanovski and Fraser, 2019]

Table 7: Overview of works which successfully incorporate extra-sentential context information in NMT. Letters s and t denote source or target-side context, and "amount" defines the number of sentences used as context. An additional * flags studies that do not use the notion of past and future, but rather use sentence groups for training and evaluation (inspired from [Maruf et al., 2019b]).

3.1.1 Single Encoder Approaches

The simplest way to feed more context in neural MT is to modify the input, i.e. to concatenate one or several surrounding sentence(s) to the current one and process the extended sentence as usual, as done in [Tiedemann and Scherrer, 2017, Jean et al., 2017, Agrawal et al., 2018, Junczys-Dowmunt, 2019]. A special token is inserted between the context and the targeted sentence to identify boundaries (e.g. BREAK). Figure 3 illustrates this approach. Here, a single encoder processes the context and the current sentence together as one long input. This approach requires no change in the model architecture; however it badly increases the computational cost of encoding, which grows quadratically with the source length in the original Transformer architecture. As discussed in Section 2, using recent variants of the Transformer can reduce this complexity. Notwithstanding these computational costs, data scarcity of a higher-dimensional input space also makes

it difficult to train the attention component with very long spans [Sukhbaatar et al., 2019].

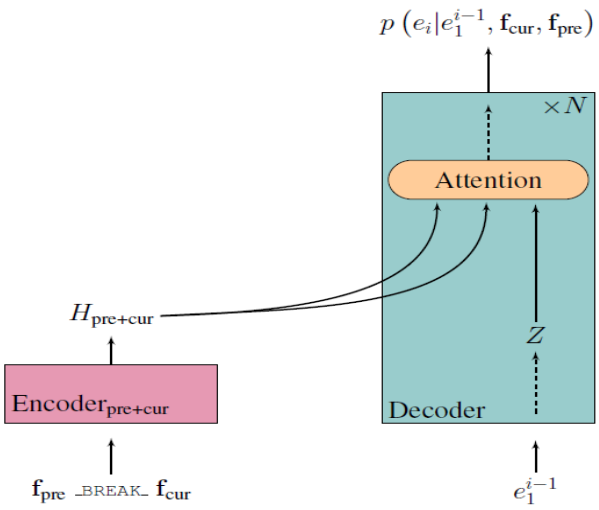

Figure 3: Single encoder approaches, figure borrowed from [Kim et al., 2019].

In an early attempt, Tiedemann and Scherrer [2017] explore two ways to use an extended context: one that adds source language history to the input, and the other that uses both the past source and target sentences. They explore the influence of a limited number of context sentences, up to two in source or target. Working on the OpenSubtitles2016 [Lison and Tiedemann, 2016] corpus for German to English translation, they reported marginal improvements in terms of BLEU, chrF3, precision and recall. Through manual evaluation, they were able to find output examples in which referential expressions across sentence boundaries could be handled properly.

Agrawal et al. [2018] extend this idea and run a comparison between RNN and Transformer-based systems. They experiment with various context sizes: up to three previous and one next sentence on the source side, up to two previous sentences on the target side. They report Transformer-based models to outperform the RNNs, attributing this to the inherent inability of RNNs to accommodate long-range dependencies. 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 seems detrimental, due to error propagation. Junczys-Dowmunt [2019] takes the idea further and uses the whole document as context for their document-level systems.

Ma et al. [2020] also present a single encoder approach where they modify the Transformer to incorporate context and source sentence in a single encoder, called the "Flat-Transformer". To this end, the encoder is separated into two parts for both the global and the local attention. Bottom encoder blocks apply self-attention to the whole sequence, while for the top blocks, it is only applied at the current source sentence. They compare their architecture with vanilla single-encoder as well as dual-encoder document-level sys-

tems and report that the encoder with a unified structure yields a gain of 1.08 in terms of BLEU and 2.03 in terms of Meteor.

The study by [Fernandes et al., 2021] confirms the overall merits of this approach; they however show that the usefulness of past sentences quickly vanishes after one or two sentences; they also suggest, following [Bawden et al., 2018b], that the target side context may be more useful than the source side context; they finally propose to use word drop out to improve the effect of the contextual model.

3.1.2 Multi-Encoder Approaches

In contrast to most single-encoder approaches, multi-encoder approaches handle context integration at the architectural level (Figure 4), meaning that the context and the current sentence are processed by distinct mechanisms. An early, pre-Transformer, approach is in [Wang et al., 2017a], who use a hierarchy of RNNs to summarize a context containing all the past sentences in the document, and explores various ways to combine it with the current sentence. A further distinction with multi-encoder architectures is whether the integration takes place inside the encoder or inside the decoder. Note that this distinction does not depend on specific types of context encodings, for which one can use recurrent or self-attentive encoders with a variable number of layers, or just word embeddings without additional hidden layers on top of them.

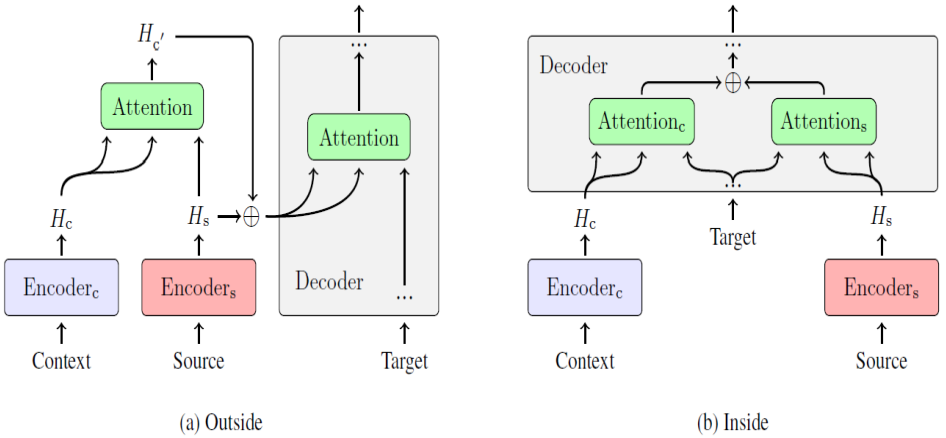

Figure 4: Multi encoder approaches, inside and outside integration of context (borrowed from [Li et al., 2020])

Integration outside the decoder In this approach, the current sentence and its context are first transformed (encoded) using a source side network, for instance an attentional model. These representations are then fused by a gated sum before being fed to the decoder [Voita et al., 2018, Miculicich et al., 2018, Zhang et al., 2018, Maruf and Haffari, 2018]. This gating mechanism allows the model to learn which additional contextual

information should be included. We will give a brief overview of each of these studies in this section.

Voita et al. [2018] propose to encode the source sentence and its context independently, a single attention layer is then used in combination with a gating function to produce a context-aware representation of the source sentence. As using separate encoders does not seem to yield an accurate model, they propose to share the parameters of the first $N-1$ layers of the context encoder with the source encoder; a special token is further used at the start of context sentences to make the shared layers know whether it is encoding a source or a context sentence. Working on English-Russian translation of subtitles, they study the effect of previous and next sentences and *found only previous sentences to be beneficial.* They also found contextual attention to be is high for the translation of function words such as "*it*", "*yours*", "*ones*", "*you*" and "*I*" and, similar to [Tiedemann and Scherrer, 2017], report words like "*yes*", "*yeah*", and "*well*" in the list of top-10 context-dependent words. They conjecture that the context is especially helpful at the beginning of a sentence, and for shorter sentences, since they found a negative correlation between the amount of attention placed on contextual history and sentence length, and between token position and the amount of attention to context. They report more improvements for sentences containing ambiguous pronouns.

The study of Miculicich et al. [2018] is the first to use hierarchical attention networks (HANs) [Wang et al., 2016] to model contextual information at the word and sentence level. They use an extended version of HANs with multi-head attention to model the context. Two separate HANs are considered for integrating respectively the source and target contexts. Training follows a two-step procedure where they first build a vanilla Transformer-based system, then optimize the parameters of the whole network including the HAN module. A context of 3 preceding sentences was used. In their evaluation, an assessment of the new model was performed separately for three types of discourse phenomena. First, an evaluation of noun and pronoun translation was performed using accuracy with respect to a human reference. For the lexical cohesion, the ratio of repeated and lexically similar words over the number of content words were measured, as suggested in [Wong and Kit, 2012]. Finally, for coherence, they computed an average cosine similarity between consecutive sentences using the metrics proposed by Miculicich Werlen and Popescu-Belis [2017a] (APT). In all cases, an improvement with respect to the baseline system was observed. In particular, their model was able to identify relevant previous sentences and words for the context prediction. Their results also suggested that the contextual information obtained from source and target sides are complementary. However, one important limitation of these conclusions is that they are based on a restrictive notion of the context, including a small number of previous source/target sentences. Chen et al. [2020] use HAN with discourse representation structures (DRS); however the improvements achieved are more or less are on the same scale as regular HANs and are thus not reported in Table 7.

Zhang et al. [2018] present an approach akin to pre-training and extend the Transformer model with a new context encoder to represent document-level context, which is then incorporated into the original encoder and decoder. They use a multi-head self-attention Transformer model to first train a sentence-level system, which is then extended with

document-level model parameters estimated on document-level parallel corpora. The default Transformer residual connections are used in each sub-layer, but to control the influence of context, the residual connections after the context attention sublayer are replaced with a position-wise context gating sublayer. Using two-step training, they report improvements with respect to the previous work, reporting contrasts with a RNN based contextual model Wang et al. [2017a], and with the cache based models of Kuang et al. [2018], presented in Section 3.1.3.

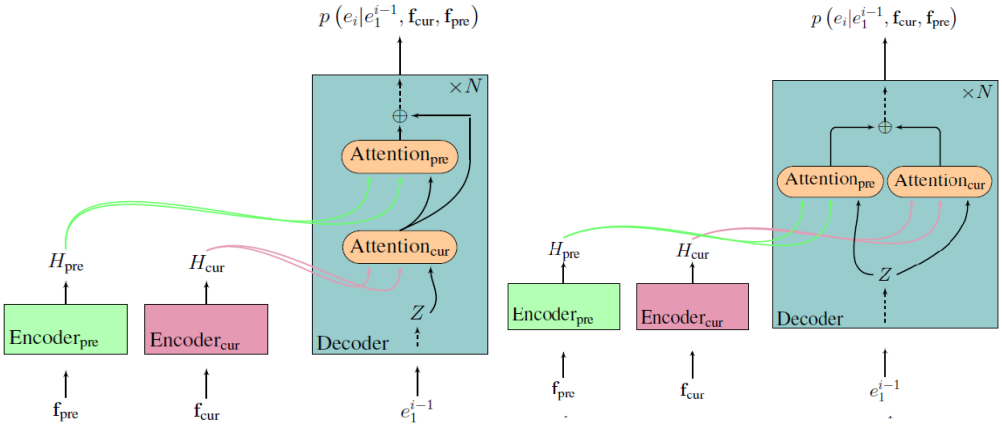

Figure 5: Multi encoder approaches: Inside integration of context, sequential attentions (left) and parallel attentions(right) (graph borrowed from [Li et al., 2020])

Integration inside the decoder These methods incorporate context inside the decoder, meaning that the target word generation process can separately attend to the source and the context representations, in addition to the available target-side prefix. Depending on the specific architecture, this combination of source and context attention can be performed *sequentially*, as in [Tu et al., 2018, Zhang et al., 2018] or *in parallel* [Jean et al., 2017, Bawden et al., 2018a, Stojanovski and Fraser, 2018]. Compared to single-encoder approaches, this strategy also enables to use simpler processing modules for the context, which arguably is less informative than the current sentence for the translation.

In sequential attention models, the two attention components are stacked such that the output of one component is the query for the other [Tu et al., 2018, Zhang et al., 2018]: this is illustrated on Figure 5 (left), where the current sentence is first attended to by the decoder, providing the input for the next cross-attention layer evaluating the context sentence(s). This augments the baseline cross-attention with supplementary contextual information. Note that the order of the attention components may also be switched. To block signals of potentially irrelevant context information, a gating mechanism can be employed between the regular sentencial and context attention outputs.

In *parallel attention models* (see Figure 5 (right)), the two attention operations are performed in parallel and only combined with gating afterwards [Jean et al., 2017, Bawden et al., 2018a, Stojanovski and Fraser, 2018, Kuang and Xiong, 2018]. By doing so,

the document context is only queried based on the current target history, independent of the current source sentence representation, which has the additional benefit to speed up the decoding.

3.1.3 Cache-based methods

Cache based models are often used to provide short-term memory, for example as an additional language model in the manner of the cache models of Kuhn and DeMori [1990], or Bertoldi et al. [2013] in a computer-assisted translation context. Such models use a specific short term memory mechanism typically aimed to boost the probability of words that have been generated in the recent past. This means that these models try to model and integrate target side context. Earlier attempts with SMT to use cache based language and translation models include [Nepveu et al., 2004, Tiedemann, 2010, Gong et al., 2011, Bertoldi et al., 2013] and the works on NMT include [Kuang et al., 2018, Tu et al., 2018, Maruf and Haffari, 2018, Yang et al., 2019, Dobreva et al., 2020].

The proposal of Tu et al. [2018] uses a *continuous representation* of the cache to store recent context history. After each sentence translation, the decoding contexts are stored in the cache as history for use in future decoding steps. Cache slots are pairs of key-value vectors, with the keys being attention context vectors, and values corresponding to the decoder states collected from previous translations (see Figure 6). Using context representations as keys in the cache, the cache lookup is performed using dot products between the stored representations and the current sentence. Cache is added to a pre-trained NMT model with fine-tuning, updating only the new parameters related to the cache.

The reported improvements are around 1 BLEU points for Subtitles and News, and about 1.5 BLEU points for TEDTalks. The authors also found smaller cache sizes (of 25 sentences) to be simlar in performace to larger caches (of size 500). They report keeping complete sentences in the cache to be more advantageous for lexical consistency as compared to keeping a few previous words [Gu et al., 2016]. A few example sentences are also presented, for which their model improves *verb tense consistency*, even though this is not reflected in the BLEU score.

Kuang et al. [2018] use both a *topic* and a *dynamic cache*. They add a new neural network layer as the scorer for the cache. During decoding final word prediction probability is computed via gating mechanism by combining the probability estimated from the cache to the original probability computed by the decoder. The static topic cache is built using topic distribution from each source document which is used to obtain the corresponding topical words on the target side. These topical words are then integrated into the topic cache. For the dynamic cache, words are retrieved from the best translation hypotheses of recently translated sentences. These caches are reset each time the decoder shifts to a new test document. Each sentence of the new test document is then translated with the cache-augmented model; the dynamic cache is augmented with the newly generated target words obtained from the best translation hypothesis for previous sentences. Significant improvements in translation quality are reported by the authors. Topic cache and dynamic cache were also found to be complementary to each other.

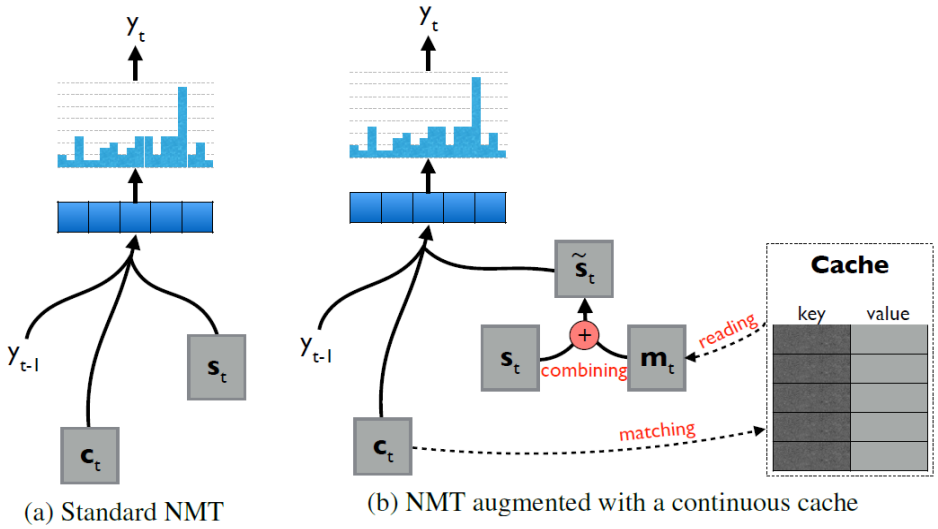

Figure 6: Architectures using Caches (figure taken from [Tu et al., 2018]).

Maruf and Haffari [2018] use memory networks to include source and target context information in their NMT. Each sentence in the document is passed through a word-level bi-directional RNN to get a complete sentencial representation. This information is then propagated across the document by passing these sentence representations through a sentence-level bi-directional RNN. The same encoding process is applied to the source and target document contexts, which are then used either to compute the next hidden state, or more directly as extra information provided to the output layer. These authors worked on French-English, German-English and Estonian-English, and report BLEU and Meteor scores, as well as results of a manual evaluation. Comparing the use of a sole source memory, of a sole target memory, and of memories, they notably report that using *both source and target memories gives the best results for all three language pairs*. This suggests that leveraging both source and target document context is indeed beneficial to MT performance.

Yang et al. [2019] retrieve helpful contextual features from (a small number of) previous sentences by applying a dynamic routing algorithm called *query guided capsule network*. They use these features by extending the source encoder of the Transformer architecture with a supplementary source-context cross-attention layer, among other architectural changes. Their experiments with three standard benchmarks (TEDTalks, News commentary, and Europarl, for the German-English language pair) show a small improvement over context-independant and context-dependant baselines.

Finally, the work of Dobreva et al. [2020] makes two contributions: one is to use a context tag to inform the encoder with information about the document structure; the other is to use a fixed-size topic cache and dynamic cache similar to the proposal of Kuang et al. [2018],. These are concatenated and passed to output layer, thereby helping to improve the estimation of the probability distribution over words. The topic model is

built using the most probable words learnt using the topic model, and the dynamic cache contains the set of unique content words from previously translated sentences from the same section. Out-of-domain Transformer models are further fine-tuned with in-domain data (in that case biographies) .

It is interesting to note the similarity between these techniques and approaches proposed in the context of resource-rich NMT, where dictionary entries, multi-word terms or other topic information have also been introduced in the form of cache LMs or related mechanisms [Yvon and Abdul Rauf, 2020].

3.2 Multi-pass systems

Multi-pass systems usually introduce an additional computational component that helps to refine translations produced by context-agnostic systems and make them more globally coherent [Xiong et al., 2019, Voita et al., 2019b, Yu et al., 2020]. Such approaches are easy to implement as they only rest on the availability of sufficiently large monolingual document corpora and do not require to change the first pass system; however, the two-stage generation process may result in cascading errors [Xiong et al., 2019].

The two-pass strategy of Voita et al. [2019b] resembles automatic post-editing (APE) [do Carmo et al., 2020]. They train a sentence level system whose output is then corrected by using a monolingual document-level model, called the *DocRepair model*. This model aims to corrects inconsistencies among the individual sentence translations in the context of the other sentences. The DocRepair model is trained using only monolingual document level data in the target language, with the inconsistent sentences produced by back translations as source and consistent sentences as target. They report improvements in terms of BLEU score, targeted contrastive evaluation for deixis, lexical cohesion, VP ellipsis and ellipsis which affects NP inflection, as well as human evaluation.

Yu et al. [2020] frame their work in the paradigm of the noisy channel model, where the best document translation $E^* = [\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(K)}]$ for the source text $F = [\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)}]$ is computed as $\text{argmax}_E \, P(E)P(F|E)$ instead of $\text{argmax} \, P(E|F)$, which allows them to take advantage of a *document-level target language model* for $P(E)$. Remarkably, this target side language model can be trained with document-level monolingual data, which are much easier to find than parallel document in multiple languages. The authors claim that this approach is advantageous for two reasons: (a) the conditional translation model and the document language model can be trained separately, the latter component being in a position to exploit large sets of monolingual data; using an additional assumption, training the translation can be performed with parallel sentences only, and does not require parallel documents which are a rather rare resource. If training is simple, decoding is hard because every $\mathbf{e}$ is of course unknown in training. As is custom for models of this type, optimal decoding is approximated with a two-step procedure: (a) generation of a lattice of possible document translations using a non-contextual forward model ($P(\mathbf{e}|\mathbf{f})$); (b) (beam) search re-scoring of this lattice using the full " inverted" model. They achieve substantial improvements as compared to existing document translation approaches and report improvements in number, lexical choice and tense.

The recent work of Kang et al. [2020] somehow departs from the other works in this

section, as they do not combine two search procedures, but instead revise the first-pass model dynamically. Their dynamic context approach is based on the claim that "different source sentences may require a different context size". They introduce an independent context selection module which selects context sentences prior to the MT step through an independent retrieval step; based on the output score of this module, a set of context sentences (possibly empty) will be input to the NMT encoder, along with the current sentence to be translated. As the selection module includes a non-differentiable step, the corresponding score function can not be trained end to end, and the authors have to resort to a reinforcement learning strategy, where the final translation quality serves as the reward signal. The authors report improvements in BLEU scores, as well as in the processing of discourse phenomena using the contrastive test sets of Voita et al. [2019a].

3.3 Discussion: is context really useful in NMT ?

In the previous sections, we have reviewed various attempts to increase the size of dependencies in neural MT, and to mitigate the effect of systematic errors incurred when translating sentences in isolation. In most of these studies, the authors have concluded positively on the usefulness of context in MT. There is however much variance on the notion of context: is it short or long distance ? just the past and also the future ? just the source or also the context ? In this section, we relay some of the contradictory findings that have been reported on this issues. Is the context useful after all ?

A first answer to those questions is in [Voita et al., 2018], which reports performance degradation when using the following sentence, a conclusion that is contradicted by the results obtained by Wong et al. [2020] on the same corpus. This study also reports that taking the future context into account not only yields significant improvements over a context-agnostic Transformer, but also yields comparable and in some cases better performance than training with the past context.

Going one step further, [Kim et al., 2019] and [Li et al., 2020] make a case *against using a document-level context*. While investigating the effect of context and parameters on the quality of document-level translation, both studies try to show that a strong sentence-level baseline (with enough data or proper regularization, drop out etc.) does not show much improvements with added context. Their main claim being that context encoders are more useful as noise generators, enabling to regularize the training, than as providers of an additional supervision signals to train the sentence-level system.

The study of Kim et al. [2019] compares the usefulness of various architectures and representations of a short term context (one to a few past sentences). One contribution is to show that these representations can be much simpler and lightweight than for the current sentence. For instance, filtering out words from the context sentences based on predefined lists or linguistic tags does not seem to harm the translation quality (measured with BLEU) as compared to using a full context. Using simpler representations also enables to increase the size of the context: in this study, it was found that only a few previous sentences are actually useful for translation quality. Another result is that the observed improvements in translation obtained with longer contexts are hard to analyze in terms of better processing of discourse phenomena (coreference, lexical choice,

etc). Finally observing that their document-level models did not outperform a sentence-level baseline trained on a large corpus, they attributed the improvements achieved by existing document-level models to a better parameter regularization. These conclusions should be somewhat nuanced given that (a) most experiments are performed in small data conditions; (b) their test domain (TED talks [Cettolo et al., 2012]) may be more appropriate to study document-wide effects (consistency, rhetorical structure) than local phenomena such as co-reference; (c) most of their analysis is based on BLEU scores, which is a poor metric for document-level MT.

Similar findings are reported by Li et al. [2020], who focus on multi-encoder approaches. They experimented with several such architectures (with or without weight sharing and pre-training, and varying depth levels) and three types of contexts; *Context*, consisting on the previous sentence, *Random*, using a set of random words and *Fixed*, using a fixed context for all sentences. They found that both *Random* and *Fixed* systems can achieve comparable or even higher BLEU scores than *Context* in most cases. These results also hint to interpreting the role of the context as a regularizer, a fact they try to confirm using a (Gaussian) random noise generator combined with a regular encoder.

According to Maruf et al. [2019b], the results reported in these two studies certainly raise intriguing questions, but are not sufficient to fundamentally question the success achieved by most studies of document-level NMT. They however suggest that document-level MT models still require further experiments and analysis before more firm conclusions can be drawn regarding the pitfall and strengths of these context-aware models.

3.4 Conclusion

In this section, we have surveyed recent approaches to document level machine translation, studying first methods focusing on integrating a larger context with the sentencial representation in a unified architecture (section 3.1), before surveying multi-pass approaches - where a baseline, sentence based translation, is further processed with a stronger target language model.

While the benefits of an extended context are somewhat questionned, especially with respect to the usual evaluation metrics such as BLEU or Meteor, the recent systematic study of Ma et al. [2021] still shows that simple approaches [Scherrer et al., 2019, Junczys-Dowmunt, 2019] complemented with additional (back-translated) target monolingual document-level data can be quite effective, both in terms of BLEU, but also in terms of the other standard metrics for document level MT.

4 Evaluation

Interpreting translation errors is not always simple and often involves a speculation on the actual cause of the error. Mistranslating a pronoun can be a morpho-syntactic or a discourse error: if the error only concerns the gender, then this may be attributed to erring on the identification of a co-referent; however a wrong case can assimilated to a grammatical error. As a general principle, when evaluating discourse phenomena, the

main focus should be on errors that are caused by ignoring the context [Guillou et al., 2016].

Traditional MT metrics such as BLEU [Papineni et al., 2002], TER [Snover et al., 2006] or Meteor [Banerjee and Lavie, 2005] are unreliable for evaluating discourse phenomena in MT. The corresponding score computations rely on sentence-level comparisons, thus ignoring essential relationships between clauses and sentences [Joty et al., 2017]. This point was also neatly made in [Vojtěchová et al., 2019], which describes the use of test set comprising audit reports from the Supreme Audit Office of Czech Republic for English, German and Czech languages as a complementary exercise during the WMT19 News Translation Task.[8] The main goal of this manual evaluation was to assess the performance of NMT trained on general purpose texts for such very specific documents, and notably to evaluate the use of terminology and as well as the document level coherence. These experiments showed that even though NMT systems perform well for automatic metrics, a lot of subtle errors remain (mistranslation, misuse of terminology, inconsistencies across sentences) that could not be evaluated based on comparison on single reference translation.

As translation quality improves, there is however a growing need for context-aware *automatic evaluation* measures to capture the discourse information that remains out of reach for sentence-level metrics. Attempts along these lines can be categorized into two main groups. A first set of metrics mimic, for specific errors categories and test sets, the reference-based approach that is used by global metrics., there have been evaluation studies focusing on errors that are caused by insufficient modeling of the sentence context and discourse structure. In particular, Popescu-Belis [2019] distinguishes between discourse errors based on lexical choice, anaphora and coreference from those that can be attributed to the general discourse structure. Reference-based assessment can however be challenging: this is because in addition to agreeing with a reference, good translations at the doucment level must also show *internal consistency*. For instance, the validity of a pronoun translation may actually depend more on target side agreement constraints that depend on previous choices than on its similarity to the reference translation(s), which is what the traditional MT metrics measure [Loáiciga et al., 2017].

Therefore, a large body of works to measure discourse phenomenon use evaluation methods based on *contrastive pairs* (section 4.1.2) which compare the likelihood of two possible translations, one correct and one deliberately made wrong. For this type of evaluation, a system gets a positive reward if it gives a higher likelihood to the correct than to the incorrect translation; else, it will get a negative reward. This technique is used for instance by Sennrich [2017] for the evaluation of grammatical errors and in Rios et al. [2018] for word sense disambiguation.

This section is organized according to the type of phenomena and reproduces the distinctions identified in the opening section (1). We thus distinguish the evaluation of coreference (Section 4.1), of deixis and ellipsis (Section 4.2), of lexical cohesion (Section 4.3), then of word-sense disambiguation (Section 4.4) and finally of discourse structures (Section 4.5). Example sentences are presented to clearly demonstrate the point

[8] http://statmt.org/wmt19/

Evaluation Type	Discourse Phenomena	Dependency	Metric	Reference
Automatic Metrics	Pronoun	Alignments, pronoun lists	AutoPRF	[Hardmeier and Federico, 2010a]
		Alignments, pronoun lists	APT	[Miculicich Werlen and Popescu-Belis, 2017b]
		Pairwise evaluation	CRC	[Jwalapuram et al., 2019]
		Alignments, pairwise evaluation		[Wong et al., 2020]
	Lexical	Lexical cohesion devices		[Wong and Kit, 2012]
	Cohesion	Topic model, lexical chain		[Gong et al., 2015]
	Structure	Alignments, dictionary	ACT	[Meyer et al., 2012, Hajlaoui and Popescu-Belis, 2013, Meyer et al., 2015]
		Re-sampling, app. randomization	MultEval	[Meyer and Poláková, 2013]
		Discourse parser		[Guzmán et al., 2014],[Joty et al., 2017]
		Discourse parser	Dis-Score	[Smith and Specia, 2018]
Test Suites	Pronouns	EN→FR	Protest	[Guillou, 2013, Guillou and Hardmeier, 2016]
		EN→FR		[Bawden et al., 2018a],[Lopes et al., 2020]
		EN→DE		[Müller et al., 2018]
		EN→DE	ControPro	[Huo et al., 2020], [Lopes et al., 2020]
		JA →EN		[Nagata and Morishita, 2020]
	Cohesion	EN→FR		[Bawden et al., 2018a]
		EN→RU		[Voita et al., 2019a]
		EN→FR, EN→DE		[Lopes et al., 2020]
		JA →EN		[Nagata and Morishita, 2020]
	Coherence	EN→FR		[Bawden et al., 2018a]
		EN,CS↔DE,EN↔CS	SAO	[Vojtěchová et al., 2019]
		EN→CS		[Rysová et al., 2019]
	Conjunction	EN/FR→De		[Popović, 2019]
	Deixis, Ellipsis	EN→RU		[Voita et al., 2019a]
	Grammatical Phenomena	EN→DE	LingEval97	[Sennrich, 2017]
		DE→EN	DFKI	[Avramidis et al., 2019]
	Word Sense Disambiguation	DE→EN/FR	ContraWSD	[Rios Gonzales et al., 2017b]
		DE→EN/FR		[Rios et al., 2018]
		EN→CS, EN↔DE/FI/LT/RU	MUCOW	[Raganato et al., 2019]
		CS↔EN		[Zouhar et al., 2020]

Table 8: Overview of works on evaluation of discourse phenomena (inspired by [Maruf et al., 2019b]).

(where available). A summary of the major works on evaluation listing the dependencies, techniques and languages are presented in Table 8.

4.1 Evaluating co-reference

4.1.1 Reference-based evaluations

A first attempt to evaluate the translation of pronouns is in [Hardmeier and Federico, 2010b], which proposed a metric based on precision and recall. Working on German-English, they identified anaphoric links using a co-reference resolution system and studied how often an anaphoric pronoun is correctly translated, i.e. matches the agreement features of its antecedent. Automatic word alignments are generated between source and reference, and source and output. A clipped count is computed for each pronoun in the source language against the reference and output translation. Precision, recall and F-score against these clipped counts are computed to evaluate credible pronoun translations.

Guillou [2016] presents a semi-automatic pronoun evaluation test suite (PROTEST for English to French) containing 250 pronoun tokens. It consists of manually selected pronoun tokens of various types: anaphoric, cataphoric, event, textual, pleonastic, speaker and addressee reference, and relies on a comparison between reference and output translations. Pronouns in the MT output and the reference are automatically compared, but a manual evaluation is required in when no match is found.

Several shared tasks have since been organised with a focus on pronoun translation. The evaluation protocols have evolved over the years due to the difficulty of this type of evaluation. The DiscoMT 2015 shared task [Hardmeier et al., 2015] started with a human evaluation but the subsequent shared tasks [Guillou et al., 2016, Loáiciga et al., 2017] switched to a semi-automatic evaluation which involved computing macro-averaged recall for pronoun classification. Tasks were designed as classification tasks, where the participants are given the source language pronoun in the context of a sentence together with a lemmatised versions of the reference translations where pronouns have been deleted. Note that in this design, using the partly lemmatized test sets provided by the shared tasks organizers requires to train models on the provided lemmatized data and results in systems that are not usable in real translation settings.

Miculicich Werlen and Popescu-Belis [2017b] propose a reference based evaluation metric (APT). Using word-level alignments, first triples of pronouns are identified in the form of (source-pronoun, reference-pronoun, candidate-pronoun). These are then compared against the corresponding reference. Counts are collected for the identical, equivalent, different/incompatible translations in the output and reference, as well as cases where candidate translation or reference translation or both are absent. Correctness of MT output given the reference is determined by assigning a weight between 0 and 1. These weights and the counts are then used to compute the final metric score. They report APT metric to reach around 0.993–0.999 Pearson correlation with human judges, while other automatic metrics such as BLEU, Meteor, or those specific to pronouns used at DiscoMT 2015 reached only 0.972–0.986 Pearson correlation.

APT is also used by Wong et al. [2020] where an evaluation of the translation of cataphoric pronouns involves multiple metrics, including also CRC (see below) [Jwalapuram et al., 2019], precision, recall, F1 scores and AutoPRF [Hardmeier and Federico, 2010a].

Source		
	context	Oh, I hate **flies**. Look, there's another one!
	current sent.	Don't worry, I'll kill **it** for you.
Target		
1	context	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct	T'inquiète, je **la** tuerai pour toi.
	incorrect	T'inquiète, je **le** tuerai pour toi.
2	context	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct	T'inquiète, je **le** tuerai pour toi.
	incorrect	T'inquiète, je **la** tuerai pour toi.
3	context	Ô je déteste les **araignées**. Regarde, il y en a une autre !
	semi-correct	T'inquiète, je **la** tuerai pour toi.
	incorrect	T'inquiète, je **le** tuerai pour toi.
4	context	Ô je déteste les **papillons**. Regarde, il y en a un autre !
	semi-correct	T'inquiète, je **le** tuerai pour toi.
	incorrect	T'inquiète, je **la** tuerai pour toi.

Table 9: Co-reference test set block from [Bawden et al., 2018a]).

A last study worth mentioning is the work of Stanovsky et al. [2019] on gender bias in MT, where the authors develop a (very challenging) test set (*Winogender*) for pronoun translation from English into multiple target languages, and show that the system is much more likely to err when the reference pronoun is feminine - but also that the error rates for masculine pronoun is also very high.

4.1.2 Contrastive Evaluation

Contrastive test pairs have been extensively used to evaluate MT performance on specific phenomenon [Sennrich, 2017], including discourse related phenomena [Bawden et al., 2018a, Müller et al., 2018]. These require specifically designed sentence pairs to check/evaluate correct translations for the discourse phenomenon being studied.

The **Co-reference** test set by Bawden et al. [2018a] comprises 50 blocks with four translation pairs as shown in Table 9. The objective is to assess the use of a linguistic context on the target side. Each block is made of a source sentence countaing an anaphoric pronoun whose antecedent appears in a preceding context sentence. The choice of the correct gender for the French pronoun can only be made if one takes the previous sentence translation into accout, as no gender information is in the source. Each block comprises four versions of the same pattern, with a strict balance between the gender of the corrrect pronoun.

Müller et al. [2018] also presents a large-scale test suite using contrastive translations

for 12,000 difficult cases of pronouns for English to German extracted from the OpenSubtitle corpus[9]. In contrastive translation, the correct pronoun is swapped with an incorrect one as illustrated in Table 71. The pronoun "*it*" in English should translate to "*sie*" in German because of the antecedent " *bat*". For the contrastive pair, contextually wrong translations are produced by replacing "*sie*" with other pronouns "*er*" and "*es*". Another factor is the distance to the antecedent which is zero if the pronoun and its antecedent are in same sentence and do not require additional context for translation. If the model scores are higher for actual references than for the erroneous variant, then the model is considered capable to detect wrong pronouns.

EN:	a bat	antecedent
GR:	eine Fledermaus (f.)	antecedent
	1	antecedent distance
EN:	**It** could get tangled in your hair	source
GR:	**Sie** könnte sich in deinem Haar verfangen.	reference
GR:	**Er** könnte sich in deinem Haar verfangen.	contrastive
GR:	**Es** könnte sich in deinem Haar verfangen.	contrastive

Table 10: A sentence pair with two contrastive translations. An antecedent distance of 1 means that the antecedent is in the immediately preceding sentence, (example taken from [Müller et al., 2018])

Jwalapuram et al. [2019] present a pronoun test set for multiple source languages (Russian, German, Chinese and French) with English as the sole target language. Crucially, this test set was collected in a semi-automatic manner as follows: using WMT system submissions for years 2011 to 2015 and 2017, each output translation was automatically aligned with its reference, allowing to collect cases of pronoun mismatches between the output and the correct translation. These challenging examples of actual pronoun errors were finally valided by human judges. Additional 'noisy' examples were obtained by replacing a correct pronoun with an incorrect one found in the system output as illustrated in Table 11. Using this test-suite they then introduce *trainable evaluation measures* termed RC and CRC. A statistical model first learns to differentiate between good and bad output translations via pairwise comparison between a system output and reference translation using the same reference context. The corresponding scores can then be used in a contrastive evaluation. One weakness of this metric is that it is limited to evaluating translation into English.

The largest contrastive evaluation set for the translation of pronominal anaphora is presented in Lopes et al. [2020]. It contains 14K examples fron OpenSubtitle Lison and

[9] http://opus.nlpl.eu/OpenSubtitles2016.php

Original French input	**Il** était créatif, généreux,drôle, affectueux et talentueux, et il va beaucoup me manquer.
Reference translation	**He** was creative, generous, funny, loving and talented, and I will miss him dearly.
MT system translation	**It** was creative, generous, funny, affectionate and talented, and **we** will greatly miss.
Generated noisy example 1	**It** was creative, generous, funny, loving and talented, and I will miss him dearly.
Generated noisy example 2	He was creative, generous, funny, loving and talented, and **we** will miss him dearly.

Table 11: (FR-EN) 'Noisy' examples of pronouns errors, where the two mismatches between the system output and the reference give rise to two additional 'noisy' containing only one mismatch, taken from [Jwalapuram et al., 2019].

Tiedemann [2016], where the English 'it' can be translated into French as 'le' or 'la'. An extra annotation layer was further introduced where professional translators were asked to identify in the preceding sentences (source and target) the words that could contribute to disambiguate the correct translation [Yin et al., 2021]. This resource, SCAT (for '*Supporting Context for Ambiguous Translations*'), has recently been open-sourced by its authors.[10]

4.2 Conjunction, Deixis and Ellipsis

Popović [2019] offers a test suite for evaluating the disambiguation of conjunctions for document-level MT systems participating to the WMT19 shared translation tasks for French to German and English to German. The test suites do not rely on extra sentence level context but only focus on sentence pairs.

The contrastive testset of Voita et al. [2019a] already mentioned above contains contrasts aimed to evaluate the translation of deixis and ellipsis phenomena for the direction English to Russian. Each instance in the testset (3000 for deixis, 1000 for ellipsis) is made of reference sentences along with contrastive translations where an error has been infused. For deixis, the most frequent error category concerns the second person pronouns and is related to the inconsistency of T-V forms; the test set therefore mostly consists of formal and informal examples. To evaluate the processing of ellipsis, two ambiguities are addressed. The first concerns the prediction of a correct morphological form and second is related to verb phrase ellipsis. Example sentences are displayed in Table 2 in Section 1.4.

[10] https://github.com/neulab/contextual-mt. Note that it also contains contrastive pairs for English-French WSD.

4.3 Evaluating Cohesion and Coherence

In discourse analysis, cohesion is often studied together with coherence which is another dimension of the linguistic structure of a text. Cohesion is related to the surface structure, and can be assessed based on the analysis of word choices accross documents; coherence is more concerned with the underlying meaning connectedness [Xiong et al., 2013].

4.3.1 Evaluating lexical cohesion with references

Wong and Kit [2012] propose to use *lexical cohesion* along with standard reference-based metrics such as BLEU, TER and Meteor to evaluate machine translation models at the document level. The lexical cohesion devices considered in this work are based on the repetitions of content words (nouns, adjectives, adverbs or main verbs) within a document, including hyperonyms and synonyms, where WordNet is used to identify the semantic relationships. To analyse MT systems two ratios are defined: LC = (lexical cohesion devices/content words) and RC = (repetition/content words), which can be computed for human and machine translations. Higher rates of LC or RC means that a high portion of content words are found to take part to lexical cohesion.

Table 12 displays examples from MetricsMATR [Przybocki et al., 2008] data set.[11] They report LC and RC ratios to correspond well with human assessments; however at the document level, they are not as good as the other metrics, and should better be used in cunjunction with those, rather than as standalone metrics.

Going further, Gong et al. [2015] introduce a *gist consistency score* and *a cohesion score* along with available evaluation metrics to measure text cohesion. The gist consistency score is measured by building a topic model [Blei et al., 2003] and computing the topic distribution of the reference and model output for the evaluation set. The cohesion score is computed on lexical chains using lexical cohesion devices for content words. Note that they use a simplified notion of lexical chains where the focus is on reiterated stem-match words, which dispense with the use of a special thesaurus. These metrics are combined with BLEU and Meteor using a weighted average.

4.3.2 Cohesion and Coherence Contrastive Evaluations

Bawden et al. [2018a] also includes a contrastive test set aimed to evaluate cohesion and coherence.[12] The task is to select the right translation for a given word, and contains two kinds of difficult cases: one where the right choice is to repete a word occurring in the context (cohesion), one where the correct word sense can only be disambiguated looking at the context (coherence). Instances of the cohesion set are in Table 13.

Another useful resource for lexical cohesion evaluation is the test set of Voita et al. [2019a] comprising contrastive examples for English into Russian. Each of the 2000 instances in this test set comprises consists two reference sentences where a given (named) entity is translated consistently, and a constrast version exhibiting cases of inconsistencies.

[11] https://www.nist.gov/itl/iad/mig/metrics-machine-translation-evaluation
[12] https://github.com/rbawden/discourse-mt-test-sets/

	MT 1
1	Chine scrambled research on 16 key *technical*
2	These techniques are from within headline everyones boosting science and *technology* and *achieving* goals and contend of delivered on time bound through *achieving* breakthroughs in essential *technology* and complimentarity resources .
	BLEU: 0.224 (1-gram:7, 2-gram:0, 3-gram:2, 4-gram:1)
	LC: 0.107 (number of lexical cohesion devices: 5)
	Human assessment: 2.67
	MT 2
1	China is accelerating research 16 *main technologies*
2	These *technologies* are within the *important* realm to promote sciences and *technology* and *achieve* national goals and must be completed in a timely manner through *achieving main* discoveries in *technology* and integration of resources .
	BLEU: 0.213 (1-gram:5, 2-gram:3, 3-gram:2, 4-gram:1)
	LC: 0.231 (number of lexical cohesion devices: 9)
	Human assessment: 4.22
	Reference
1	China Accelerates Research on 16 Main Technologies
2	These technologies represent a significant part in the development of science and technology and the achievement of national goals. They must be accomplished within a fixed period of time by realizing breakthroughs in essential technologies and integration of resources.

Table 12: MT outputs of different quality (examples from [Wong and Kit, 2012]). N-grams that match the reference translation are underlined and italicized words represent lexical cohesion devices. The second MT output is better according to human assessment and their LC ratios present more variation. (see text for details).

4.4 Word Sense Disambiguation

Word sense disambiguation (WSD) is a well studied standalone task for which multiple test sets and test conditions have been developped, notably in the context of the Semeval shared tasks.[13] The ability to select the correct sense for polymous words also participates to the overall coherence of a translation, and we focus below solely on WSD in translation.

Adopting the same methodology of contrastive evaluation as Sennrich [2017], Rios Gonzales et al. [2017b] present ContraWSD, a set of test pairs aimed to evaluate the capability of a MT system to generate the correct sense of polysemous words in context. They pair a human reference translation with a set of contrastive examples, which include incorrect translations of semantically ambiguous source words. Using this data set, they are able to identify specific errors types of the model and to perform a quantitative analysis the model's ability to perform *lexical disambiguation*.

Using a smaller version of ContraWSD, this approach was offered as a supplementary

[13] https://semeval.github.io/

Source	
context	What's **crazy** about me?
current sent.	Is this **crazy**?
Target	
context	Qu'est-ce qu'il y a de **dingue** chez moi ?
correct	Est-ce que ça c'est **dingue** ?
incorrect	Est-ce que ça c'est fou ?
Source	
context	What's **crazy** about me?
current sent.	Is this **crazy**?
Target	
context	Qu'est-ce qu'il y a de **fou** chez moi ?
correct	Est-ce que ça c'est **fou** ?
incorrect	Est-ce que ça c'est dingue ?

Table 13: Cohesion and Coherence test set block (repetition) (taken from [Bawden et al., 2018a]). In both examples, the repetition of **crazy** in English has to be reproduced in French, either with the translation **dingue** or **fou**.

test suite at WMT18[14] [Rios et al., 2018] for the German to English translation direction. They evaluated all German-English systems submitted to the WMT18 News Translation task and found that the accuracy of the best system for the lexical disambiguation task improved from 81%to 93%, compared to the 2016 systems, and that this metrics had a strong correlation with BLEU scores.

Raganato et al. [2019] presented a multilingual test suite named as MUCOW, that includes contrastive sentence pairs for 16 language pairs built using word alignments and sense inventory of BabelNet. The contrastive test suite was used to evaluate ambiguous words for WMT19 News translation task.

Marvin and Koehn [2018] examined the representations of polysemous words at different levels in NMT encoding layers. They studied 4 polysemous words (*right*, *like*, *last*, and *case*) using the Europarl and News Commentary corpora. Their results did not reveal any clear conclusion due to the small size of the test data. Nonetheless, their methodology is a rare case of a fine-grained study of the word sense disambiguation capabilities of NMT systems.

4.5 Evaluating Discourse Structures

The evaluation of discourse structure translation has mostly given rise to reference-based metrics. Three of them are described below. Hajlaoui and Popescu-Belis [2013] introduce "Accuracy of Connective (ACT)" for evaluating translation of discourse connectives based on word alignment between the source, reference, and output sentences. Translations are automatically or semi-automatically scored using a dictionary of equivalent

[14] http://statmt.org/wmt18.

connectives. The metric first identifies discourse connectives in the source, reference and output sentence. If there are more than one translations, alignment information is used.

In the case of irrelevant alignment, word position is compared (six cases are considered for comparison). Using a dictionary of equivalents, the translations are scored automatically, or semi-automatically.

The precision of the metric is assessed by human judges on sample data for English/French and English/Arabic translations: the ACT scores are on average within 2% of human scores.

A more recent test suite for for evaluating discourse phenomena in MT systems is finally introduced in [Rysová et al., 2019] to assess the WMT19 News Translation Tasks for the English to Czech language direction.[15] Along with sentence-level errors, they identified discourse phenomenon at the document level that resulted in translation errors by manually inspecting output translations; these errors are related to topic articulation, lexicalization of connectives and discourse connectives.

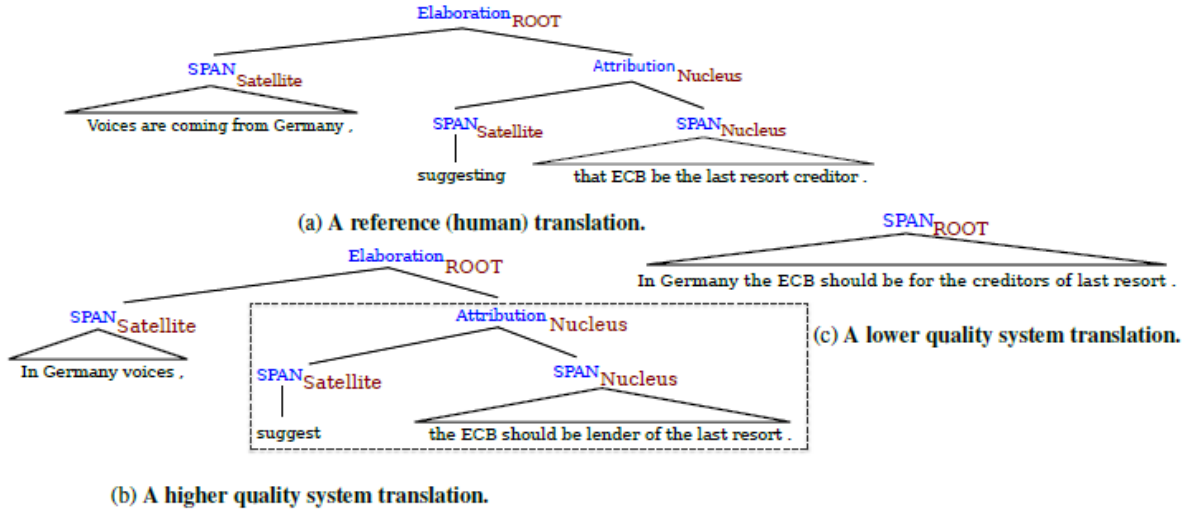

Figure 7: Using discourse trees to compare two system translations (reproduced from [Guzmán et al., 2014])

The Rhetorical Structure Theory (RST) [Mann and Thompson, 1988] is one of the most widely used discourse theories in computational linguistics. RST uses discourse relations between parts of texts to represent coherence in a form of hierarchical discourse trees (DTs). A DT represents text as a set of labelled hierarchical structures including linguistic information in layers capturing predicate-argument relationship and syntax.Guzmán et al. [2014] and Joty et al. [2017] have used discourse structures to design MT evaluation metrics specifically targeting discourse in MT. Evaluation is performed by measuring the similarity between the discourse trees of the reference and the automatic translation, with

[15] https://statmt.org/wmt19.

the assumption that good translation would preserve the discourse relations detected in the text. Figure 7 gives an example, where (a) is the reference translation, and (b) and (c) are two translations from participants to the WMT12 shared task.[16] Leaves in the DT represent text spans, where adjacent spans are joined through coherence relations (attribution, elaboration etc.) forming larger discourse units. Nuclei are the core parts while satellites are supportive parts. We see that nuclearity and labels are retained in translation (b) and not in (c) which probably makes (b) a better output than (c). The comparison between discourse trees is performed using tree kernels (TKs) [Collins and Duffy, 2002]. Guzmán et al. [2014] report improvements in overall MT metrics as well as a better correlation with human judgements even when using discourse information only at the sentence level.

Smith and Specia [2018] also propose to evaluate the output translation on the basis of a comparison with the source text; this comparison assesses the degree to which the discourse elements are preserved and correctly translated. The corresponding metric represents and compares actual and predicted discourse connectives based on pre-trained word embeddings, and combine this score with a discourse relationship "match score".

4.6 A global evaluation metrics: BLonDe

As explained above, global evaluation measures for context-aware MT are not many. The work of [Jiang et al., 2022] intends to close this gap, and introduces two new resources. First, a new collection of translated documents for the Chinese-English language pairs, comprising mostly fiction works (the "Bilingual Web Book". Part of this resource is annotated with translation errors.[17] Second, and more to the point for our purposes, the authors also introduce a new global metric that aggregates the performance of the MT system over a pre-defined set of discurse phenomenons.[18] In their study, the authors consider four main sources of errors defining *discourse categories*: (a) inconsistencies in refering to named entities, which happens when the same entity is translated in multiple ways across the document; (b) inconsistencies in the use of verb tense, which is especially important for Chinese-English translation; (c) errors in the translation of pronouns; (d) errors in the translation of discourse markers. In addition, the authors also consider an ad-hoc category for lexical content, based on n-grams occurrences.

For each category, they automatically annotate spans in the target and reference texts, where each span contains one occurrence of the target category (eg. a name entity, a tensed verb, etc). BlonDe is then defined based on automatic comparisons between the vector of spans in the hypothesis and in the reference translations, where the comparison score can be operationalized in several ways (eg. with precision and recall scores) that the authors present and discuss. In a systematic comparison with several MT systems and a diverse set of metrics, the authors show that BlonDe correlates well with human quality judgements.

[16] https://statmt.org/wmt12
[17] https://github.com/EleanorJiang/BlonDe#-the-bwb-dataset
[18] See https://github.com/EleanorJiang/BlonDe

4.7 Conclusion

Traditional MT metrics, such as BLEU, TER and Meteor are based on comparisons performed at the sentence level and are quite insensitive to changes aimed at improving a machine translation for complete documents. This has fostered the development of multiple proposals aimed at evaluating specific discourse-related phenomena in NMT. These initiatives, to date, have mostly been designed independently for each targeted linguistic phenomenon, with the study of pronominal anaphora attracting most of the attention.

As we have seen, the design of a new evaluation protocol can be achieved with two main strategies: comparison with a reference, or design of contrastive test sets. In both cases, designing a new metrics or test set requires a great amount of linguistic expertise, sophisticated annotation tools (eg. a RST parser) and human intervention, to capture the indented discourse phenomenon for the targeted language pair(s), which hinders their deployment as a universal solution for document-level MT evaluation.

Overall, it therefore appears that the development of generic evaluation strategies for context-aware MT by and large remains an open problem, a view that we share with Maruf et al. [2019b]. This suggests that human evaluation, no matter how costly, still remain the most effective ways to systematically detect inconsistencies and errors at the supra sentential level, and warrants the adoption of new protocols for direct assessment evaluations [Läubli et al., 2020].

5 Useful Resources

5.1 Corpora

In this section, we survey some resources available for building document-level NMT systems. These have also been mentioned and referred to in the corresponding sections. Most of corpora used in MT are only aligned at the sentence level without any metadata, which makes it difficult, if possible at all to reconstruct complete documents. This severely restricts the possibilities to focus on local and global discourse phenomenon, as there is no guarantee that the phenomenon under study is significantly represented in publicly available datasets. Even the prepared parallel dialogue datasets such as subtitle corpora sometimes lack speaker information. Table 14 lists some useful resources for building context-aware systems.

1. OpenSubtitles corpus consists of movies subtitles [Lison and Tiedemann, 2016] and can be used for building context aware systems. For example, Yun et al. [2020] used timestamped based approach to use context sentences in Opensubtitle corpus, considering start and end time of proceedings sentences.

2. TED Talks is collection of multilingual transcriptions of TED talks [Cettolo et al., 2012]. For document level MT, usually each talk is considered as a document. XML files for source and target language contains meta-data about document and sentence ids against each talk. TED Talks are also used for speech translation

Document Aligned Corpora:	
OpenSubtitle	http://opus.nlpl.eu/OpenSubtitles2016.php
Tilde Rapid	http://www.statmt.org/wmt20/translation-task.html
News Commentary	http://www.statmt.org/wmt20/translation-task.html
Europarl	http://www.statmt.org/wmt20/translation-task.html
Rotowire	https://github.com/harvardnlp/boxscore-data
TED Talks	https://wit3.fbk.eu/
ParaNatCom	https://www2.nict.go.jp/astrec-att/member/mutiyama/paranatcom/
Toolkits:	
[Miculicich et al., 2018]	https://github.com/idiap/HAN_NMT
[Zhang et al., 2018]	https://github.com/THUNLP-MT/Document-Transformer
[Kim et al., 2019]	https://github.com/ducthanhtran/sockeye_document_context
[Ye et al., 2019]	https://github.com/yzh119/BPT
[Li et al., 2020]	https://github.com/libeineu/Context-Aware

Table 14: Some useful resources for building context aware systems.

evaluation campaign as a recurrent shared task of the International Workshop on Spoken Language Translation (IWSLT).[19]

3. Tilde Rapid is parallel (EN-DE, CS-EN) data-set compiled from European Commission press releases between 1975 and 2016. Each press release is considered as a document, metadata includes document ids.

4. Rotowire consists of NBA basketball game summaries. There are 4853 distinct game summaries which are randomly split into train, dev and test set. Dataset is available both in json and text files. Each text file contains one game summary and considered as single document.

5. New Commentary-14 is multilingual data-set released for WMT series of shared task. Source and Target sentences are aligned into separate text files and documents are separated with blank lines.

6. Europarl v9 parallel data-set is extracted from proceedings of European Parliament. Europarl provides document distinctions for each document with document id.

7. ParaNatCom [Utiyama, 2019] is a parallel corpus (JA-EN) of abstract of scientific papers, used for the document level evaluation campaign run as a shared task of the Workshop on Asian Translation (WAT)[20].

[19] http://iwslt.org
[20] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/aspec_doc.html

5.2 Implementations

Several toolkits are available to implement context-aware NMT models. An overview of the architecture of each of these research works is given in Section 3.1. All of them use the Transformer [Vaswani et al., 2017a] architecture. *HAN_NMT* by Miculicich et al. [2018] extends OpenNMT-py [Klein et al., 2017] to incorporate a two-level model built by using hierarchical attention at sentence and document level. Kim et al. [2019] base their implementation on Sockeye [Hieber et al., 2018] and provide *outside-decoder*, *inside-decoder-sequential-attention* and *inside-decoder-parallel-attention* architectures. *BPT* by Ye et al. [2019] is written in DGL with PyTorch as backend. It builds graph neural networks allowing small scale spans for the local context and large scale span for the global or longer contexts. Li et al. [2020] is based on fairseq [Ott et al., 2019] allowing inference using *inside-context*, *outside-context* and *gaussian* models.

6 Conclusions

In this report, we have presented a comprehensive survey of recent research aimed to better take discourse phenomena into account in neural translation models. In section 1, we have presented a brief overview of the underlying linguistic phenomena, and of attempts to address them with the tools and concepts of statistical MT. From a bird's eye view, a recurring issue is the need to integrate source and target side information beyond the sentence/segment level which remains the default translation context in most systems.

In the two subsequent sections, we have detailled the computational and statistical challenges of representing and using a larger context in neural architectures: in Section 2, we notably discuss the basic Transformer architecture and various extensions aimed to improve its computational complexity, while Section 3 is more focused on better models of the document context, as well as improvements in the decoding architecture.

The last two sections have been devoted to resources: in Section 4, we have surveyed the existing test set and evaluation procedures that have been specifically designed with the purpose to assess the impact of using larger contexts in MT; useful training corpora recording document information in the test and/or train data as well as open implementations of document level MT are listed in Section 5.

As we discussed, the term of "context-aware" model is somewhat misleading, and covers a number of unrelated linguistic issues, some usually requiring a rather local view of the context (eg. anaphora resolution), while other (e.g. lexical consistency) need to have a much global vision of the document, spanning paragraphs, section, or even the entire document. Another important distinction is between phenomena that would need an analysis of the past target, and those which can be solved with just an extended source side window. All these phenomena are furthermore unevenly distributed across textual genres and registers: while the handling of pronouns, diexis and ellipsis is critical for systems translating dialogs or chats [Farajian et al., 2020], they are virtually unobserved for other document-level translation tasks (eg. TED talks or medical texts), for which issues related to lexical consistency and argumentative structure are more important. Another obstacle that still hinders the development is the lack of automatic tools that

would help identify and quantify the main types of errors, and guide progresses towards their resolution.

In the near future, as new architectures capable of effectively handling multi-sentencial contexts become more mature, it is likely that the number of errors caused by a too narrow analysis window will progressively decrease; handling document-level context is more challenging and would require to analyze, represent and integrate large chunks of information, so as to correctly generate consistent documents and to reproduce the main rhetorical structures and their overall organization.

Bibliography

Ruchit Agrawal, Marco Turchi, and Matteo Negri. Contextual handling in neural machine translation: look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 11–20, Alacant, Spain, 2018. European Association for Machine Translation. [Cited on pages 27, 28, and 29.]

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. doi: 10.18653/v1/W19-5351. URL https://www.aclweb.org/anthology/W19-5351. [Cited on page 39.]

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL http://arxiv.org/abs/1409.0473. [Cited on pages 6, 18, and 20.]

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*, pages 65–72, Ann Arbor, Michigan, 2005. URL http://www.aclweb.org/anthology/W/W05/W05-0909. [Cited on pages 7 and 38.]

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, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://www.aclweb.org/anthology/N18-1118. [Cited on pages 10, 17, 27, 32, 39, 41, 44, and 46.]

Rachel Bawden, Rico Sennrich, and Barry Birch, Alexandraand 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.

Association for Computational Linguistics, 2018b. doi: 10.18653/v1/N18-1118. URL http://aclweb.org/anthology/N18-1118. [Cited on pages 6, 28, and 30.]

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurélie Névéol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5403. URL https://www.aclweb.org/anthology/W19-5403. [Cited on page 6.]

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL http://arxiv.org/pdf/2004.05150. [Cited on page 23.]

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. [Cited on page 18.]

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of the Machine Translation Summit*, MT Summit XIV, pages 35–42, Nice, France, 2013. URL http://www.mtsummit2013.info/files/proceedings/main/mt-summit-2013-bertoldi-et-al.pdf. [Cited on page 33.]

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG supertags in factored statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W07-0702. [Cited on pages 14 and 17.]

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning research*, 3:993–1022, 2003. URL https://jmlr.csail.mit.edu/papers/v3/blei03a.html. [Cited on pages 7 and 44.]

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990. URL https://www.aclweb.org/anthology/J90-2002. [Cited on page 18.]

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL https://www.aclweb.org/anthology/J93-2003. [Cited on pages 18 and 21.]

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell,

Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL http://arxiv.org/pdf/2005.14165. [Cited on page 23.]

Marine Carpuat. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-2404. [Cited on pages 7 and 14.]

Marine Carpuat and Michel Simard. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-3156. [Cited on page 14.]

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D07-1007. [Cited on pages 7, 15, and 17.]

Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May 2012. [Cited on pages 37 and 49.]

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.autosimtrans-1.5. URL https://www.aclweb.org/anthology/2020.autosimtrans-1.5. [Cited on pages 27 and 31.]

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219873. URL https://www.aclweb.org/anthology/P05-1033. [Cited on page 18.]

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. [Cited on page 23.]

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings*

of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111, Doha, Qatar, October 2014a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W14-4012. [Cited on pages 18 and 20.]

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014b. [Cited on page 18.]

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, pages 1724–1734, Doha, Qatar, 2014c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D14-1179. [Cited on page 20.]

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1102. URL https://www.aclweb.org/anthology/N16-1102. [Cited on page 21.]

Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *Advances in neural information processing systems*, pages 625–632, 2002. [Cited on page 48.]

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. SYSTRAN's pure neural machine translation systems. *CoRR*, abs/1610.05540, 2016. URL http://arxiv.org/pdf/1610.05540. [Cited on page 18.]

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, 2014. [Cited on page 18.]

Shuoyang Ding, Hainan Xu, and Philipp Koehn. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on*

Machine Translation (Volume 1: Research Papers), pages 1–12, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5201. URL https://www.aclweb.org/anthology/W19-5201. [Cited on page 21.]

Félix do Carmo, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 2020. doi: 10.1007/s10590-020-09252-y. URL https://doi.org/10.1007/s10590-020-09252-y. [Cited on page 35.]

Radina Dobreva, Jie Zhou, and Rachel Bawden. Document sub-structure in neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3657–3667, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.451. [Cited on pages 33 and 34.]

M. Amin Farajian, António V. Lopes, André F.T. Martins, Sameen Maruf, and Gholamreza Haffari. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.3. [Cited on page 51.]

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.505. URL https://aclanthology.org/2021.acl-long.505. [Cited on page 30.]

Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1004. [Cited on page 21.]

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. URL https://www.aclweb.org/anthology/D11-1084. [Cited on pages 7 and 33.]

Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural*

Language Learning, pages 276–285, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1026. [Cited on page 7.]

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. doi: 10.18653/v1/W15-2504. URL https://www.aclweb.org/anthology/W15-2504. [Cited on pages 15, 17, 39, and 44.]

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1154. URL https://www.aclweb.org/anthology/P16-1154. [Cited on page 33.]

Liane Guillou. Analysing lexical consistency in translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3302. [Cited on pages 14, 17, and 39.]

Liane Guillou and Christian Hardmeier. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1100. [Cited on page 39.]

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 525–542, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2345. URL https://www.aclweb.org/anthology/W16-2345. [Cited on pages 6, 10, 38, and 40.]

Liane Kirsten Guillou. *Incorporating pronoun function into statistical machine translation*. PhD thesis, The University of Edinburgh, 2016. [Cited on pages 7, 10, 17, and 40.]

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-transformer. 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 1315–1325, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1133. URL https://www.aclweb.org/anthology/N19-1133. [Cited on page 8.]

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1065. URL https://www.aclweb.org/anthology/P14-1065. [Cited on pages 39, 47, and 48.]

Najeh Hajlaoui and Andrei Popescu-Belis. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 236–247. Springer, 2013. [Cited on pages 10, 39, and 46.]

Anissa Hamza. *La détection et la traduction automatiques de l'ellipse. Enjeux théoriques et pratiques*. PhD thesis, UFR Sciences du Langage, Université de Strasbourg, 2019. [Cited on pages 11 and 12.]

Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation, Paris.*, page 283–289, 2010a. URL http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-154490. [ed] Marcello Federico et al. [Cited on pages 6, 7, 39, and 40.]

Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT, pages 283–289, Paris, France, 2010b. [Cited on pages 10 and 40.]

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 2013a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-4033. [Cited on pages 6 and 7.]

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 2013b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1037. [Cited on page 7.]

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2501. URL https://www.aclweb.org/anthology/W15-2501. [Cited on pages 6, 7, 10, and 40.]

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. Dynamic topic adaptation for phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–337, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1035. URL https://www.aclweb.org/anthology/E14-1035. [Cited on page 7.]

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*, 2018. [Cited on page 6.]

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/W18-1820. [Cited on page 51.]

Carla Huls, Edwin Bos, and Wim Claassen. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79, 1995. URL https://www.aclweb.org/anthology/J95-1003. [Cited on page 10.]

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 602–614, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020wmt-1.71. [Cited on page 39.]

Sébastien Jean and Kyunghyun Cho. Context-aware learning for neural machine translation, 2019. URL http://arxiv.org/pdf/1903.04715. [Cited on page 28.]

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. Does neural machine translation benefit from larger context?, 2017. URL http://arxiv.org/abs/1704.05135. [Cited on pages 28 and 32.]

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.111. URL https://aclanthology.org/2022.naacl-main.111. [Cited on page 48.]

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722, Decem-

ber 2017. doi: 10.1162/COLI_a_00298. URL https://www.aclweb.org/anthology/J17-4001. [Cited on pages 38, 39, and 47.]

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. doi: 10.18653/v1/W19-5321. URL https://www.aclweb.org/anthology/W19-5321. [Cited on pages 17, 27, 28, 29, and 37.]

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 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1294. URL https://www.aclweb.org/anthology/D19-1294. [Cited on pages 10, 39, 40, 42, and 43.]

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 1700–1709, Seattle, Washington, USA, 2013. URL http://aclweb.org/anthology/D13-1176. [Cited on page 18.]

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.175. [Cited on page 35.]

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. doi: 10.18653/v1/D19-6503. URL https://www.aclweb.org/anthology/D19-6503. [Cited on pages 8, 27, 29, 36, 50, and 51.]

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *Proceedings of the International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB. [Cited on page 24.]

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012. [Cited on page 51.]

Philip Koehn. *Neural Machine Translation*. Cambridge University Press, 2020. [Cited on page 18.]

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010. [Cited on pages 18 and 19.]

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1091. [Cited on page 13.]

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3204. URL http://aclweb.org/anthology/W17-3204. [Cited on page 21.]

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistic*, pages 127–133, Edmondton, Canada, 2003. [Cited on page 18.]

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic, 2007. [Cited on page 18.]

Robert Krovetz. More than one sense per discourse. *NEC Princeton NJ Labs., Research Memorandum*, 23, 1998. [Cited on page 11.]

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. URL https://www.aclweb.org/anthology/C18-1051. [Cited on pages 28 and 32.]

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, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1050. [Cited on pages 6, 18, 27, 28, 32, 33, and 34.]

Roland Kuhn and Renato DeMori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6): 570–583, 1990. [Cited on page 33.]

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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://www.aclweb.org/anthology/D18-1512. [Cited on page 6.]

Hai-Son Le, Alexandre Allauzen, and François Yvon. Continuous space translation models with neural networks. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1005. [Cited on page 18.]

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. URL https://www.aclweb.org/anthology/W10-1737. [Cited on pages 7, 9, and 17.]

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.322. URL https://www.aclweb.org/anthology/2020.acl-main.322. [Cited on pages 8, 27, 30, 32, 36, 37, 50, and 51.]

Hongzheng Li, Philippe Langlais, and Yaohong Jin. Translating implicit discourse connectives based on cross-lingual annotation and alignment. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 93–98, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4812. URL https://www.aclweb.org/anthology/W17-4812. [Cited on pages 14 and 17.]

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1124. URL https://www.aclweb.org/anthology/P19-1124. [Cited on page 21.]

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1147. [Cited on pages 29, 42, and 49.]

Frederick Liu, Han Lu, and Graham Neubig. Handling homographs 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 1336–1345, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-1121. URL https://www.aclweb.org/anthology/N18-1121. [Cited on pages 16 and 17.]

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018b. URL https://openreview.net/forum?id=Hyg0vbWC-. [Cited on page 23.]

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4801. URL https://www.aclweb.org/anthology/W17-4801. [Cited on pages 9, 10, 38, and 40.]

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André Martins. Document-level neural MT: A systematic comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, 2020. [Cited on pages 6, 8, 17, 39, and 42.]

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Review*, 67:653–672, 2020. [Cited on page 49.]

Shuming Ma, Dongdong Zhang, and Ming Zhou. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.321. URL https://www.aclweb.org/anthology/2020.acl-main.321. [Cited on pages 8 and 29.]

Zhiyi Ma, Sergey Edunov, and Michael Auli. A comparison of approaches to document-level machine translation. arXiv preprint arXiv:1910.07481, 2021. [Cited on pages 6 and 37.]

Valentin Macé and Christophe Servan. Using whole document context in neural machine translation. arXiv preprint arXiv:1910.07481, 2019. URL http://arxiv.org/abs/1910.07481. [Cited on page 28.]

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988. [Cited on page 47.]

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. doi: 10.18653/v1/P18-1118. URL https://www.aclweb.org/anthology/P18-1118. [Cited on pages 27, 28, 30, and 33.]

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. doi: 10.18653/v1/W18-6311. URL https://www.aclweb.org/anthology/W18-6311. [Cited on page 28.]

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, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1313. URL https://www.aclweb.org/anthology/N19-1313. [Cited on pages 17, 27, and 28.]

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level machine translation: Methods and evaluation. arXiv preprint arXiv:1912.08494, 2019b. URL http://arxiv.org/abs/1912.08494. [Cited on pages 6, 8, 28, 37, 39, and 49.]

Rebecca Marvin and Philipp Koehn. Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 125–131, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/W18-1812. [Cited on pages 6, 15, 17, and 46.]

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, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4813. URL https://www.aclweb.org/anthology/W17-4813. [Cited on pages 15 and 17.]

Thomas Meyer. Disambiguating temporal-contrastive connectives for machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 46–51, Portland, OR, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-3009. [Cited on pages 7, 13, and 17.]

Thomas Meyer and Lucie Poláková. Machine translation with many manually labeled discourse connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia, Bulgaria, August 2013. Association for Computational Lin-

guistics. URL https://www.aclweb.org/anthology/W13-3306. [Cited on pages 7, 13, 14, 17, and 39.]

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, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-0117. [Cited on pages 14 and 17.]

Thomas Meyer and Bonnie Webber. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3303. [Cited on pages 12 and 13.]

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, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2022. [Cited on page 17.]

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012. [Cited on pages 13, 17, and 39.]

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:1184–1197, 2015. [Cited on pages 7 and 39.]

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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://www.aclweb.org/anthology/D18-1325. [Cited on pages 17, 27, 28, 30, 31, 50, and 51.]

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 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-4802. URL https://www.aclweb.org/anthology/W17-4802. [Cited on page 31.]

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

2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-4802. URL https://www.aclweb.org/anthology/W17-4802. [Cited on pages 39 and 40.]

Ruslan Mitkov. Introduction: special issue on anaphora resolution in machine translation and multilingual NLP. *Machine translation*, pages 159–161, 1999. [Cited on page 7.]

Ruslan Mitkov et al. Anaphora resolution in machine translation. In *Proceedings of the Sixth International conference on Theoretical and Methodological issues in Machine Translation*. Citeseer, 1995. [Cited on pages 9 and 17.]

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991. URL https://www.aclweb.org/anthology/J91-1002. [Cited on page 14.]

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. doi: 10.18653/v1/W18-6307. URL https://www.aclweb.org/anthology/W18-6307. [Cited on pages 10, 39, 41, and 42.]

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://www.aclweb.org/anthology/W18-6322. [Cited on page 26.]

Masaaki Nagata and Makoto Morishita. A test set for discourse translation from Japanese to English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.457. [Cited on page 39.]

Laurent Nepveu, Guy Lapalme, Philippe Langlais, and George Foster. Adaptive language and translation models for interactive machine translation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 190–197, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3225. [Cited on page 33.]

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073133. URL https://www.aclweb.org/anthology/P02-1038. [Cited on page 18.]

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. URL https://www.aclweb.org/anthology/J03-1002. [Cited on page 21.]

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL https://www.aclweb.org/anthology/N19-4009. [Cited on page 51.]

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040. [Cited on pages 7 and 38.]

Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors. *Anaphora Resolution: Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer Verlag, Gmbh, Berlin, Heidelberg, 2016. [Cited on page 8.]

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381, 2020. doi: 10.1038/s41467-020-18073-9. URL https://doi.org/10.1038/s41467-020-18073-9. [Cited on page 6.]

Andrei Popescu-Belis. Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:1901.09115*, 2019. URL http://arxiv.org/pdf/1901.09115. [Cited on pages 8 and 38.]

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. doi: 10.18653/v1/W19-5353. URL https://www.aclweb.org/anthology/W19-5353. [Cited on pages 39 and 43.]

Mark Przybocki, Kay Peterson, and Sébastien Bronsart. NIST metrics for machine translation (MetricsMATR08) development data, 2008. [Cited on page 44.]

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. doi:

10.18653/v1/W19-5354. URL https://www.aclweb.org/anthology/W19-5354. [Cited on pages 39 and 46.]

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. doi: 10.18653/v1/W18-6437. URL https://www.aclweb.org/anthology/W18-6437. [Cited on pages 6, 38, 39, and 46.]

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 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-4702. URL https://www.aclweb.org/anthology/W17-4702. [Cited on pages 6 and 7.]

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 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-4702. URL https://www.aclweb.org/anthology/W17-4702. [Cited on pages 15, 17, 28, 39, and 45.]

Kateřina 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. doi: 10.18653/v1/W19-5352. URL https://www.aclweb.org/anthology/W19-5352. [Cited on pages 39 and 47.]

Danielle Saunders, Felix Stahlberg, and Bill Byrne. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7764–7770, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.693. URL https://www.aclweb.org/anthology/2020.acl-main.693. [Cited on page 27.]

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. doi: 10.18653/v1/D19-6506. URL https://www.aclweb.org/anthology/D19-6506. [Cited on pages 28 and 37.]

Holger Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012: Posters*, pages 1071–1080, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://www.aclweb.org/anthology/C12-2104. [Cited on page 18.]

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. URL https://www.aclweb.org/anthology/E17-2060. [Cited on pages 38, 39, 41, and 45.]

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162. [Cited on page 25.]

Karin Sim Smith and Lucia Specia. Assessing crosslingual discourse relations in machine translation. *arXiv preprint arXiv:1810.03148*, 2018. [Cited on pages 39 and 48.]

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA, 2006. URL http://www.mt-archive.info/AMTA-2006-Snover.pdf. [Cited on page 38.]

Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Review*, 69:343–418, 2020. [Cited on page 18.]

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://www.aclweb.org/anthology/D19-1331. [Cited on page 26.]

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://www.aclweb.org/anthology/P19-1164. [Cited on page 41.]

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, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6306. URL https://www.aclweb.org/anthology/W18-6306. [Cited on page 32.]

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, August

2019. European Association for Machine Translation. URL https://www.aclweb.org/anthology/W19-6614. [Cited on page 28.]

Sara Stymne, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. Feature weight optimization for discourse-level SMT. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 60–69, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3308. [Cited on page 7.]

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 459–468, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P12-1048. [Cited on page 7.]

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1032. URL https://www.aclweb.org/anthology/P19-1032. [Cited on page 29.]

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 3104–3112. 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf. [Cited on page 6.]

Gongbo Tang, Rico Sennrich, and Joakim Nivre. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6304. URL https://www.aclweb.org/anthology/W18-6304. [Cited on pages 6 and 17.]

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient Transformers: A survey, 2020. URL http://arxiv.org/pdf/2009.06732. [Cited on pages 23 and 24.]

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, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-2602. [Cited on page 33.]

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, September 2017. Association for Computational Linguistics.

doi: 10.18653/v1/W17-4811. URL https://www.aclweb.org/anthology/W17-4811. [Cited on pages 8, 27, 28, 29, and 31.]

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. doi: 10.1162/tacl_a_00029. URL https://www.aclweb.org/anthology/Q18-1029. [Cited on pages 8, 18, 27, 28, 32, 33, and 34.]

Masao Utiyama. ParaNatCom — Parallel English-Japanese abstract corpus made from Nature Communications articles, 2019. [Cited on page 50.]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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., 2017a. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. [Cited on pages 6, 20, and 51.]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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., 2017b. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. [Cited on page 22.]

Elena Voita, 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://www.aclweb.org/anthology/P18-1117. [Cited on pages 6, 18, 28, 30, 31, and 36.]

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 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://www.aclweb.org/anthology/P19-1116. [Cited on pages 10, 11, 12, 15, 17, 18, 28, 36, 39, 43, and 44.]

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 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1081. URL https://www.aclweb.org/anthology/D19-1081. [Cited on pages 10, 11, 15, 17, 27, 28, and 35.]

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. doi: 10.18653/v1/W19-5355. URL https://www.aclweb.org/anthology/W19-5355. [Cited on pages 38 and 39.]

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 2017a. Association for Computational Linguistics. doi: 10.18653/v1/D17-1301. URL https://www.aclweb.org/anthology/D17-1301. [Cited on pages 18, 28, 30, and 32.]

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. Memory-enhanced decoder for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 278–286, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1027. URL https://www.aclweb.org/anthology/D16-1027. [Cited on page 31.]

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020. URL http://arxiv.org/pdf/2006.04768. [Cited on page 24.]

Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. Translating phrases in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Copenhagen, Denmark, September 2017b. Association for Computational Linguistics. doi: 10.18653/v1/D17-1149. URL https://www.aclweb.org/anthology/D17-1149. [Cited on page 6.]

Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. Improving conditioning in context-aware sequence to sequence models, 2019. URL http://arxiv.org/abs/1911.09728. [Cited on page 28.]

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, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1097. [Cited on pages 31, 39, 44, and 45.]

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.530. URL https://www.aclweb.org/anthology/2020.acl-main.530. [Cited on pages 6, 36, 39, and 40.]

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. URL https://arxiv.org/abs/1609.08144. [Cited on page 18.]

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. URL https://www.aclweb.org/anthology/D13-1163. [Cited on pages 15, 17, and 44.]

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345, 2019. [Cited on pages 18, 27, 28, and 35.]

Hayahide Yamagishi and Mamoru Komachi. Improving context-aware neural machine translation with target-side context, 2019. URL http://arxiv.org/abs/1909.00531. [Cited on page 28.]

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. Enhancing context modeling with a query-guided capsule network for document-level 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 1527–1537, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1164. URL https://www.aclweb.org/anthology/D19-1164. [Cited on pages 18, 27, 33, and 34.]

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. BP-Transformer: Modelling long-range context via binary partitioning, 2019. URL http://arxiv.org/pdf/1911.04070. [Cited on pages 23, 24, 50, and 51.]

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F.T. Martins, and Graham Neubig. Do context-aware translation models pay the right attention? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 788–801, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.65. URL https://aclanthology.org/2021.acl-long.65. [Cited on page 43.]

Rika Yoshii. JETR: A robust machine translation system. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 25–31, Stanford, California, USA, July 1987. Association for Computational Linguistics. doi: 10.3115/981175.981179. URL https://www.aclweb.org/anthology/P87-1004. [Cited on pages 11 and 17.]

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360, 2020.

doi: 10.1162/tacl_a_00319. URL https://www.aclweb.org/anthology/2020.tacl-1.23. [Cited on pages 28 and 35.]

Hyeongu Yun, Yongkeun Hwang, and Kyomin Jung. Improving context-aware neural machine translation using self-attentive sentence embedding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 9498–9506. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6494. [Cited on page 49.]

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. doi: 10.18653/v1/W15-2519. URL https://www.aclweb.org/anthology/W15-2519. [Cited on pages 14 and 17.]

François Yvon and Sadaf Abdul Rauf. Utilisation de ressources lexicales et terminologiques en traduction neuronale. Research Report 2020-001, LIMSI-CNRS, July 2020. URL https://hal.archives-ouvertes.fr/hal-02895535. [Cited on page 35.]

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences, 2020. URL http://arxiv.org/pdf/2007.14062. [Cited on page 23.]

Francisco Zamora-Martinez, Maria José Castro-Bleda, and Holger Schwenk. N-gram-based machine translation enhanced with neural networks for the French-English btec-iwslt'10 task. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2010*, Paris, France, 2010. [Cited on page 18.]

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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1049. URL https://www.aclweb.org/anthology/D18-1049. [Cited on pages 8, 18, 28, 30, 31, 32, and 50.]

Rong Zhang and Abraham Ittycheriah. Novel document level features for statistical machine translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 153–157, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2520. URL https://www.aclweb.org/anthology/W15-2520. [Cited on pages 15, 16, and 17.]

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. Towards making the most of context in neural machine translation. In Christian Bessiere,

editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/551. URL https://doi.org/10.24963/ijcai.2020/551. [Cited on pages 6, 18, and 27.]

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 369–378, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.41. [Cited on pages 16, 17, and 39.]