MeMAD

Methods for Managing Audiovisual Data

memad.eu
info@memad.eu

Twitter – @memadproject
LinkedIn – MeMAD Project

MeMAD Deliverable

D4.2 Report on Discourse-Aware Machine Translation for Audiovisual Data

Grant agreement number	780069
Action acronym	MeMAD
Action title	Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy
Funding scheme	H2020–ICT–2016–2017/H2020–ICT–2017–1
Version date of the Annex I against which the assessment will be made	3.10.2017
Start date of the project	1.1.2018
Due date of the deliverable	31.12.2019
Actual date of submission	30.12.2019
Lead beneficiary for the deliverable	University of Helsinki
Dissemination level of the deliverable	Public

Action coordinator's scientific representative
Prof. Mikko Kurimo
AALTO–KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics
mikko.kurimo@aalto.fi

MeMAD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069. This document has been produced by theMeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Authors in alphabetical order		
Name	Beneficiary	e-mail
Maija Hirvonen	University of Helsinki Tampere University	maija.hirvonen@helsinki.fi
Maarit Koponen	University of Helsinki	maarit.koponen@helsinki.fi
Umut Sulubacak	University of Helsinki	umut.sulubacak@helsinki.fi
Jörg Tiedemann	University of Helsinki	jorg.tiedemann@helsinki.fi

Internal reviewers in alphabetical order		
Name	Beneficiary	e-mail
Sebastian Andersson	Lingsoft	sebastian.andersson@lingsoft.fi
Sabine Braun	University of Surrey	s.braun@surrey.ac.uk

Abstract

Machine translation is conventionally based on processing isolated textual sentences, disregarding the broader context around them, as well as any cues that are not explicit in text. This formulation is unable to reliably determine referential discourse phenomena such as anaphora and connectives, which creates a barrier to the production of cohesive and coherent language. While exploiting the audio and visual modalities provide a window into the context, machine translation must also venture beyond the sentence, and become aware of the discourse in the entire document. There has been a surge of research in discourse-aware machine translation in the last two decades, primarily focusing on the design and evaluation of document-level machine translation systems, and expanding the textual context in machine translation architectures without compromising computational efficiency. Following the recent advances, we have likewise put substantial effort into the development of discourse-aware machine translation systems, albeit to no considerable improvement. In this deliverable, we start by breaking down the most salient phenomena to explain their relevance, and presenting a brief survey of successful approaches to discourse-aware machine translation, followed by descriptions of the models we have developed within the WP4 of the MeMAD project. We dedicate the rest of the report to our analysis of user evaluation data collected in an experiment where professional translators tested post-editing of machine translation for subtitling. Finally, we conclude our report with discussions of future directions in utilising dedicated subtitle translation systems, speaker and dialogue information, and end-to-end speech translation models in the last year of the project.

Contents

1 Introduction

Developments in machine translation (MT) in the last two decades have led to significant improvements in translation quality. The success and popularity of statistical machine translation (SMT) systems such as Moses (Koehn et al., 2007) were matched and eventually surpassed by neural machine translation (NMT) architectures (Cho et al., 2014; Sutskever et al., 2014). The paradigm shift became even more widespread with the introduction of the attention mechanism (Bahdanau et al., 2015) to address some of the well-recognised drawbacks of NMT, culminating in the Transformer architecture (Vaswani et al., 2017), which has now become a staple in the MT community. For all their successes, these advances in MT came at the cost of drifting away from research being done in the field of Translation Studies (Hardmeier, 2014). From a Translation Studies oriented perspective, questions of quality, for example, tend to take a broader view exploring definitions of quality in specific contexts, for specific users and specific purposes, as well as view points on not only the quality of the *product* of translation, but also *process* quality and *social* quality (for a recent overview, see Moorkens et al., 2018). The discrepancy between the Translation Studies and Machine Translation fields stems from the fact that, in contrast to the wider range of approaches in Translation Studies to what translation may entail, MT is a highly conventionalised task with a strict formulation and specialised evaluation methods. This entails an inherent bias, meaning that MT accounts for particular aspects of translation, while tending to neglect others simply owing to the conventions around how it is formulated as a task.

In deliverable D4.1, we previously addressed how multimodal extensions to conventional text-based MT motivates research by opening up various possibilities in deliverable D4.1. Similar to how multimodality relaxes the restriction of using text input, another undercurrent of research proposes superseding the constraint of using sentences as translation units in order to attain discourse-awareness (Hardmeier, 2014; Loáiciga, 2017). Recently, Bawden (2018) emphasises that the translation of sentences in isolation is among the most striking approximations in MT, motivated by computational restrictions and the relatively niche utility of extra-sentential context. For the most part, the body of work on the utility of leveraging document-level context is concentrated on a number of linguistic concepts (see Section 2). Nevertheless, these concepts are often pervasive in language (e.g. using correct genders, producing cohesive language), and an MT system's ability to consider such phenomena could dictate its credibility as an emerging technology.

In this report, we first discuss phenomena relevant for discourse particularly in the context of (machine) translation, as well as the connection between discourse and multimodality (Section 2), followed by the state of the art in approaches to discourse-aware machine translation and issues related to the evaluation of discourse phenomena in machine translation (Section 3). We then present the work carried out in MeMAD on discourse-aware machine translation (Section 4) and focus on a user evaluation experiment involving a comparison of sentence-level and document-level machine translation for subtitling (Section 5). Finally, we present directions for future work and conclusion (Sections 6 and 7).

2 Phenomena relevant for discourse

The vast majority of the data used in MT research comprise text documents or spoken language transcripts in some form. Of these, text datasets typically have aligned segments that correspond to parallel sentences. The same is true for spoken language transcripts, where the

data would be formatted as a series of aligned segments (e.g. video subtitles). For consistency, we will refer to all types of segments as sentences and aligned segments as translation units. Many of the most prominent MT datasets (such as Koehn, 2005; Lison and Tiedemann, 2016) also organise sentences in cohesive divisions (e.g. session proceedings, movie texts), which we will collectively refer to as documents. Although a document appears to be natural sequences of sentences, processing them independently comes at the cost of obscuring complex linguistic dependencies that should rather be considered in translation (Hardmeier, 2012, 2014). Regardless, conventional MT systems are built on sentence-level architectures that are unable to process document-level context, in the same way monolingual MT systems disregard speech audio while translating speech transcripts.

The fact that both natural language understanding and generation rely on the accessibility of discourse information to be fully realised is highly consequential for translation. The larger context in which the translation is taking place is the only cue for certain aspects of language, such as inferring the register, speaker intent, attitude and style, choosing the correct words for the occasion and maintaining a consistent vocabulary, and maintaining the flow in discursive settings like conversations (Bawden, 2018). Separate from natural language processing, discourse analytical theories and methods have achieved popularity in the field of Translation Studies as a useful way to analyse linguistic structure and meaning in the context of human translation (Munday, 2012). Work on discourse in human translation has focused on linguistic themes such as cohesion and coherence in translation, genre and register analysis, but in recent years e.g. semiotic issues and multimodality (see Section 2.2) as well as extralinguistic themes such as power and ideology have gained prominence (Zhang et al., 2015).

2.1 Cohesion and coherence

The linguistic concepts of cohesion and coherence (see Sanders and Pander Maat, 2006) constitute two essential discourse properties that establish congruence within the document. Going beyond intra-sentential context, both properties are concerned with elements that register incrementally to dictate the interpretation of the whole document (Halliday and Hasan, 2014). Cohesion is specifically linked with surface properties such as word choice and coreference relations. Some examples of incohesive translation could be translating repeating words differently, or a polysemous word with the wrong sense, and misattributing the gender or number agreement of a pronoun with respect to its antecedent. An example of cohesive vs incohesive translation of the polysemous word *party* can be seen in the following sample from a political debate. The first MT output (by the document-level system described in Section 5.1) has correctly rendered it with the Finnish *Puolueeni* 'political faction', while the second (sentence-level system, see Section 5.1) mistranslates it as *juhlani* 'my festivities', instead:

Source	Strict border control and humanitarian responsibility. My party did it.
doc-MT	Tiukka rajavalvonta ja humanitaarinen vastuu. Puolueeni teki sen.
sent-MT	Tiukka rajavalvonta ja humanitaarinen vastuu. Minun juhlani tekivät sen.

Conversely, coherence involves maintaining a consistent mental model throughout the document, relating to semantics and understanding. For instance, mistranslating ambiguous verb tenses, and omitting or inappropriately using discourse connectives, could result in incoherent translations. In the following example, the question tag (*Ai et vai?*) following the question is intended to communicate surprise. While both MT versions are likely to be comprehensible, the use of the auxiliary verb *didn't* in the second one (sentence-level system) breaks the coherence of the text:

Source	Oot sä käyny ikinä metrossa? Ai et vai?
doc-MT	Have you ever been to the subway? Haven't you?
sent-MT	Have you ever been to the subway? Oh, you didn't?

From a natural language processing standpoint, lexical disambiguation is at the crux of most practical studies on leveraging discourse context (e.g. Xiao et al., 2011; Loáiciga et al., 2017a). Pronouns often comprise a fitting case study of this problem, since discourse context plays a large role in determining how they should be translated. A pronoun is not always best translated as a pronoun (e.g. expletive pronouns as in *"It is snowing"*), and even in case of direct correspondence, the ambiguity from the source language to the target language can be extreme. For example, Loáiciga (2017) notes that the English pronoun *"it"* has 14 options when translated into French (*"il"*, *"elle"*, *"la"*, *"le"*, *"l'"*, *"lui"*, *"cela"*, *"celui"*, *"celui-ci"*, *"celle-là"*, *"ce"*, *"c'"*, *"en"*, *"y"*). Moreover, other cases like pronoun dropping in languages such as Finnish and Turkish can pose a contextually-conditioned choice that creates nuances in meaning, further complicating this issue.

Despite the emphasis on pronouns, there are other discourse phenomena that discourse-aware MT could resolve by integrating document context and multimodality. In his dissertation on the subject, Hardmeier (2014) presents an in-depth examination of the subject matter in four main categories: (1) Pronominal anaphora, (2) Noun phrase definiteness, (3) Verb tense and aspect, and (4) Discourse connectives. While the relative prevalence of each item varies across different languages and contexts, they comprise the recurring challenges in discourse-aware MT. Multimodality could be useful, for example, in aiding the disambiguation of e.g. gender of pronominal anaphora.

Pronominal anaphora Anaphora means the use of a word to refer back to a previous word or constituent, typically realised with the use of personal pronouns and demonstratives. Anaphora is tackled as part of the task of coreference resolution along with its counterpart, cataphora, which conversely involves referring to a future entity. Although anaphoric reference is a frequently-used device in language, particular usages and distributions of pronouns show significant variance across languages (Russo et al., 2012), which poses a challenge in the context of discourse-aware MT (Guillou, 2016). This becomes especially important when the source and target languages handle the gender and number agreement of anaphoric pronouns in different ways, whereby a small mistake made by an MT system might create a situation where the output might seem incohesive, or even offensive, to the reader. For example, Finnish makes no gender distinction in pronouns, using only one pronoun – *hän* – in the 3rd person. A common challenge when translating from Finnish into languages like English and Swedish is disambiguating the correct gender. A further challenge may arise in informal language, where Finnish commonly uses the pronoun *se* 'it' to refer also to humans.

Noun phrase definiteness Definiteness marking is a morphosyntactic phenomenon associated with nominal references in the context of discourse. A language may have no overt definiteness marking (e.g. Russian), or mark degrees of definiteness in many different ways, such as through articles (e.g. English), affixes (e.g. Swedish), verb forms (e.g. Hungarian), or nominal declension (e.g. Turkish). Furthermore, even closely-related pairs of languages can exhibit diversity in their situational use of definiteness, leaving disparate amounts of information to be deduced from the discourse setting. Such cases create referential ambiguity that could leave sentence-level MT guessing, especially when translating from a source language with no definiteness marking to a target language with heavy marking. This issue could only be realistically addressed by discourse-aware MT with access to a larger context. For a human analogue, Knight and Chander (1994) conduct a study, where they show a series of noun phrases to English speakers, asking them to determine whether they refer to definite or

indefinite entities. The results from this experiment show that human accuracy reaches 95% when the annotators are provided the discourse context, in contrast to 80% when they are shown isolated phrases.

Verb tense and aspect Verbs constitute another challenge for translational determinacy in a restricted context. A verb's tense and aspect denote the temporal situation of the action indicated by the verb. Languages use various inventories of tenses and aspects to possibly denote when the action starts and ends, whether or not it is ongoing, how repetitive it is, and so on. For example, Meyer (2011) observe that the English simple past can correspond to the imparfait, passé simple or passé composé owing to the more multifarious aspect marking system of French. Similarly, the Finnish present tense be translated into English as simple present, present progressive, simple future or future progressive. In contrast, Gong et al. (2012a,b) examine ways in which Chinese verbs, which are not marked for tense, could be translated to the correct morphological forms of English verbs. Such cases where a language has less overt marking than the other are analogous to the cases of pronouns and nouns discussed so far, in that the discourse could inform translation on what is situationally appropriate.

Discourse connectives Hardmeier (2014) also touches upon the case of discourse connectives, which are words or phrases occurring in a clause, which clarify how it relates to another clause (e.g. *"but"*, *"so"*, *"because"*, *"moreover"*). It is quite common to see discourse connectives linking together entire sentences (or even spans of sentences), which renders their intepretation a distinctly discourse-level phenomenon. Cartoni et al. (2013) take the Europarl corpus (Koehn, 2005) as a study, and compare parts that were originally written in French, with those translated into French from English, German, Italian, and Spanish. Comparing different subsets of the vocabulary, they find that discourse connectives show a significantly variable distribution depending on the original source language, and also report that the translators have introduced new connectives more liberally than leaving out existing connectives in the source language.

2.2 Discourse and multimodality

As noted above, considerations related to discourse in (human or machine) translation have often focused on mainly linguistic issues. Because the MeMAD project involves specifically audiovisual data, such as television programmes, where different modalities (language/verbal, auditory and visual) combine, discourse-related questions necessarily extend beyond the language/verbal mode. We have previously discussed issues and approaches related to multimodal MT in deliverable D4.1. In this section, we examine how multimodal considerations affect discourse-related questions in general.

In broad terms, discourse is embedded in multimodality. This means that language is used in the presence of other communication modes, such as images and sounds, and these modes are simultaneously affecting our interpretation of the message. In fact, language itself necessarily materialises in some modality, and the modalities have particular modal features: The spoken form is produced in sound and perceived via hearing, and it has a variety of auditory means for meaning-making (e.g. pitch, rhythm, intonation and other prosodic elements of language). The signed form is produced with the body and perceived via sight (or via touching), and it has a multitude of visual and embodied means for meaning-making (e.g. handshape, movement and location). Finally, the written language form, which is the most recent invention of the three modalities, also takes the visual modality to make meaning (e.g. typography and text layout). In addition, spoken language is embedded in audiovisual multimodality because as we speak, we use gaze, gestures, movements of mouth and our bodies to accompany the

words - or sometimes vice versa: what we say accompanies other action. This last condition characterises the concept and perspective of multimodality: language is not necessarily the primary means for meaning-making.

These different modes have distinct representational qualities. Images in audiovisual texts (hence, typically in cinema or videos) represent unique items (*certain* man, woman, house, scenery), whereas words can merely make reference to these on an abstract level. Another difference is that images are organised holistically and provide multiple points of interest, while language is sequentially organised, one element following another, and the sequentiality is pivotal in order to make sense (i.a. word order). Nonetheless, the visual form comes closer to the linguistic one in audiovisual texts that follow the rules of cinematic representation: shots are organised in mutually-dependant sequences (Hirvonen, 2014).

To overcome the dominance of the linguistically-oriented study of translation, multimodality has become a central object of study and an important theoretical framework in Translation Studies and in the field of Audiovisual Translation (AVT) in particular (O'Sullivan, 2013; Kaindl, 2013; Pérez-González, 2014). The criticisms presented concern the past trend in research to place the verbal elements of an audiovisual text above all other modes whereas the focus should often be the contrary: the verbal components are complementary to image, sound, and music.

A basic multimodal inquiry involves Multimodal Discourse Analysis (MDA) and the study of text-image relations (e.g. Unsworth and Cléirigh, 2011). MDA departs from the theory of social systemic functional grammar (Halliday, 1985) and it posits a linguistically-oriented meaning-making system to multimodal entities, particularly emphasising a syntax-level ranking to elements on different levels (e.g. colours serve certain function, so does movement, etc.) (Jewitt, 2011, 34-35). MDA has, however, been criticised for explaining non-linguistic meaning-making with linguistic repertoire, and there are other approaches to understand multimodality in discourse (Ketola, 2018).

In AVT, the relation between subtitles (language) and shots/sequences (image, sound) has produced knowledge of different strategies and tactics for translating. Text and image or text, image and sound can be in synchrony and thus mutually complementary (e.g. Lautenbacher, 2018, on redundancy in multimodal translation) but they can also be dissonant (providing different information at the same time) and even contradictory (e.g. famous scenes from *Clockwork Orange* where pleasant classical music is heard while violent action is depicted). These findings are relevant for discourse-aware MT of audiovisual data in that the information coming from distinct modes is not always mutually supporting. Thus, for instance, image does not always help in translating the subtitle, nor does the subtitle always talk about what is visible in the image.

Finally, relevant understanding of speech and spoken language for the development of discourse-aware MT comes from the discipline of multimodal conversation/interaction analysis. Multimodal CA (see Deppermann, 2013) studies human sociality and the use of language and other modes in social interaction by analysing the construction of interaction (talk, but also common activities) between people moment-by-moment. Here, too, speech is not necessarily the primary or the most central mode of meaning-making, and the research has shown, among other issues, how bodily activities are constructed, how objects are involved in activities, how interactants orient to spatial surroundings, and how human-machine interaction occurs. Hence, in order to understand what is being said, the multimodal context needs to be analysed in detail. Dialogue and action in audiovisual data are only representations of natural interaction but, given the premise that these representations are interpreted by humans who are used to making meaning in social interaction, knowledge of this type of multimodality be-

comes handy. For instance, the knowledge of multimodal classroom interaction (how people use gestures, gaze, visualisations and artefacts in combination with speech to teach and learn) is useful when the multimodality of instructive audiovisual texts (e.g. assemblage videos) is analysed in order to understand the resources for meaning-making and the text-image-sound relations.

In the context of MeMAD, development of MT systems for various use cases involves audiovisual material, such as the translation of video subtitles discussed in Section 5. For such use case, augmenting subtitles with information from the visual and auditory modalities could help improve translation accuracy in general, and handling of discourse-relevant phenomena in particular. As one potential use, visual information could be helpful in resolving ambiguity in the cases of pronominal anaphora by providing information about whether *they* refers to humans or inanimate objects (which take different pronoun in e.g. Finnish). In addition to linguistic and textual content of the subtitles, multimodal discourse-related information such as speaker turns and shot changes could be useful for predicting segmentation and timing of subtitles (see Section 5.2). MeMAD work on multimodal analysis software is reported in more detail in deliverable D2.2.

3 Approaches to discourse-aware machine translation

Over the years, a number of practical approaches have been proposed for tackling discourse-aware MT based on the current state of the art. While said approaches have been predominantly based on the statistical MT paradigm in the recent past (Hardmeier, 2014), the ideas behind them often translate well into the neural MT architectures more common today. As a relatively niche subject within the MT community, published research on discourse-aware MT is clustered in a few specialised venues.

The Conference on Machine Translation (WMT) is the most well-known conference for the MT community, drawing many submissions featuring a variety of research topics. While dating back to the Workshop on Statistical Machine Translation (Koehn and Monz, 2006), it has been held as a large-scale conference housing several shared tasks related to various recurring problems in MT since 2016. Although WMT has attracted some work specialising in discourse-level MT, such as a pronoun test suite by Guillou et al. (2018), there has been no shared task tailored specifically for discourse-aware MT until this year. The news translation task at WMT 2019 included a separate track for document-level translation of multilingual news articles (Barrault et al., 2019). We include our paper describing the MeMAD submission to this track (Talman et al., 2019) in Appendix B.1.

The International Workshop on Spoken Language Translation (IWSLT) is the leading venue for research on spoken language translation systems. The workshop is held annually, and has organised various shared tasks on spoken language translation since 2004 (Akiba et al., 2004). Evaluation campaigns from the last two years of IWSLT (Niehues et al., 2018, 2019) have featured tracks on end-to-end speech translation, with the 2019 shared task also including an experimental track on multimodal spoken language translation in the form of subtitle translation using the full audiovisual modality of videos. These aspects make the research targeted by the workshop relevant for discourse-aware MT. We discuss more details on IWSLT in Sulubacak et al. (2019), an expanded survey based on our previous MeMAD deliverable D4.1.

As a more focused venue, albeit with a smaller scope, the Workshop on Discourse in Machine Translation (DiscoMT) was first held in 2013 and has been repeated biennially since (Webber et al., 2013, 2015, 2017; Popescu-Belis et al., 2019). The workshop has served to incentivise

research on document-level MT systems and theoretical studies on discourse-level phenomena. So far, the workshop has organised shared tasks on pronoun translation (Hardmeier et al., 2015) and cross-lingual pronoun prediction (Loáiciga et al., 2017b). We include our MeMAD submission to 2019 DiscoMT workshop (Scherrer et al., 2019) in Appendix B.3.

3.1 Document-level machine translation methods

The early literature in document-level MT contains some influential studies based on the phrase-based statistical MT architectures common at the time. Following in the footsteps of earlier proposals to integrate translation memories to achieve a consistent MT output (e.g. Veale and Way, 1997), Alexandrescu and Kirchhoff (2009) present one such study, utilising a graph-based semi-supervised learning method that goes beyond the confines of the sentence to enforce a more structured MT output. While the approach comes with relatively high computational costs, they report significant improvements in general translation quality.

Later, Xiao et al. (2011) propose a two-pass decoding approach as a more efficient means to achieve document-level cohesion. Their approach involves cascading sentence-level translation with a second document-wide decoding step that replaces inconsistently-translated words across sentences with the optimal candidates based on their relative frequencies in the document. Ture et al. (2012) build upon their work by factoring in the overall rarity of words in their forced decoding routine, alleviating the bias on word senses that occur frequently in the most common text domains. Their improved method also allows multiple senses of words within a document through a heuristic process that detects when to enforce decoding consistency.

Hardmeier et al. (2012) present a more radical criticism of the locality assumptions built into the state-of-the-art MT architectures. In their proposed solution, they introduce a novel method of decoding entire documents as opposed to sentences, showing that transcending the independence assumption between sentences can be done in an efficient way. In Hardmeier et al. (2013), the authors follow up on their previous study by presenting their implementation of a document-level decoder. Although this innovation facilitates the design of a broad range of document-level MT models by allowing the use of cross-sentence features in decoding, it was overshadowed by the rapidly-rising interest in neural MT systems.

As a more recent study in the realm of neural MT, Tiedemann and Scherrer (2017) experiment with using extra-sentential context windows of various sizes for both the source and target languages. Their method involves concatenating a fixed number of sentences using a special delimiter that indicates a boundary between context and the current sentence. This approach contrasts with previous work that focuses on maintaining lexical consistency (as discussed previously in Section 2.1) by leaving the training procedure free to choose how to utilise the extra information. In their evaluations, the authors observe minor improvements in general translation quality when using extra source context, which we investigate further in our document-level MT experiments (see Section 4.1). In addition, the authors investigate patterns in attention weights that may indicate discourse-awareness. They report some cases where the decoding of a target word causes attention to antecedent words in the source context which are coreferent with it, which might be a learned behaviour that conditions the decoding step to prioritise coherence over word-for-word accuracy.

More recently, a number of studies have focused on directly conditioning the attention mechanism to leverage the context outside the sentence. Miculicich et al. (2018) introduce an hierarchical attention structure implemented as an extra layer of abstraction, modelling extra-sentential context in a consistent manner. This method comprises an extension to the encoder-

decoder structure of neural MT that allows arbitrarily-sized context windows to factor into attention, allowing flexibility in designing document-level MT models. Maruf et al. (2019) consequently apply this approach to build an hierarchical attention model on entire documents, rather than considering a fixed-size context window from neighbouring sentences. Both studies report significantly increased translation quality according to reference-based evaluation metrics, though we were not able to replicate the results in our own experiments (see Section 4.1). Jiang et al. (2019) very recently propose using associated memory networks to extract attention distributions between sentences, which allows them to automatically classify the relevance of sentences to each other as part of their contexts. The authors rigorously compare their results with various document-level neural MT approaches, demonstrating an overall superior performance.

Finally, it is also worth noting that the new frontier in document-level neural MT did not preclude further research in maintaining consistency in the output of MT. A recent notable example is the study by Voita et al. (2019), which proposes an offline automatic post-editing layer specifically tailored to repair inconsistencies in the output of sentence-level MT models. While the authors do not include performance comparisons with end-to-end document-level MT systems, they do report improvements based on contrastive test sets revolving around particular linguistic challenges, and also report positive results based on human evaluation.

3.2 Evaluation of discourse-aware machine translation

Evaluating discourse-related features and phenomena in MT forms a challenge of its own. Commonly used automatic MT evaluation measures such as BLEU (Papineni et al., 2002), which calculate similarity scores based on the overlap of words or n-grams (usually up to 4-grams) between an MT hypothesis and one or more reference translations, do not reliably capture textual features like discourse extending beyond the n-grams, much less beyond sentences. As long-range dependencies are not captured by these metrics, they cannot reflect discourse-level phenomena. It has therefore been argued by various authors that different approaches are needed to assess discourse phenomena and inform the on-going work on discourse-aware MT development (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Meyer et al., 2012). Recent work has indeed been aimed at developing metrics and test sets focusing on various discourse-related issues, mainly pronouns and cohesive devices.

Pronouns and pronominal anaphora has been one of the most studied aspects of discourse phenomena in MT research, and various authors have recently proposed metrics or implementations for evaluating pronoun translation. Hardmeier and Federico (2010) use precision and recall of pronoun translations to assess the correctness of pronominal anaphora in English-German MT. Comelles et al. (2010) present automatic metrics for assessing anaphoric relations between a subject noun and personal pronoun, possessive adjective and subject noun or personal pronoun, demonstrative pronoun and its referent, as well as main verb and auxiliary by comparing relevant translations in the MT and human reference.

Cohesive devices have been another discourse phenomenon targeted in evaluation. Wong and Kit (2012) focus on lexical cohesion and propose automatic metrics for assessing lexical cohesive devices by identifying repetitions of the same content words and stems as well as synonyms, near synonyms and superordinate terms. Their metrics use stemming and WordNet (Miller, 1995), and can also be combined with n-gram similarity metrics. Lexical consistency or cohesion is also employed by Hardmeier et al. (2013) in the context of evaluating the readability of English-Swedish MT. For assessing lexical consistency, they implement

type-token ratio as well as specific readability metrics. In addition to lexical cohesion, discourse markers such as connectives have been included in evaluation metrics, for example by Comelles et al. (2010). A set of automatic and semi-automatic evaluation metrics called ACT is presented by Meyer et al. (2012) to evaluate the accuracy of discourse connective translation, and further implemented for English–French and English–Arabic (Meyer et al., 2012; Hajlaoui and Popescu-Belis, 2012, 2013). The performance of their metrics is also assessed and validated against human evaluations.

Based on Rhetorical Structure Theory, Guzmán et al. (2014) extend existing lexical similarity metrics using discourse parsing and comparing discourse parse tree kernels between MT and human reference sentences. They tested the approach on four language pairs with English as target language, and suggest that integrating discourse information into lexical similarity metrics can improve correlation with manual evaluation. The work on discourse tree representations is further implemented in Joty et al. (2014), where the authors test and evaluate a combination of discourse-based metrics for various language pairs with English as the target.

In addition to metrics, some specific test suites focusing on discourse-related phenomena have been proposed. Guillou and Hardmeier (2016) present a test suite for evaluating pronoun translations containing 250 hand-selected pronoun tokens in English and an automatic evaluation method which compares the translations of pronouns in MT output with those in the reference translation. An English-German adaptation of that test suite was released later as part of the WMT 2018 shared task (Guillou et al., 2018), comparing 16 NMT systems with both general (BLEU) and targeted pronoun evaluation (APT), and demonstrating that (i) translation quality varies greatly between MT systems, and (ii) pronoun translation is a challenge even for the best MT models. Two hand-crafted test sets are presented by Bawden et al. (2018), one tailored to evaluate anaphoric pronoun translation, and the other coherence and cohesion. The test sets proposed consist of ambiguous English source sentences translated into contrasting French target sentences with varying degrees of correctness. A recent paper by Bawden et al. (2019) presents a bilingual test set consisting of 144 spontaneous informal dialogues (5,700+ sentences) between native English and French speakers mediated by MT. The test set also contains human evaluations of the MT outputs of each speaker's turns, rated by the other speaker for various aspects of quality, including coherence (Bawden et al., 2019).

The metrics discussed here provide some interesting tools for examining specific discourse-relevant issues. For the purposes of our experiment involving machine-translated subtitles (see Section 5), we have at this stage chosen to take a more qualitative, manual analysis of discourse features to identify issues arising in the audiovisual data addressed by the MeMAD project and in the specific language pairs in question. Possible selection and use of suitable metrics will be explored at a later stage of the project informed by the issues identified.

4 MeMAD work on discourse-aware machine translation

4.1 Document-level machine translation models

The international conference on machine translation (WMT) featured a document-level news translation task this year[1] covering English-to-German, English-to-Czech and Chinese-to-English tasks. The University of Helsinki participated in the English-German task as one of the relevant language pairs in the MeMAD project. The conference organisers provided training data with document boundaries and the test data are sampled from news snippets with

[1] http://www.statmt.org/wmt19/translation-task.html

short coherent blocks of text. Document boundaries are provided for Europarl, News Commentary and the Rapid corpus in the approved training data. Note that these documents refer to units that are quite different from the target test data, for example, in Europarl they refer to speech boundaries in the proceedings of the European Parliament.

For our submission to the WMT task, we experimented with concatenation approaches (Tiedemann and Scherrer, 2017) and hierarchical attention models (Miculicich et al., 2018; Maruf et al., 2019). The latter were basically unsuccessful and we did not submit their outputs to the official competition. More details are given in Appendix B.1. For the concatenation models, we experimented with extended source context models and models that extend both source and target language context. Furthermore, we added a model that includes target language context as input to the encoder instead of decoding larger chunks and cutting the output later on. All those models use sliding window approaches with at most three sentences in one chunk. Note alse that the models with target context sentences in the input are only run as oracle experiments with the correct contextual target sentence taken from the reference data. The results are summarised in Table 1.

	BLEU on news2018	
System	Shuffled	Coherent
Baseline	38.96	38.96
prev+cur → cur	36.62	37.17
2 prev+cur → cur	33.90	34.30
prev+cur+next → cur	34.14	34.39
prev-target+cur → cur	36.82	37.24
prev+cur → prev+cur	38.53	**39.08**

Table 1: Comparison of concatenation approaches for English–German document-level translation on the WMT news2018 test set; different combinations of previous (prev), current (cur) and subsequent (next) sentences.

In order to study the effect of contextualisation, we created a shuffled test set that artificially destroys the coherence by moving sentences out of place. The table contrasts the results on both versions of the test data and we can see that all document-level models have a positive effect on the outcome of the translation process even though the differences are rather modest. Also note that only the 2+2 setup beats the single sentence baseline, which means that the other context-extension models have problems to learn a proper treatment of the concatenated units.

To further study these concatenation models, we also carried out a study on two different domains with a careful analyses of the impact of document-level models on machine translation Scherrer et al. (2019). In this study, we also included another variant that has been proposed at WMT by Junczys-Dowmunt (2019), in which the data is chunked into blocks of roughly the same size with additional labels for marking document and sentence boundaries as well as document continuations. We applied all models in a news translation task and in a subtitle translation task; the latter with particular interest in relation to the MeMAD project. Subtitles are in general an interesting domain as they include substantial contextual relations due to their connections to intensive interactions of speakers and the pragmatic nature of conversational language. Sentences and sentence fragments are also much shorter, which makes it easier to capture context outside of sentence boundaries even with rather limited window sizes that can be managed properly by modern NMT architectures.

We added an analyses of some discourse-level properties that can be identified in the original human-translated data and we found a striking difference between the two scenarios

especially in terms of pronoun usage and referential expressions. Details about the analyses are illustrated in Figure 1. It shows that there is a significant difference in the two domains we consider, for example, in terms of pronouns and coreferential chains. These characteristics are likely to influence the results that can be obtained by document-level MT. More detailed discussions are included in the paper attached as Appendix B.3.

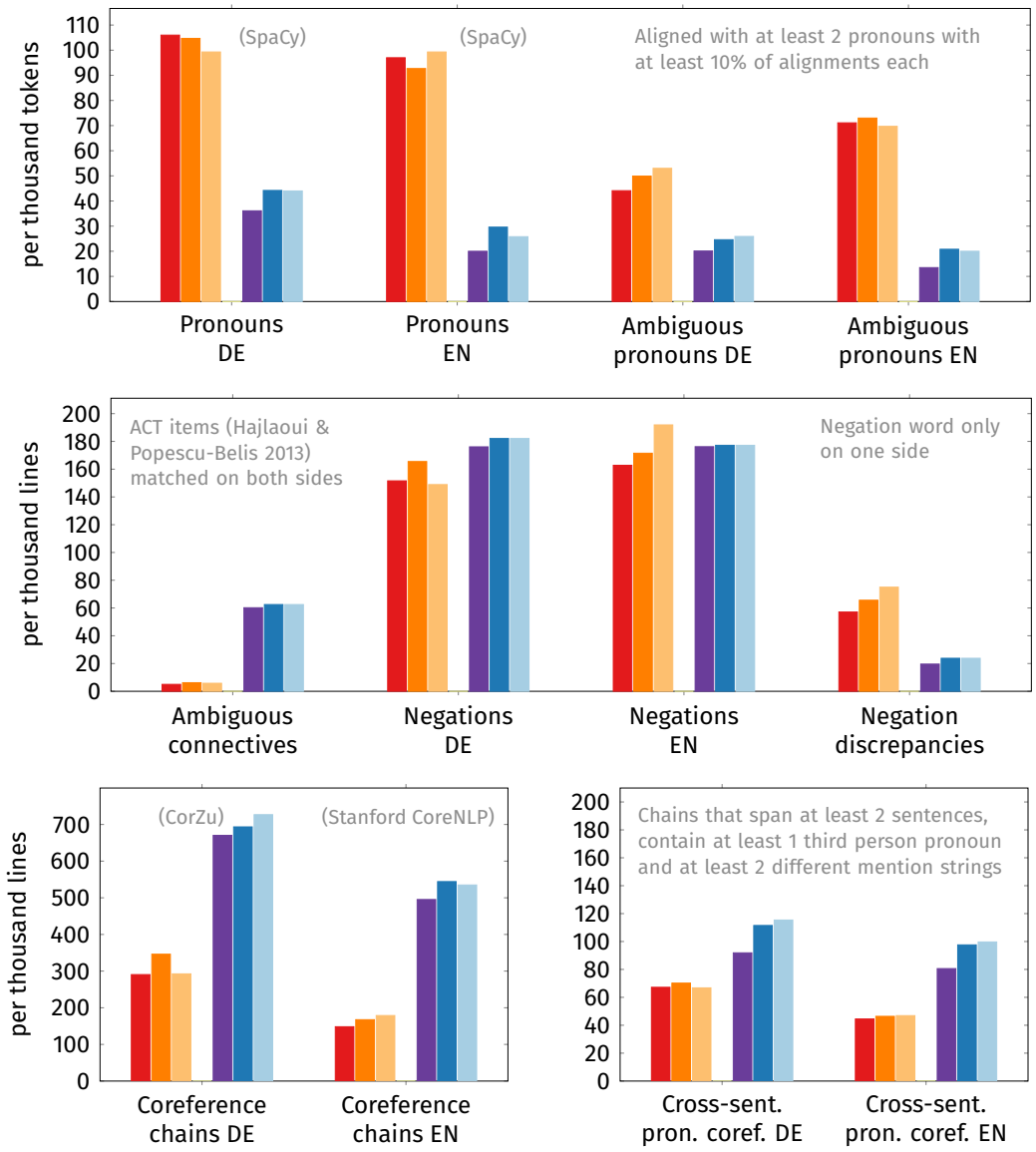

Figure 1: Discourse-level properties identified in two different translation scenarios: The subtitle translation task (in orange-red) and WMT news task (blue-purple). The colour intensity refers to the subset that is considered with training data, validation and test date from left to right.

In the translation experiments, we again compared shuffled (inconsistent context), coherent and also zero-context models (in terms of no surrounding sentences) and we observed a significant improvement with the fixed-size context model on subtitle data. This is true for both translation directions, English-to-German and German-to-English. The positive effect in terms of automatic evaluation can also be confirmed by our YLE test experiments described below. On news data, the effect is not clearly visible confirming our findings from WMT. Figure 2 visualises the results for the English-to-German translation task for the different settings that we tried in that paper.

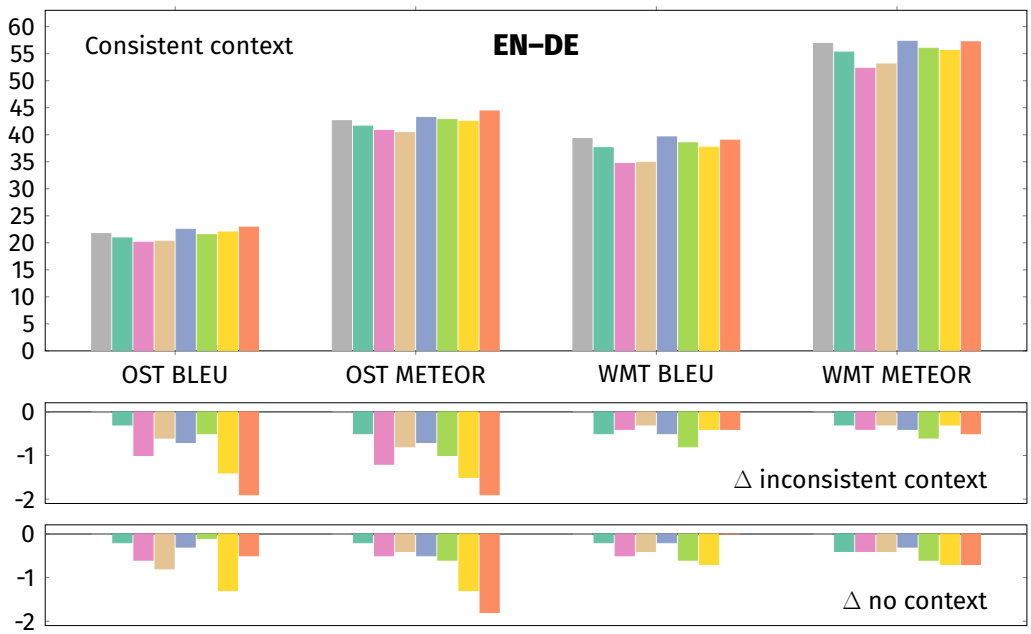

Figure 2: Translation results in terms of BLEU and METEOR scores for different concatenation models in English-to-German subtitle translation (OST) and news translation (WMT) tasks. From left to right: baseline, previous+current source sentences, previous+current+next source sentences, 2-previous+current source sentences, previous+current source and target sentences, previous+current+next source to previous+current target sentences, previous (oracle) target+current source sentences, fixed-size chunk with a maximum of 100 tokens. The delta-plots below the main graph show the difference to the scores obtained with shuffled data (inconsistent context) and test results with zero context.

More details about the experiments and results can be found in the paper attached in Appendix B.3. Overall, the differences between all models are pretty small but there are encouraging results that motivate further studies in this direction especially in the case of subtitle translation. For example, a more sophisticated chunking into coherent blocks would be an interesting extension of the concatenation approach. Adding more explicit information about dialogue structure and speaker identity together with the contextual information would be another point to explore.

4.2 Speech translation models

Audio as a source of non-textual context has been mention earlier in section 2.2. We continue our efforts in developing tools for multimodal machine translation currently focusing on speech translation. With the aim of producing a tool capable of performing end-to-end speech-to-text translation, we have extended the attention bridge model (Vázquez et al., 2019), an inner-attention-based multilingual translation model. The model was originally proposed for learning fixed-size sentence representations via multilingual training of language-specific encoders and decoders that share the parameters of the attention bridge (see Figure 3). However its modular architecture can further be extended to even include audio encoders to be part of the multitask environment it implements. Processing audio features certainly requires a different architecture but the attention-bridge model allows to create independent encoder modules as long as they can be attached to the shared intermediate attention layer. In our initial experiments we integrated a modern audio encoder in our setup training an end-to-end speech translation task using publicly available data sets.

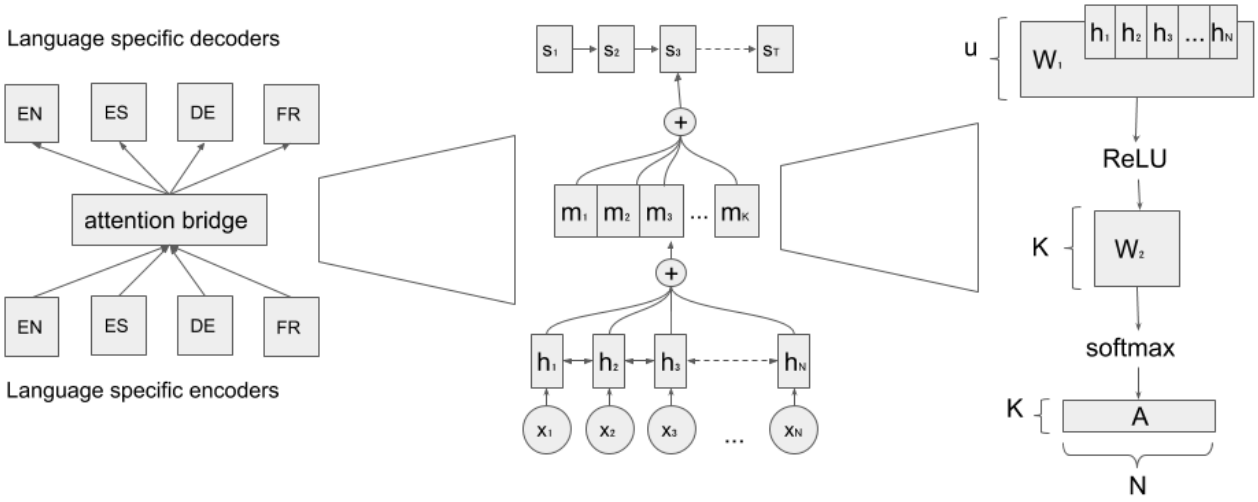

Figure 3: The attention bridge model illustrated in terms of a multilingual NMT model with language-specific encoders and decoders coupled together with a shared intermediate layer. The figure zooms into the details of the architecture from left to right. *Left:* the attention bridge connects the language-specific encoders and decoders. *Center:* input $x_1 x_n$ is translated into the decoder states $s_1 s_t$ via the encoder states $h_1 h_n$ and the attention bridge $M = m_1 m_k$. *Right:* Computation of the fixed-size attentive matrix A.

Those initial experiments suggest that including multimodal information during training is indeed beneficial. This allows us to hypothesise that training modality- and language-specific encoders and decoders allow for specialised handling of the input and output information, respectively, at the time that the shared layer provides a natural alignment of the latent representations and enriches the sentence/utterance-representations. This results in rather promising end-to-end speech-to-text translation capabilities.

For the experiments we have used the English to German part of the MuST-C dataset (Di Gangi et al., 2019) which contains around 400 hours of audio recordings of English TED talks and 229,703 sentences of the English transcripts and the corresponding German translations. As noted by the participants of the IWSLT 2019 speech-to-text evaluation campaign (Niehues et al., 2019), this is far from being sufficient information to train a successful speech-to-text, so our preliminary experiments are to be treated with a grain of salt.

As another stepping stone towards building reliable speech translation systems, Aalto University is developing end-to-end speech recognition systems. Particularly, Aalto has focused on speaker-aware training, motivated by parallel work in speaker verification and diarisation. Both of these developments are further elaborated in MeMAD deliverable D2.2. The next step in our collaboration is to properly integrate improved speech encoding facilities derived from the ASR system into the multitask environment of our attention-bridge model. This will be the focus of the coming period in work package four, together with the contextualised translation models that we discuss in further detail below.

5 MeMAD Use Case 4: User evaluation of discourse-aware MT for subtitling

One of the use cases (Use Case 4.3.4) defined for the MeMAD tools involves manual correction of auto-translated subtitles or post-editing of MT by subtitlers. The rationale for using MT and post-editing is that it has been found to increase productivity in various translation

scenarios (e.g. Plitt and Masselot, 2010; Guerberof Arenas, 2014). However, apart from some prior research projects exploring the use of MT for subtitling, post-editing appears to be less commonly used in the field of audiovisual translation (see Díaz-Cintas and Massidda, 2019). To examine how the use of MT and post-editing in the subtitling workflow affects the work and productivity of subtitlers, we carried out a process study pilot in November 2019. In this pilot study, 12 subtitlers worked on a series of tasks where they created interlingual (translated) subtitles for short video clips both without MT output and by correcting machine-translated subtitles. To assess productivity and effort, keylogging data were recorded during these tasks. Task time and technical effort represented by keystrokes were compared between post-editing and translation from scratch as well as post-editing of two different MT model outputs.

5.1 MT models used for evaluation

For the assessment of MT in subtitle translation we created sentence-level and document-level translation models from all the parallel data available in OPUS.[2] For Finnish-Swedish, this includes a bit over 30 million training examples and for Finnish-English it amounts to roughly 44 million translation units. The training data comes from a very diverse background with sources ranging from Bible translations to software localisation data, official EU publications and data mined from unrestricted web crawls. Altogether, the collection includes 19 different sub-corpora for each, Finnish-Swedish and Finnish-English. The quality and size varies a lot and we did not make any attempt to further clean or filter the data. However, for both language pairs the biggest sub-corpus refers to the collection of movie and TV show subtitles derived from the OpenSubtitles (v2018) data set. For Finnish-English, this collection contains almost 30 million translation units and for Finnish-Swedish it contains over 15 million translation units. Even though this sub-corpus is quite noisy as well it fits the task rather well and, therefore, we can expect that our models should have a decent performance in the subtitle translation task even without further fine-tuning.

The models we trained rely on the popular Transformer architecture, the current state-of-the-art in neural machine translation. We apply the implementation from the open source toolkit MarianNMT (Junczys-Dowmunt et al., 2018), which offers fast training and decoding with the latest features of production-ready NMT. We use the standard setting of a multi-layer transformer with 6 layers on both encoder and decoder side with 8 attention heads in each layer. We enable label smoothing, dropout and use tied embeddings with a shared vocabulary. For text segmentation we apply SentencePiece (Kudo and Richardson, 2018) with models that are trained independently for source and target language for a vocabulary size of 32,000 in each language. There are no further pre-processing steps to keep the setup as general as possible apart from some basic normalisation of unicode punctuation characters and parallel corpus filtering using standard scripts from the Moses SMT package (Koehn et al., 2007).

For the document-level models we apply the concatenative models proposed by Tiedemann and Scherrer (2017); Junczys-Dowmunt (2019) using units of maximum lengths of 100 tokens. Note that sentences and sentence fragments in subtitles are typically very short and 100 tokens typically cover substantial amounts of context beyond sentence boundaries. We mark sentence boundaries with special tokens as described earlier in Section 4.1 and chunk the training and test data sequentially from the beginning to the end without any overlaps. This procedure creates roughly 3.3 million pseudo documents for Finnish-Swedish and 4.7 million documents for Finnish-English. This means that we have on average about 9 sentences per pseudo document, which are concatenated into one long string with sentence boundary

[2] http://opus.nlpl.eu

markers in between.

During test time, we proceed in the same way creating pseudo documents from the original input by concatenating subsequent sentences and splitting when a segment exceeds 100 tokens. Sentence-level models are translated in the usual way. In order to examine the translation quality, we applied our models to the YLE benchmark test set. This test set is taken from a larger set of subtitles from YLE programmes with audio in Finnish, Swedish or English. Intralingual subtitles in the language original audio were aligned with interlingual subtitles of the same programme in one of the other two languages. However, it should be noted that the interlingual subtitles are not direct translations of the intralingual subtitles, as such. Alignment of the subtitle segments in the test set was manually checked and non-corresponding segments were removed. The Finnish and Swedish parts of the dataset also contain intralingual subtitles for the deaf or hard-of-hearing, which were separated in the test set as their own subsets.

The results are shown in Table 2. Scores are shown separately for different subsets. Subsets containing subtitles for the deaf or hard-of-hearing are labelled as FIH (Finnish) and SWH (Swedish). Note, that the document-level results needs to be treated in a special way as they do not automatically match the sentence-level reference translations even when splitting on generated sentence boundary markers. To ensure that reference and system output correspond to each other, we apply a standard sentence alignment algorithm implemented in the open source package hunalign (Varga et al., 2005). We use the re-alignment flag to enable lexical matching as well, which is very beneficial in this monolingual alignment task. BLEU score might be negatively affected by this procedure as this alignment is certainly not perfect.

benchmark	**sentence-level**		**document-level**	
	BLEU	**chrF$_2$**	**BLEU**	**chrF$_2$**
FIH → SWE	17.0	0.435	17.4	0.444
FIN → SWE	21.6	0.466	22.2	0.470
FIN → SWH	17.8	0.429	18.4	0.439
fi → sv (avg.)	**18.8**	**0.443**	**19.3**	**0.451**
SWE → FIH	11.3	0.418	12.9	0.433
SWE → FIN	20.5	0.489	21.6	0.501
SWH → FIN	15.2	0.441	15.9	0.451
sv → fi (avg.)	**15.7**	**0.449**	**16.8**	**0.462**
FIH → ENG	18.6	0.439	21.5	0.459
FIN → ENG	24.4	0.476	25.6	0.484
fi → en (avg.)	**21.5**	**0.458**	**23.6**	**0.472**
ENG → FIH	11.7	0.404	12.8	0.416
ENG → FIN	20.3	0.483	21.3	0.492
en → fi (avg.)	**16.0**	**0.444**	**17.1**	**0.454**

Table 2: Comparison of BLEU and chrF$_2$ scores on the YLE benchmark testset for the sentence-level and document-level systems in the language pairs Finnish-Swedish, Swedish-Finnish, Finnish-English and English-Finnish

The results indicate that document-level models seem to be beneficial in the subtitle translation case. The automatic evaluation scores consistently show an improvement over the corresponding sentence-level models for both language pairs and in all directions. However, this encouraging result does, unfortunately, not carry over to the manual assessment (see Sec-

Original subtitles in SRT format:

```
16
00:01:05,960 --> 00:01:12,360
We have to make readmission
agreements with other countries, -

17
00:01:12,440 --> 00:01:17,440
so that they would be willing.
We have to cooperate closely.
```

Converted to sentence-level segments for machine translation:

```
  <s id="13">
    <time id="T16S" value="00:01:05,960" />
We have to make readmission agreements with other countries, -
    <time id="T16E" value="00:01:12,360" />
    <time id="T17S" value="00:01:12,440" />
so that they would be willing.
  </s>
  <s id="14">
We have to cooperate closely.
    <time id="T17E" value="00:01:17,440" />
  </s>
```

Mapped back to subtitle frames after translation:

```
16
00:01:05,960 --> 00:01:12,360
Meidän on tehtävä
takaisinottosopimuksia muiden maiden kanssa,

17
00:01:12,440 --> 00:01:17,440
jotta ne olisivat halukkaita.
Meidän on tehtävä tiivistä yhteistyötä.
```

Figure 4: Pre- and post-processing of subtitle data before and after translation. Sentences may run over several subtitle frames like in the second sentence in this example. Multiple sentences and sentence fragments can also appear in one frame as it happens in the first one in the example above. The translation comes from a document-level model.

tion 5.4). A reason for this may at least be partially related to the problem of segmentation and time-frame alignment, which we introduce below.

5.2 Subtitle frame alignment

In both cases, sentence-level and document-level translation, we have to treat the results in a way that maps the translations back into the time slots allocated for the original subtitles. Remember that those time slots may include more than one sentence and sentences may stretch over multiple time slots as well. Because our translation models are trained on sentence-aligned data, we need to extract sentences first from subtitles, too. We do this using the techniques proposed by Tiedemann (2008), which are also applied to the OpenSubtitles corpus in our training data. An example of the pre-processing procedure is shown in Figure 4.

Mapping back to subtitle frames and their time allocations is implemented as another alignment algorithm. We apply a simple length-based algorithm for this assuming that there is a strong length correlation between source and target language subtitles. The difference to traditional sentence alignment is that we are now only interested in 1-to-x alignments, meaning that each existing subtitle frame in the original input should be filled with one or more segments from the translation. The segments on the target side that we consider are clauses from the generated sentences. For simplicity, we split on any punctuation in the output that is followed by space to approximate the structural segmentation. We then apply the traditional Gale&Church algorithm (Gale and Church, 1993) to optimise the global alignment between source segments (original subtitle frame data) and target segments. For this, we adjust the parameters of the algorithm in two ways: (i) we remove priors and apply a uniform distribution over possible alignment types and (ii) we change the set of alignment types to include all possible mappings from one source segment to a maximum of four target segments. The mapping between source and target is than created using the original dynamic programming algorithm that ensures a globally optimal mapping according to the model. The result of this procedure is shown in figure 4 on the example from the pre-processing steps discussed earlier.

The algorithm is then also extended by further segmentation and line-break insertion procedures. Subtitles need to follow certain length and formatting constraints. In our current implementation, we added a simplistic strategy to insert line-breaks. We introduce a hard length limit and a soft length limit that regulate the segmentation of text into lines. Furthermore, we set a hard coded length limit penalty that is added to the cost function of the alignment algorithm. This penalty is added for alignment candidates that exceed the hard limit. Additionally, we add another penalty for splitting subtitle frames in the middle of a sentence at some potential clause boundary. This is used to prefer full sentences within one frame. Finally, the soft length limit is used to split sentences at arbitrary white space characters in cases where a text to be added exceeds the hard length limit. The algorithm tries to find a splitting point around the soft limit that balances the length on both sides of that newly introduced splitting point.

The implementation of the entire procedure described above is available from github as part of the subalign package.[3] The heuristics applied in this algorithm still require some development and optimisation as we will see in the assessment of the subtitle translation experiments described below. Especially, the segmentation and line-break insertion procedures do not seem to work very well and may have influenced the results in a rather negative way, especially for the document-level models.

5.3 User data collection set-up

The subtitling tasks for productivity data collection were carried out at YLE premises in November 2019. A total of 12 translators (3 for each language pair) participated in the tasks. In addition to 9 in-house translators working for the project partner YLE, 3 freelancers were recruited as participants. All participants are professional translators with between 4 and 30 years of professional subtitling experience in the language pair in question. Of the 12 participants, only 2 indicated they had previously used MT specifically for subtitling, although 7 had used MT for other purposes.

The subtitling tasks were carried out using the subtitlers' preferred software environment (Wincaps Q4 or Spot). To replicate their normal working environment, an external monitor and keyboard were provided, and the subtitlers had access to the internet as well as termi-

[3] https://github.com/Helsinki-NLP/subalign

nology and other resources normally used in their work. During the subtitling tasks, process data were logged using Inputlog (Leijten and Van Waes, 2013), which records all keyboard and mouse activity. Screen recording software provided as part of Windows 10 was used to capture video of the process to support further analysis. Pre- and post-task questionnaires were used to collect background information from the participants and subjective assessment of the MT output and PE experience after the tasks. After the tasks, a brief semi-structured interview was also carried out to collect more detailed feedback regarding problems in the workflow and the participants' views on potential improvements.

Based on availability of material and participants, subtitling tasks were carried out in four of the MeMAD language pairs: Finnish-English, Finnish-Swedish, English-Finnish and Swedish-Finnish. For each source language (Finnish, English, Swedish), six clips were selected from two different genres: three clips from unscripted European election debates (Finnish and Swedish from 2014, English from 2019), and three clips from semi-scripted lifestyle or cultural programmes. The individual clips were selected so each clip (i) forms a coherent, self-contained section of the programme, (ii) is approximately 3 minutes long, and (iii) contains approximately 30–35 subtitle segments. The total length and number of clips used was limited due to limited availability of the participants.

Each participant completed a total of six tasks where they subtitled a short video clip: two clips "from scratch" without MT output, two clips using output from a sentence-level MT system, and two clips using output from a document-level MT system. The clips and MT outputs were rotated in a round-robin format so that each clip was subtitled once in each condition (no MT output, sentence-level MT output, document-level MT output) by a different participant. Task order was also varied to minimise facilitation effect. The participants were instructed to produce subtitles that would be acceptable for broadcasting, and to use the resources (e.g. the internet, terminology resources) they normally would for their work, but to not spend excessive time in "polishing" any given wording or researching information. No explicit time limit was given for each task, rather, the participants were instructed to work at their own pace.

In the from scratch condition, the participants also created the segmentation and timing of the subtitles following their normal work process (subtitling templates, which are commonly used in e.g. DVD translation are not used by YLE for these content types). In the post-editing condition, the participants worked with output that was pre-segmented and timed based on the intralingual subtitles used as source text for the MT (see Section 5.2.

5.4 Analysis of productivity and changes in post-editing

To assess productivity, the process logs recorded during the subtitling tasks were analysed using Inputlog's analysis functions. Task time and number of keystrokes logged were used as productivity measures. Based on the final subtitles produced, edit distance between the MT output and the final versions were calculated using HTER (Snover et al., 2006) and characTER (Wang et al., 2016). These measures were then compared between the tasks of creating interlingual subtitles from scratch and post-editing MT, as well as between post-editing the sentence-level and document-level MT outputs described in Section 5.1.

Figure 5 shows the average total task time and average subtitling task time. The total task time reflects the total time taken by the participant to complete an individual task. In addition to creating the subtitles themselves, activity during the task also includes, for example, searching for terminology or other information online. Inputlog also measures time spent working with specific software during the logging, and this function was used to focus on the

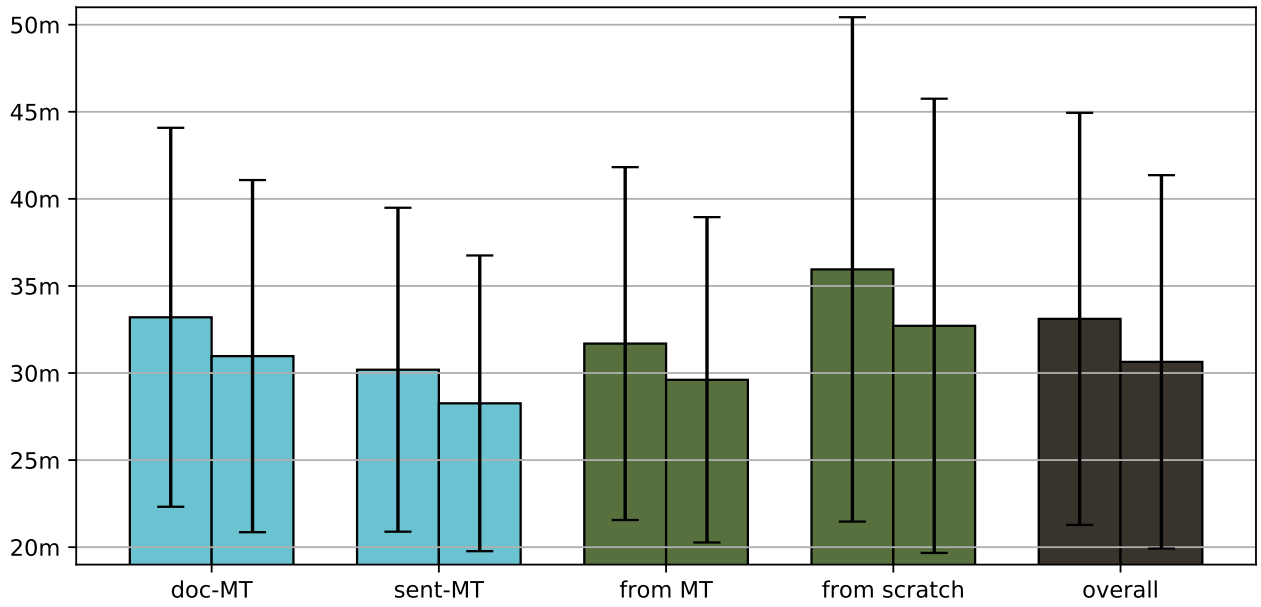

Figure 5: Average total task times (left) and task times subtitling (right).

time the participant used the subtitling software, specifically, excluding internet searches and other activity. The comparison of average task times indicates that, on average, 1) post-editing machine-translated subtitles (regardless of MT output) was faster than creating subtitles from scratch, and 2) post-editing output by the sentence-level MT system was slightly faster than post-editing the document-level MT output.

Figure 6 shows a comparison of technical effort in terms of the average number of keystrokes used when producing subtitles. For the comparison of keystrokes, we focused only on the keystrokes created in the subtitling software, excluding online searches and other activity. In addition to the total number of keystrokes, the distribution of types of keystrokes is shown in Figure 6. Text production includes alphanumeric keys related to producing characters, and text deletion includes keystrokes related to removing characters. Keystrokes related to editing the subtitle frames (creating, deleting, splitting or merging frames etc.), as well as changing the timing of subtitle frames, are shown separately. Keystrokes related to other activity such as navigation and video controls are grouped as miscellaneous keystrokes. Similarly to task times, the comparison of keystrokes indicates that, on average, 1) post-editing machine-translated subtitles (regardless of MT output) involved fewer keystrokes than creating subtitles from scratch, and 2) post-editing the sentence-level MT involved fewer keystrokes than post-editing the document-level MT output.

Some differences can also be seen in distribution of keystroke types. Intuitively, post-editing reduces the need for text producing keystrokes compared to writing the subtitles from scratch. However, the share of text deleting keystrokes is somewhat higher, as correcting the output also involves removing words or characters. The number of keystrokes related to editing subtitle frames can also be expected to be higher in the from scratch mode, as the participants needed to create and set the timing for each subtitle segment themselves. Although in the post-editing condition, the MT output was already segmented and timed based on the intralingual subtitles used as source text, the number of keystrokes in this category show that the participants found it necessary to change both the segmentation and timing. These edits are discussed in more detail below.

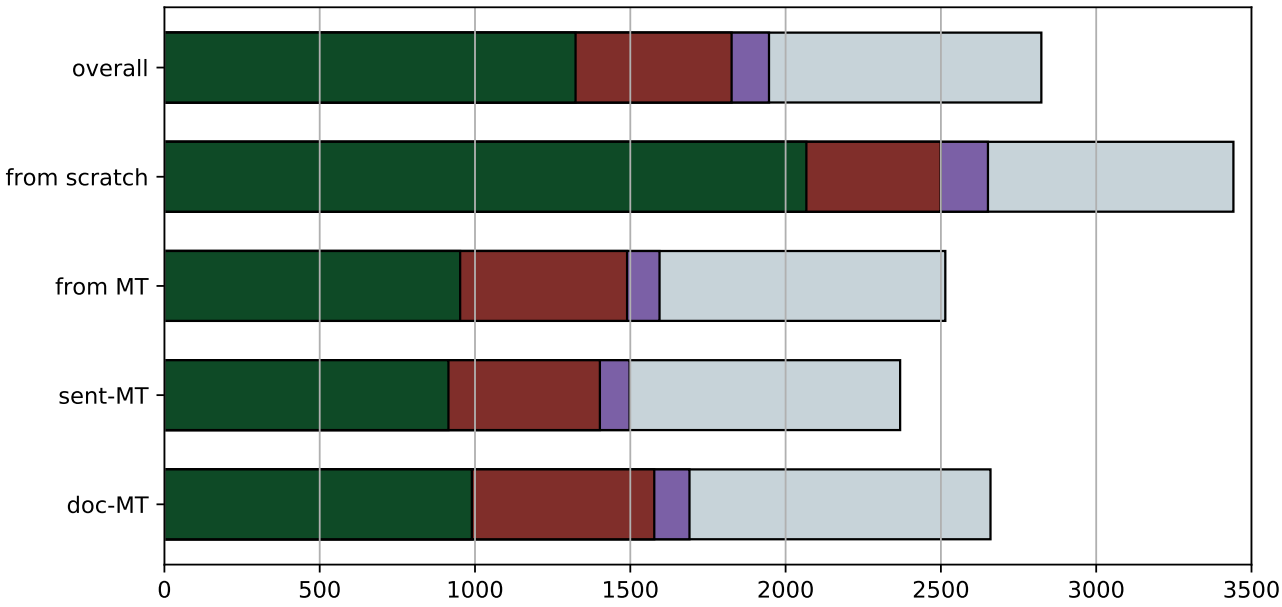

Figure 6: Average numbers of keystrokes, divided into keystrokes for text production (leftmost), text deletion (mid-left), timestamp and segmentation edits (mid-right), and miscellaneous actions (rightmost).

Although overall, post-editing machine-translated subtitles can be seen to lead to productivity gains in terms of shorter task times and reduced keystrokes, considerable variation was observed between different participants. Figure 7 shows the average subtitling task times for each participant across all the subtitling tasks while post-editing either MT output (left) and while creating subtitles from scratch (right). Some participants (fi-en A, en-fi A, en-fi B, sv-fi B, sv-fi C) show marked gains in productivity, particularly those with the overall longest average task times. However, 5 out of the 12 participants (fi-en B, fi-en C, en-fi C, fi-sv A, sv-fi A) are in fact slower when post-editing MT.

A further comparison can be made with the average number of keystrokes. Figure 8 shows the average number of subtitling keystrokes for post-editing either MT output (left) and for translating from scratch (right). Although most participants use, on average, fewer keystrokes when post-editing, fi-sv A and sv-fi A actually use slightly more keystrokes. While these two are among the participants who are also slower when working with MT output, it should be noted that for fi-en B, fi-en C and en-fi C, a lower number of keystrokes does not correspond to shorter task time in post-editing.

These findings are in line with other process studies (see e.g. Plitt and Masselot, 2010; Koponen et al., 2012) which have demonstrated that potential productivity gains from post-editing vary, with slower participants often showing the largest gains and translators who are already fast benefiting less. A lower number of keystrokes not necessarily leading to increased productivity in terms of time has also been observed in other studies. While the number of keystrokes captures the amount of technical effort needed to make corrections to the MT output, it does not capture the amount of cognitive effort involved in recognising potential errors and deciding on necessary changes. Another factor is post-editing experience. The participants in this study indicated they had little prior experience with MT particularly in the subtitling context, so unfamiliarity with the task may play a role.

In addition to the process-based productivity measures, the edit distance metrics HTER (word-based) and characTER (character-based) were used to compare the number of changes

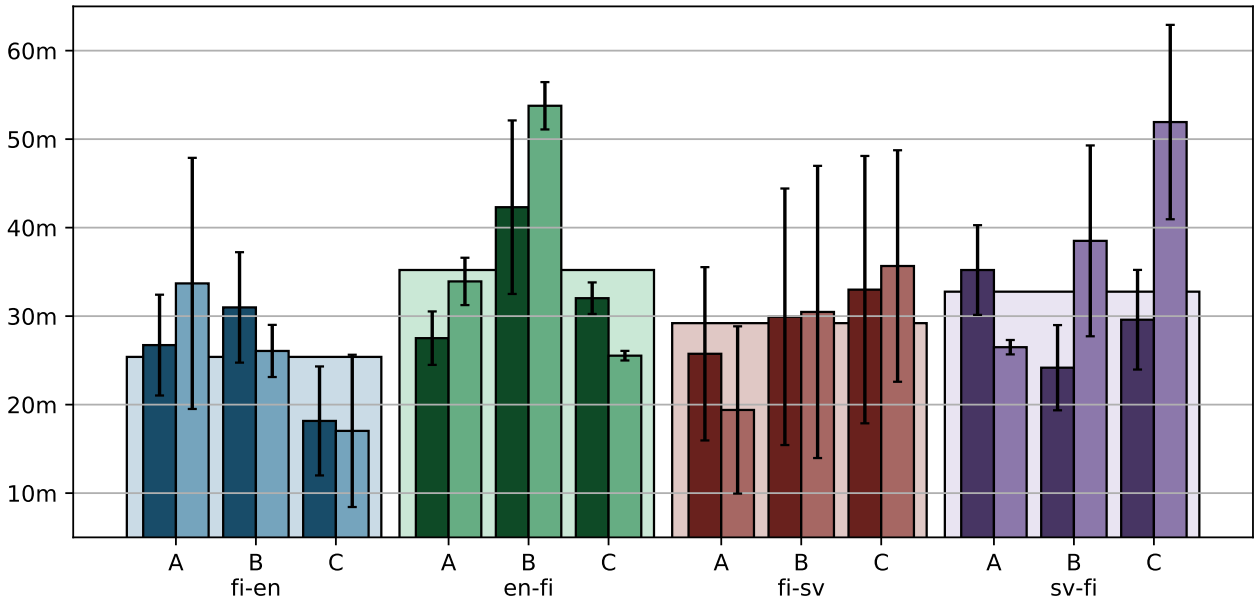

Figure 7: Average task times subtitling through post-editing (left) and subtitling from scratch (right) for each participant (labelled as A, B, C for each language pair), compared to language pair averages (behind).

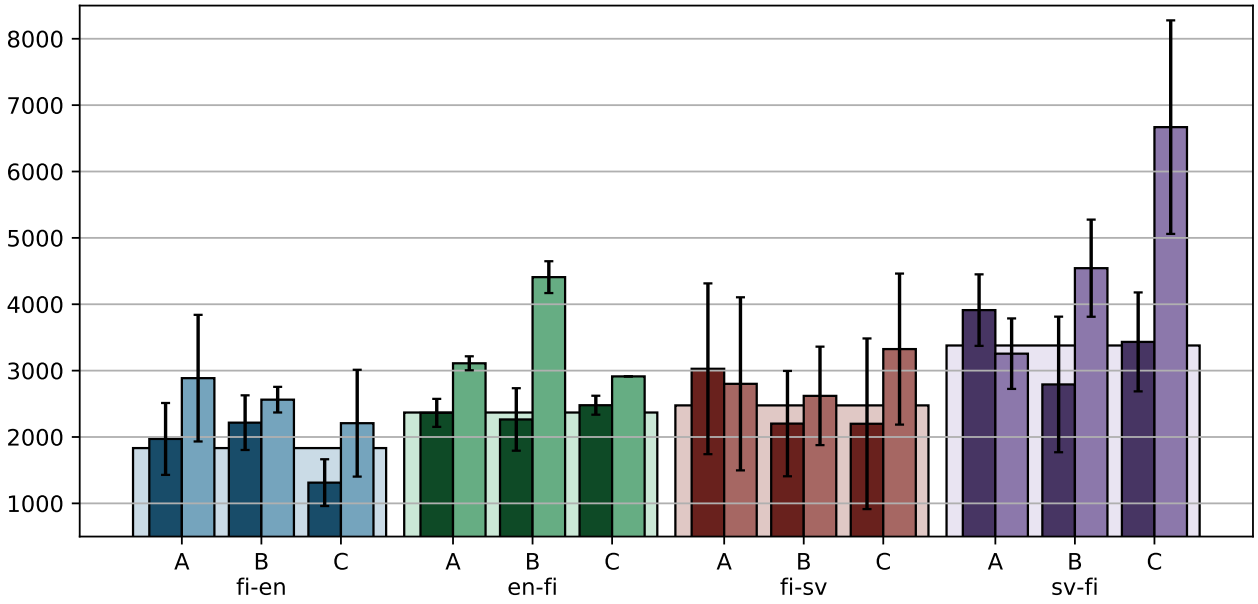

Figure 8: Average numbers of keystrokes subtitling through post-editing (left) and subtitling from scratch (right) for each participant (labelled as A, B, C for each language pair), compared to language pair averages (behind).

(edit rate) between the MT outputs and their post-edited versions. As the participants had often changed the division of text into subtitle frames by adding or deleting frames as well as moving words between frames, subtitle segmentation was ignored for calculating the edit distances. Instead, document-level scores were calculated to focus on edits affecting the textual content. Table 3 shows the HTER and characTER scores for the sentence-level and document-level MT across all four language pairs, as well as for each language pair and each participant.

	HTER	**characTER**
sent-level	**55.1 ± 17.7**	**45.0 ± 12.3**
doc-level	**60.3 ± 16.1**	**48.7 ± 11.1**
fi → en	**45.6 ± 17.7**	**39.3 ± 13.5**
post-editor A	39.4 ± 9.9	39.3 ± 15.5
post-editor B	65.4 ± 14.6	48.6 ± 8.1
post-editor C	31.9 ± 4.2	29.9 ± 11.4
en → fi	**74.1 ± 12.7**	**48.9 ± 6.4**
post-editor A	78.6 ± 3.7	52.6 ± 4.5
post-editor B	59.0 ± 2.2	41.8 ± 1.7
post-editor C	84.8 ± 9.4	52.3 ± 5.0
fi → sv	**52.7 ± 13.2**	**44.1 ± 11.4**
post-editor A	51.0 ± 12.4	45.4 ± 12.7
post-editor B	63.8 ± 6.3	50.4 ± 8.1
post-editor C	43.4 ± 12.8	36.4 ± 11.2
sv → fi	**58.4 ± 9.5**	**55.1 ± 9.2**
post-editor A	63.5 ± 9.7	59.4 ± 10.7
post-editor B	55.0 ± 12.4	51.6 ± 10.7
post-editor C	56.8 ± 5.5	54.3 ± 6.2
overall	**57.7 ± 16.9**	**46.8 ± 11.8**

Table 3: Comparison of word-level (HTER) and character-level (characTER) edit rates divided by MT system (sentence-level vs document-level), language pair (Finnish-English, English-Finnish, Finnish-Swedish, Swedish-Finnish) and post-editor.

Corresponding to the process metrics, average edit rate for the sentence-level MT output is slightly lower than for the document-level MT. Both scores, however, are over 50, indicating considerable rewriting. Variation can also be seen between the different language pairs, particularly in the case of English-Finnish, which has a much higher HTER score. In part, the higher score in this language pair may be due to the fact that for Finnish as a target language, HTER does not distinguish between changed words and changed word forms. The character-based score (characTER) shows a smaller gap between the English-Finnish case and the other language pairs, which may indicate more frequent word form edits in this language pair. However, a similar effect is not seen in the Swedish-Finnish case. This may be explained by the fact that the participants working on this language pair appear to have added a words into the post-edited versions.

Considerable variation can again be seen in the edit rates between different participants. For example, in the language pair Finnish-English one of the participants (fi-en B) has a much higher edit rate than the other two, while the two participants with the overall highest average edit rates (over 80) both worked on the language pair English-Finnish. Since the participants post-edited different MT versions, differences in output quality are possible, but to some extent

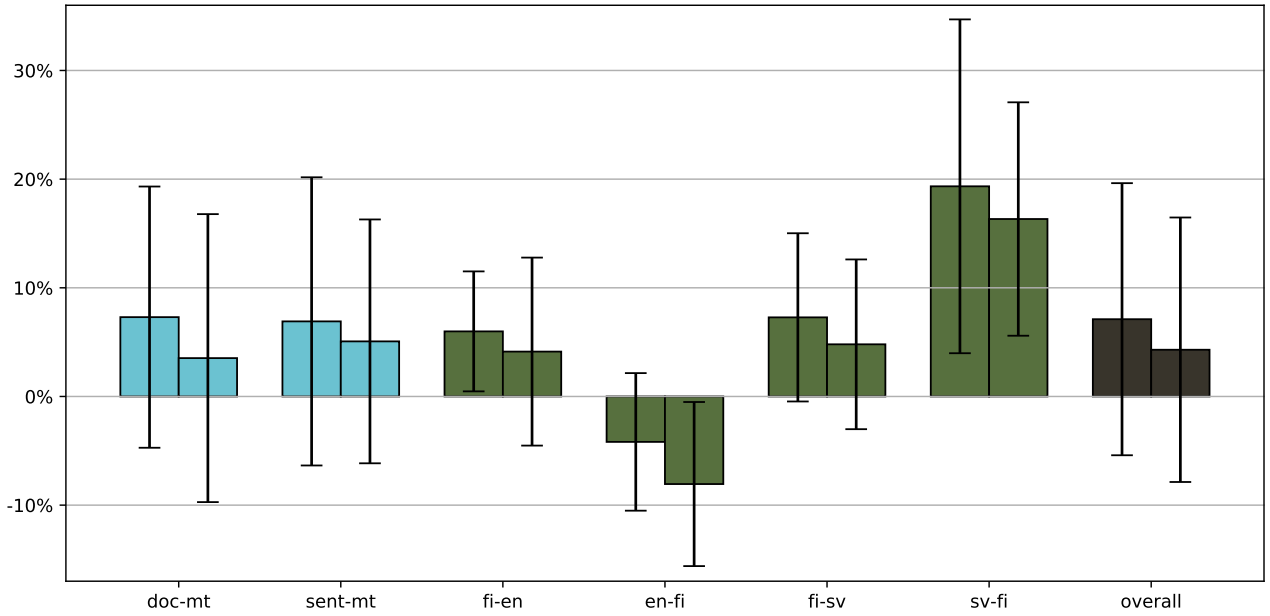

Figure 9: Average differences in number of segments (left) and lines (right).

these differences may also reflect individual preferences rather than clear errors in the MT output. Preliminary qualitative observations regarding the changes made suggest that while some edits relate to clear MT errors, many are also caused by what appear to be preferential edits; for example, in the Finnish-English language pair, one participant accepts the term "financial discipline" while another replaces it with "austerity".

In addition to the textual content of the MT subtitles, the participants edited both the splitting of that content into subtitle frames and timing of the frames. As described above, in Section 5.2, the MT output was mapped into subtitle frames from the intralingual subtitles used as source text. To estimate the changes made to segmentation of the textual content into the frames, we calculated first the number of subtitle frames and lines of text within those frames in the MT version and in the post-edited version of that MT. Figure 9 shows the average differences in number of segments (frames) and number of lines. Overall, the participants have added both frames (+7%) and lines (+4%), with no clear difference between the sentence-level and document-level outputs. However, variation is evident both across different language pairs and different clips, as indicated by the standard deviation. In the language pair Swedish-Finnish, the tendency to add frames (+19%) and lines (+16%) is particularly noticeable. This seems to arise mostly from the three clips from the lifestyle programming genre, where participants made considerable additions (for discussion of these additions, see Section 5.5). English-Finnish appears to be the only language pair where the participants have mostly reduced the number of subtitle frames (-4%) and lines (-8%). As noted above, two of the participants working with English-Finnish had done significant rewriting, often condensing the suggestion provided by the MT, which is likely also reflected in the number of segments.

Changes to the timing of subtitles frames were further analysed by comparing the in and out times of frames in the MT version of the subtitle file and the post-edited version of that MT output. Figure 10 shows the average percentage of frames in the MT version with both in and out timestamps preserved in the PE, frames with either in or out time preserved while the other was changed in PE, and frames with no matching timestamps in PE. Overall, only about 24% of the original timed frames remained in the post-edited versions, a further 27%

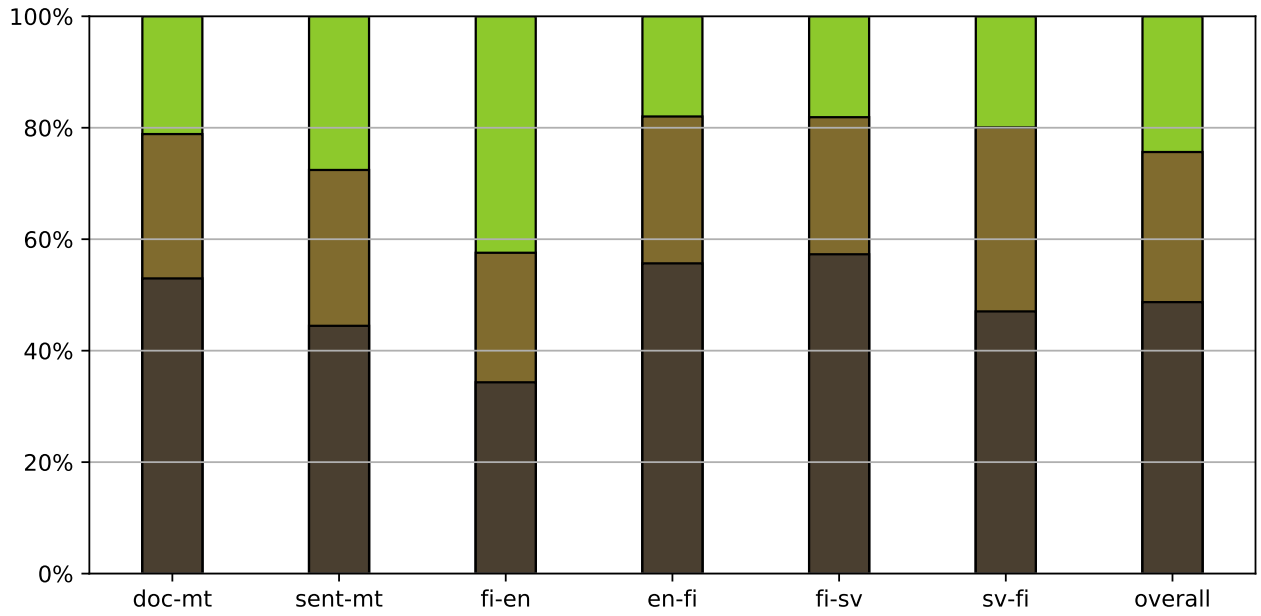

Figure 10: Average percentages of segments that were preserved (top), extended or contracted (middle), and overridden (bottom) during post-editing.

had either the in or out time changed, and 49% of frame timings were completely changed. Comparing the language pairs, a difference can also be seen in Finnish-English, where the participants appear to have accepted a larger percentage (42%) of the suggested frame timings. On average, slightly more frame timestamps were preserved in the files with sentence-level MT output than document-level output. One explanation for this is that the document-level output contained more cases of repeated fragments, which sometimes caused the segmentation of the translation into the frames to become out of sync with the audio. The large number of completely changed timestamps is likely to be connected to the fact that the participants added, or in the case of English-Finnish, deleted subtitle frames, leading to necessary changes in the timing of other frames.

5.5 Qualitative examples of discourse phenomena

A more detailed manual analysis of the MT outputs and changes carried out during post-editing is ongoing to identify specific issues related to the discourse phenomena discussed in Section 2. For example, issues related to pronouns and cohesion are being examined. To some extent, these issues are affected also by the features of the programmes in question. The largely unscripted (EU debates) and relatively informal semi-scripted (lifestyle programming) spoken language of the clips differs from written discourse. The multimodal context also brings its own challenges, for example, in terms of pronominal anaphora where the referent appears only in visual form. No definitive conclusions can yet be made regarding the frequency of specific features or errors in the MT outputs, however, some preliminary qualitative observations can be made of the types of issues arising in this dataset.

Pronominal anaphora With Finnish as source language, 3rd person singular pronouns need to be disambiguated for gender when translating into English (*she* vs *he*) and Swedish (*han* vs *hon*). As noted, this issue is further complicated by the spoken language and the colloquial style of the clips coming from the more youth-oriented lifestyle programming, where

the pronoun *se* 'it' is used for humans instead of the more formal *hän* 'he/she'. The Finnish clips contain 4 such cases where *se* is used to refer to a person (male in each case). In the Finnish-English language pair, the sentence-level MT has one correct *he* and three *it*, while the document-level MT has all but one case correctly translated as *he*. In the Finnish-Swedish language pair, the sentence-level MT has two correctly as *han* and two incorrectly as *den* or *det*, while the document-level MT has all but one correctly as *han*. The only sentence both systems in both language pairs get right is *Sillä on silmälasit.* 'It has glasses.' (translated as "he").

Explicitation of pronouns As noted in Section 2, pronouns are not necessarily always translated as pronouns. Even in cases where the pronoun is not an incorrect translation, a translator may choose replace it with the antecedent or in some cases to explicitate the implied reference. While the MT systems render pronouns as pronouns, such cases of explicitation can be seen in the participants' edits. For example, in the Finnish-English language pair, the expression *yrityksillemme* 'for our companies' is translated correctly by both MT systems, but in both cases changed to *Finnish companies* by the post-editor using the situational context (Finnish politicians discussing the economy) to explicitate what is implied by "our". Another example related to the multimodal context appears in the Swedish-Finnish language pair where the expression *de ska kokas mjuka* 'they should be cooked soft' is (correctly) translated in both the sentence-level and document-level output using the corresponding pronoun *ne*. In both cases, the participant post-editing the output replaces the pronoun with *hedelmät* 'fruit', referring to the fruit shown being cooked in the programme.

Noun phrase definiteness As Finnish does not mark definiteness using articles like English and Swedish, translating noun phrases may be ambiguous with regard to definiteness. One example appears in the Finnish-English language pair:

Source	- Ensin pitää selvittää, kuka on oikea nörtti.
doc-MT	- First we have to find out who's a real nerd.
sent-MT	First, we need to find out who the real nerd is.

In the dialogue, reference is being made to an unspecified person in a way that corresponds to the English indefinite noun phrase, as shown in the document-level MT output. The sentence-level MT output is incorrect for the meaning of the source text, and the participant post-editing that version corrects both the definiteness of the noun phrase and the word order as "who is an authentic nerd".

Multimodality and condensation in subtitles Condensation of the subtitle content by, for example, omitting repeated words can also be observed particularly in the English-Finnish language pair, possibly due to the longer average length of words in Finnish. A more detailed analysis of the edits performed would be needed to establish the extent to which edits relate to errors vs preferential changes or decisions like condensation. In contrast to omissions caused by condensation, the participants can also be seen to add textual content. While there are some cases where the additions correspond to missing words in the MT output, some in fact involve content that was also missing from the source text used to create that output. This is explained by the fact that intralingual subtitles used as source text for the MT already involve some condensation and other editing, and therefore are not an exact match with the spoken audio of the clip. This difference was particularly noticeable in the case of three of the Swedish "lifestyle" clips used for the Swedish-Finnish language pair. The intralingual subtitles of these clips appear to have been unusually condensed, leading the participants to add both textual content and new subtitle frames. These additions show one effect of the multimodal context on the participants (see 2.2). If they had been asked to translate the intralingual subtitles as

text only, they would have been unlikely to make additions not present in the written source text. However, because information missing in writing was available to them in the audio, it was added during post-editing.

6 Future work

Our previous experiments have shown that document-level machine translation is difficult and that concatenation approaches are currently the most effective method to be applied. Nevertheless, the results are still not very satisfactory and various directions of future research are possible.

First of all, the segmentation into appropriate chunks to be translated as a unit is an important issue that needs to be addressed. In relation to subtitle translation, the use of time information could be a useful approach to split the data into coherent blocks separated by significant breaks. Multimodality can also pay a crucial role in segmentation as visual and auditory cues may help to improve the division of verbal content into discourse units. For a baseline, we will implement an end-to-end system specifically tailored for subtitle translation after Matusov et al. (2019), and investigate how well such a system could generate organic subtitles.

Another direction is to integrate speaker information into the translation engines. Other features from the scenes and shots may also be useful in optimising translation quality. This is, again, an interesting idea for the incorporation of multimodal features in translation in connection with non-linguistic context for language grounding and disambiguation. For example, Lison and Meena (2016) introduce an approach for segmenting subtitles into dialogue turns by leveraging speaker labels or diarisation output. The combined MeMAD framework might allow the application of such a method to our training datasets, opening up new ways to experiment with discourse-aware MT by optimising limited document context windows.

Finally, the combination of speech recognition and translation in a discourse-aware model is another goal of future research. End-to-end models that are able to properly make use of cues from various features and signals available in video clips in connection with their narrative structure would be the ultimate goal in the MeMAD setting and our work in the final period of the project will be devoted to the development in that direction.

A detailed analysis of subjective feedback collected from the translators who participated in the post-editing experiments is also underway. Preliminary results regarding the user experience and feedback related to machine translation will also be reported as part of MeMAD deliverable D6.6. In the third year of the project, further user evaluation is planned to assess the effect of the new developments of our machine translation approaches.

7 Conclusion

This paper presents the current state-of-the-art in document-level machine translation and presents the efforts of developing and evaluating discourse-aware translation models in the MeMAD project. In this report, we focus on the analyses of professional subtitle translation supported by machine translation in the case of YLE data sets and Finnish, Swedish and English subtitles. We present the challenges of document-level MT and its shortcomings in terms of proper segmentation and integration in translation workflows. The assessment provides valu-

able insights to the remaining issues to be addressed and pinpoints the importance of manual evaluation when evaluating the utility of automatic translation in real-world applications.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT 2004 evaluation campaign. In *Proceedings of the 2004 International Workshop on Spoken Language Translation*, Kyoto, Japan.

Andrei Alexandrescu and Katrin Kirchhoff. 2009. Graph-based learning for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 119–127, Boulder, Colorado. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, pages San Diego, CA, USA.

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

Rachel Bawden. 2018. *Going Beyond the Sentence: Contextual Machine Translation of Dialogue*. Ph.D. thesis, University of Paris-Sud.

Rachel Bawden, Sophie Rosset, Thomas Lavergne, and Eric Bilinski. 2019. DiaBLa: A corpus of bilingual spontaneous written dialogues for machine translation.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. 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. Association for Computational Linguistics.

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

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Computing Research Repository*, arXiv:1409.1259.

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic MT evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338, Uppsala, Sweden. Association for Computational Linguistics.

Arnulf Deppermann. 2013. Multimodal interaction from a conversation analytic perspective. *Journal of Pragmatics*, 46:1–7.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Jorge Díaz-Cintas and Serenella Massidda. 2019. Technological advances in audiovisual translation. In Minako O'Hagan, editor, *The Routledge Handbook of Translation and Technology*, pages 255–270. Routledge, London.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. 2012a. 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. Association for Computational Linguistics.

ZhengXian Gong, Min Zhang, ChewLim Tan, and GuoDong Zhou. 2012b. Classifier-based tense model for SMT. In *Proceedings of COLING 2012: Posters*, pages 411–420, Mumbai, India. The COLING 2012 Organizing Committee.

Ana Guerberof Arenas. 2014. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28:165–186.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis.

Liane Guillou and Christian Hardmeier. 2016. 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. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. 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. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2012. Translating English discourse connectives into Arabic: A corpus-based analysis and an evaluation metric. In *Proceedings of the Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, CONF.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 236–247. Springer.

Michael A.K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. Routledge.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours (online)*.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. 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. Association for Computational Linguistics.

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

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. 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. Association for Computational Linguistics.

Maija Hirvonen. 2014. *Multimodal Representation and Intermodal Similarity. Cues of Space in the Audio Description of Film*. Ph.D. thesis, University of Helsinki.

Carey Jewitt. 2011. Different approaches to multimodality. In Carey Jewitt, editor, *The Routledge Handbook of Multimodal Analysis*, 2nd edition, pages 31–43. Routledge, London.

Shu Jiang, Rui Wang, Zuchao Li, Masao Utiyama, Kehai Chen, Eiichiro Sumita, Hai Zhao, and Bao liang Lu. 2019. Document-level neural machine translation with inter-sentence attention.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. 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. Association for Computational Linguistics.

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

Klaus Kaindl. 2013. Multimodality and translation. In Carmen Millán and Francesca Bartrina, editors, *The Routledge Handbook of Translation Studies*, pages 257–269. Routledge, London.

Anne Ketola. 2018. Word – image interaction in technical translation: students translating an illustrated text. (April).

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI)*, volume 94, pages 779–784.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. pages 79–86.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Philipp Koehn, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, and Christine Moran. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, Czech Republic. Association for Computational Linguistics.

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*, pages 11–20, San Diego, California.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Olli Philippe Lautenbacher. 2018. The relevance of redundancy in multimodal documents. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 17:215–230.

Ronan Le Nagard and Philipp Koehn. 2010. 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. Association for Computational Linguistics.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3):358–392.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation for movie & TV subtitles. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 245–252.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Sharid Loáiciga. 2017. *Pronominal Anaphora and Verbal Tenses in Machine Translation*. Ph.D. thesis, University of Geneva.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017a. What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark. Association for Computational Linguistics.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017b. 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. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. 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. Association for Computational Linguistics.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Thomas Meyer. 2011. Disambiguating temporal-contrastive connectives for machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 46–51, Portland, OR, USA. Association for Computational Linguistics.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, CONF.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. 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. Association for Computational Linguistics.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. 2018. Introduction. In *Translation Quality Assessment: From Principles to Practice*, Machine Translation: Technologies and Applications, pages 1–6. Springer International Publishing, Cham.

Jeremy Munday. 2012. *Introducing translation studies*, 3rd edition. Routledge, Oxford and New York.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 2018 International Workshop on Spoken Language Translation*, Bruges, Belgium.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loïc Barrault, Lucia Specia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*.

Carol O'Sullivan. 2013. Multimodality as challenge and resource for translation. *Journal of Specialised Translation*, 20:2–14.

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

Luis Pérez-González. 2014. *Audiovisual Translation: Theories, Methods and Issues*. Routledge, London.

Mirko Plitt and François Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong, editors. 2019. *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. Association for Computational Linguistics, Hong Kong, China.

Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. 2012. Improving machine translation of null subjects in Italian and Spanish. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–89, Avignon, France. Association for Computational Linguistics.

Ted Sanders and Henk Pander Maat. 2006. *Cohesion and Coherence: Linguistic Approaches*.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. 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. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, 6.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2019. Multimodal machine translation through visuals and speech.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Aarne Talman, Umut Sulubacak, Raúl Vázquez, Yves Scherrer, Sami Virpioja, Alessandro Raganato, Arvi Hurskainen, and Jörg Tiedemann. 2019. The university of Helsinki submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 412–423, Florence, Italy. Association for Computational Linguistics.

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

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). Http://www.lrec-conf.org/proceedings/lrec2008/.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 417–426, Montréal, Canada. Association for Computational Linguistics.

Len Unsworth and Chris Cléirigh. 2011. Multimodality and reading: the construction of meaning through image-text interaction. In Carey Jewitt, editor, *The Routledge Handbook of Multimodal Analysis*, pages 176–187. Routledge, London.

D. Varga, L. NiE9¿meth, P. HaliE1¿csy, A. Kornai, V. TriF3¿n, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 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.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Tony Veale and Andy Way. 1997. Gaijin: A template-driven bootstrapping approach to example-based machine translation. In *Proceedings of the 1997 Conference on New Methods in Natural Language Processessing (NeMNLP'97)*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. 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. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier, editors. 2015. *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal.

Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark.

Billy T. M. Wong and Chunyu Kit. 2012. 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. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of the 13th Machine Translation Summit*, volume 13, pages 131–138.

Meifang Zhang, Hanting Pan, Xi Chen, and Tian Luo. 2015. Mapping discourse analysis in translation studies via bibliometrics: A survey of journal publications. *Perspectives*, 23(2):223–239.

A Dissemination activities

- **Workshop presentation** 6.5.2019: FCAI Research Insight: Natural Language Processing (6/5/2019 Aalto University, Espoo) - presentation on "Recent Advances in Aalto ASR Group" (Mikko Kurimo), "Found in translation - learning to understand human languages with multilingual data" (Jörg Tiedemann) and "NLP components in industry service solutions" (Sebastian Andersson)
- **Conference presentation** 1.-2.8.2019: WMT 2019 news translation task paper and poster presentation (programme: http://www.statmt.org/wmt19/program.html, paper: https://www.aclweb.org/anthology/W19-5347.pdf)
- WMT 2019: The University of Helsinki Submission to the WMT19 Parallel Corpus Filtering Task (https://www.aclweb.org/anthology/W19-5441.pdf)
- **Guest lecture** 1.10.2019: "Machine Translation and Post-editing: Researching Quality and Effort", at Linnaeus University, Växjö, Sweden (Maarit Koponen).
- **Workshop presentation** 23.10.2019: Huawei Workshop on Cloud Computing and Data Security (Helsinki, Finland) - presentation on "Opportunities and Challenges in Natural Language Understanding - in a Multilingual Setup"
- **Workshop presentation** 3.11.2019: DiscoMT 2019 paper and poster presentation + poster booster presentation (programme: https://www.idiap.ch/workshop/DiscoMT/program, paper: https://www.aclweb.org/anthology/D19-6506.pdf)
- **Conference presentation** 7.11.2019: HELDIG Summit - presentation on "Learning to understand languages with neural networks" (programme: https://www.helsinki.fi/en/helsinki-centre-for-digital-humanities/heldig-digital-humanities-summit-2019, slides: https://a3s.fi/heldig/summit-2019/07-tiedemann.pdf)
- **Conference presentation** 26.11.2019: AI Day (Espoo, Finland) - presentation on "Language (Technology) is the key to (Artificial) Intelligence" (programme: https://fcai.fi/ai-day-2019-program, slides: https://static1.squarespace.com/static/59d528238419c2639782a4eb/t/5dde6354f0214772d10303d7/1574855524833/2.+J\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{o\global\mathchardef\accent@spacefactor\spacefactor}\accent127o\egroup\spacefactor\accent@spacefactorrg+Tiedemann%2C+AI+Day+2019%2C+Finland.pdf)
- arXiv submission of Survey on Multimodal MT through Visuals and Speech (Sulubacak et al., 2019) (url: https://arxiv.org/abs/1911.12798)
- **Media** 2.12.2019: Interview published in news section of the University of Helsinki website https://www.helsinki.fi/en/news/language-culture/machine-translation-no-match-for-humans-machines-translate-words-humans-the-underl (Maarit Koponen)
- **Workshop presentation** 16.12.2019: Invited talk on "Konekääntimen rakentamisen tekijänoikeusnäkökohtia" (Copyright aspects of building a machine translation engine) in seminar *Kun kone kääntää, kuka on tekijä?* (When a machine translates, who is the author?) organised by the research project "Art, Copyright and the Transformation of Authorship" University of Helsinki (https://blogs.helsinki.fi/taidejatekijanoikeus/) (Maarit Koponen)

B Appendices

B.1 MeMAD WMT document-level MT task paper

The University of Helsinki submissions to the WMT19 news translation task

Aarne Talman,[*] [†] **Umut Sulubacak,**[*] **Raúl Vázquez,**[*] **Yves Scherrer,**[*] **Sami Virpioja,**[*]
Alessandro Raganato,[*] [†] **Arvi Hurskainen**[*] and **Jörg Tiedemann**[*]

[*]University of Helsinki
[†]Basement AI
{name.surname}@helsinki.fi

Abstract

In this paper, we present the University of Helsinki submissions to the WMT 2019 shared task on news translation in three language pairs: English–German, English–Finnish and Finnish–English. This year, we focused first on cleaning and filtering the training data using multiple data-filtering approaches, resulting in much smaller and cleaner training sets. For English–German, we trained both sentence-level transformer models and compared different document-level translation approaches. For Finnish–English and English–Finnish we focused on different segmentation approaches, and we also included a rule-based system for English–Finnish.

1 Introduction

The University of Helsinki participated in the WMT 2019 news translation task with four primary submissions. We submitted neural machine translation systems for English-to-Finnish, Finnish-to-English and English-to-German, and a rule-based machine translation system for English-to-Finnish.

Most of our efforts for this year's WMT focused on data selection and pre-processing (Section 2), sentence-level translation models for English-to-German, English-to-Finnish and Finnish-to-English (Section 3), document-level translation models for English-to-German (Section 4), and a comparison of different word segmentation approaches for Finnish (Section 3.3). The final submitted NMT systems are summarized in Section 5, while the rule-based machine translation system is described in Section 3.4.

2 Pre-processing, data filtering and back-translation

It is well known that data pre-processing and selection has a huge effect on translation quality in neural machine translation. We spent substantial effort on filtering data in order to reduce noise—especially in the web-crawled data sets—and to match the target domain of news data.

The resulting training sets, after applying the steps described below, are for 15.7M sentence pairs for English–German, 8.5M sentence pairs for English–Finnish, and 12.3M–26.7M sentence pairs (different samplings of back-translations) for Finnish–English.

2.1 Pre-processing

For each language, we applied a series of pre-processing steps using scripts available in the Moses decoder (Philipp Koehn, 2007):

- replacing unicode punctuation,
- removing non-printing characters,
- normalizing punctuation,
- tokenization.

In addition to these steps, we replaced a number of English contractions with the full form, *e.g.* *"They're"* → *"They are"*. After the above steps, we applied a Moses truecaser model trained for individual languages, and finally a byte-pair encoding (BPE) (Sennrich et al., 2016b) segmentation using a set of codes for either language pair.

For English–German, we initially pre-processed the data using only punctuation normalization and tokenization. We subsequently trained an English truecaser model using all monolingual English data as well as the English side of all parallel English–German datasets except the Rapid corpus (in which non-English characters were missing from a substantial portion of the German sentences). We also repeated the same for German. Afterwards, we used a heuristic cleanup script[1] in

[1]Shared by Marcin Junczys-Dowmunt. Retrieved

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 412–423
Florence, Italy, August 1-2, 2019.

order to filter suspicious samples out of Rapid, and then truecased all parallel English–German data (including the filtered Rapid) using these models. Finally, we trained BPE codes with 35 000 symbols jointly for English–German on the truecased parallel sets. For all further experiments with English–German data, we applied the full set of tokenization steps as well as truecasing and BPE segmentation.

For English–Finnish, we first applied the standard tokenization pipeline. For English and Finnish respectively, we trained truecaser models on all English and Finnish monolingual data as well as the English and Finnish side of all parallel English–Finnish datasets. As we had found to be optimal in our previous year submission (Raganato et al., 2018), we trained a BPE model using a vocabulary of 37 000 symbols, trained jointly only on the parallel data. Furthermore, for some experiments, we also used domain labeling. We marked the datasets with 3 different labels: ⟨*NEWS*⟩ for the development and test data from 2015, 2016, 2017, ⟨*EP*⟩ for Europarl, and ⟨*WEB*⟩ for ParaCrawl and Wikititles.

2.2 Data filtering

For data filtering we applied four types of filters: (i) rule-based heuristics, (ii) filters based on language identification, (iii) filters based on word alignment models, and (iv) language model filters.

Heuristic filters: The first step in cleaning the data refers to a number of heuristics (largely inspired by (Stahlberg et al., 2018)) including:

- removing all sentence pairs with a length difference ratio above a certain threshold: for CommonCrawl, ParaCrawl and Rapid we used a threshold of 3, for WikiTitles a threshold of 2, and for all other data sets a threshold of 9;
- removing pairs with short sentences: for CommonCrawl, ParaCrawl and Rapid we required a minimum number of four words;
- removing pairs with very long sentences: we restricted all data to a maximum length of 100 words;
- removing sentences with extremely long words: We excluded all sentence pairs with words of 40 or more characters;
- removing sentence pairs that include HTML or XML tags;
- decoding common HTML/XML entities;
- removing empty alignments (while keeping document boundaries intact);
- removing pairs where the sequences of non-zero digits occurring in either sentence do not match;
- removing pairs where one sentence is terminated with a punctuation mark and the other is either missing terminal punctuation or terminated with another punctuation mark.

Language identifiers: There is a surprisingly large amount of text segments in a wrong language in the provided parallel training data. This is especially true for the ParaCrawl and Rapid data sets. This is rather unexpected as a basic language identifier certainly must be part of the crawling and extraction pipeline. Nevertheless, after some random inspection of the data, we found it necessary to apply off-the-shelf language identifiers to the data for removing additional erroneous text from the training data. In particular, we applied the Compact Language Detector version 2 (CLD2) from the Google Chrome project (using the Python interface from *pycld2*[2]), and the widely used *langid.py* package (Lui and Baldwin, 2012) to classify each sentence in the ParaCrawl, CommonCrawl, Rapid and Wikititles data sets. We removed all sentence pairs in which the language of one of the aligned sentences was not reliably detected. For this, we required the correct language ID from both classifiers, the reliable-flag set to "True" by CLD2 with a reliability score of 90 or more, and the detection probability of *langid.py* to be at least 0.9.

Word alignment filter: Statistical word alignment models implement a way of measuring the likelihood of parallel sentences. IBM-style alignment models estimate the probability $p(f \mid a, e)$ of a foreign sentence f given an "emitted" sentence e and an alignment a between them. Training word alignment models and aligning large corpora is very expensive using traditional methods

from https://gist.github.com/emjotde/4c5303e3b2fc501745ae016a8d1e8e49

[2] https://github.com/aboSamoor/pycld2

and implementations. Fortunately, we can rely on *eflomal*[3], an efficient word aligner based on Gibbs sampling (Östling and Tiedemann, 2016). Recently, the software has been updated to allow the storage of model priors that makes it possible to initialize the aligner with previously stored model parameters. This is handy for our filtering needs as we can now train a model on clean parallel data and apply that model to estimate alignment probabilities of noisy data sets.

We train the alignment model on Europarl and news test sets from previous WMTs for English–Finnish, and NewsCommentary for English–German. For both language pairs, we train a Bayesian HMM alignment model with fertilities in both directions and estimate the model priors from the symmetrized alignment. We then use those priors to run the alignment of the noisy data sets using only a single iteration of the final model to avoid a strong influence of the noisy data on alignment parameters. As it is intractable to estimate a fully normalized conditional probability of a sentence pair under the given higher-level word alignment model, eflomal estimates a score based on the maximum unnormalized log-probability of links in the last sampling iteration. In practice, this seems to work well, and we take that value to rank sentence pairs by their alignment quality. In our experiments, we set an arbitrary threshold of 7 for that score, which seems to balance recall and precision well according to some superficial inspection of the ranked data. The word alignment filter is applied to all web data as well as to the back-translations of monolingual news.

Language model filter: The most traditional data filtering method is probably to apply a language model. The advantage of language models is that they can be estimated from monolingual data, which may be available in sufficient amounts even for the target domain. In our approach, we opted for a combination of source and target language models and focused on the comparison between scores coming from both models. The idea is to prefer sentence pairs for which not only the cross-entropy of the individual sentences ($H(S, q_s)$ and $H(T, q_t)$) is low with respect to in-domain LMs, but also the absolute difference between the cross-entropies ($abs(H(S, q_s) - H(T, q_t))$) for aligned source and target sentences is low. The intuition is that both models should be roughly similarly surprised when observing sentences that are translations of each other. In order to make the values comparable, we trained our language models on parallel data sets.

For English–Finnish, we used news test data from 2015-2017 as the only available in-domain parallel training data, and for English–German we added the NewsCommentary data set to the news test sets from 2008-2018. As both data sets are small, and we aimed for an efficient and cheap filter, we opted for a traditional n-gram language model in our experiments. To further avoid data sparseness and to improve comparability between source and target language, we also based our language models on BPE-segmented texts using the same BPE codes as for the rest of the training data. *VariKN* (Siivola et al., 2007b,a)[4] is the perfect toolkit for the purposes of estimating n-gram language models with subword units. It implements Kneser-Ney growing and revised Kneser-Ney pruning methods with the support of n-grams of varying size and the estimation of word likelihoods from text segmented in subword units. In our case, we set the maximum n-gram size to 20, and the pruning threshold to 0.002. Finally, we computed cross-entropies for each sentence in the noisy parallel training data and stored 5 values as potential features for filtering: $H(S, q_s)$, $H(T, q_t)$, $avg(H(S, q_s), H(T, q_t))$, $max(H(S, q_s), H(T, q_t))$ and $abs(H(S, q_s) - (T, q_t))$. Based on some random inspection, we selected a threshold of 13 for the average cross-entropy score, and a threshold of 4 for the cross-entropy difference score. For English–Finnish, we opted for a slightly more relaxed setup to increase coverage, and set the average cross-entropy to 15 and the difference threshold to 5. We applied the language model filter to all web data and to the back-translations of monolingual news.

Applying the filter to WMT 2019 data: The impact of our filters on the data provided by WMT 2019 is summarized in Tables 1, 2 and 3.

We can see that the ParaCrawl corpus is the one that is the most affected by the filters. A lot of noise can be removed, especially by the language model filter. The strict punctuation filter also has a strong impact on that data set. Naturally, web data does not come with proper com-

[3]Software available from https://github.com/robertostling/eflomal

[4]VariKN is available from https://vsiivola.github.io/variKN/

	EN–DE	EN–FI
CommonCrawl	3.2%	
Europarl	0.8%	2.8%
News-Commentary	0.2%	
ParaCrawl	0.6%	
Rapid	13.2%	5.2%
WikiTitles	8.0%	4.0%

Table 1: Basic heuristics for filtering – percentage of lines removed. For English–Finnish the statistics for ParaCrawl are not available because the cleanup script was applied after other filters.

	% rejected		
Filter	CC	ParaCrawl	Rapid
LM average CE	31.9%	62.0%	12.7%
LM CE diff	19.0%	12.7%	6.9%
Source lang ID	4.0%	30.7%	7.3%
Target lang ID	8.0%	22.7%	6.2%
Wordalign	46.4%	3.1%	8.4%
Number	15.3%	16.0%	5.0%
Punct	0.0%	47.4%	18.7%
total	66.7%	74.7%	35.1%

Table 2: Percentage of lines rejected by each filter for English–German data sets. Each line can be rejected by several filters. The total of rejected lines is the last row of the table.

	% rejected			
	ParaCrawl		Rapid	
Filter	strict	relax	strict	relax
LM avg CE	62.5%	40.0%	50.7%	21.4%
LM CE diff	35.4%	25.7%	44.8%	31.1%
Src lang ID	37.2%	37.2%	11.9%	11.9%
Trg lang ID	29.1%	29.1%	8.5%	8.5%
Wordalign	8.3%	8.3%	8.3%	8.3%
Number	16.8%	16.8%	6.7%	6.7%
Punct	54.6%	3.3%	23.7%	7.6%
total	87.9%	64.2%	62.2%	54.8%

Table 3: Percentage of lines rejected by each filter for English–Finnish data sets. The strict version is the same as for English–German, and the relax version applies relaxed thresholds.

plete sentences that end with proper final punctuation marks, and the filter might remove quite a bit of the useful data examples. However, our final translation scores reflect that we do not seem to lose substantial amounts of performance even with the strict filters. Nevertheless, for English–Finnish, we still opted for a more relaxed setup to increase coverage, as the strict version removed over 87% of the ParaCrawl data.

It is also interesting to note the differences of individual filters on different data sets. The word alignment filter seems to reject a large portion of the CommonCrawl data set whereas it does not affect other data sets that much. The importance of language identification can be seen with the ParaCrawl data whereas other corpora seem to be much cleaner with respect to language.

2.3 Back-translation

We furthermore created synthetic training data by back-translating news data. We translated the monolingual English news data from the years 2007–2018, from which we used a filtered and sampled subset of 7M sentences for our Finnish–English systems, and the Finnish data from years 2014–2018 using our WMT 2018 submissions. We also used the back-translations we generated for the WMT 2017 news translation task, where we used an SMT model to create 5.5M sentences of back-translated data from the Finnish news2014 and news2016 corpora (Östling et al., 2017).

For the English–German back-translations, we trained a standard transformer model on all the available parallel data and translated the monolingual German data into English. The BLEU score for our back-translation model is 44.24 on newstest 2018. We applied our filtering pipeline to the back-translated pairs, resulting in 10.3M sentence pairs. In addition to the new back-translations, we also included back-translations from the WMT16 data by Sennrich et al. (2016a).

3 Sentence-level approaches

In this section we describe our sentence-level translation models and the experiments in the English-to-German, English-to-Finnish and Finnish-to-English translation directions.

3.1 Model architectures

We experimented with both NMT and rule-based systems. All of our neural sentence-level models are based on the transformer architecture (Vaswani et al., 2017). We used both the OpenNMT-py (Klein et al., 2017) and MarianNMT (Junczys-

Dowmunt et al., 2018) frameworks. Our experiments focused on the following:

- Ensemble models: using ensembles with a combination of independent runs and save-points from a single training run.
- Left-to-right and right-to-left models: Transformer models with decoding of the output in left-to-right and right-to-left order.

The English-to-Finnish rule-based system is an enhanced version of the WMT 2018 rule-based system (Raganato et al., 2018).

3.2 English–German

Our sentence-level models for the English-to-German direction are based on ensembles of independent runs and different save-points as well as save-points fine-tuned on in-domain data. For our submission, we used an ensemble of 9 models containing:

- 4 save-points with the lowest development perplexity taken from a model trained for 300 000 training steps.
- 5 independent models fine-tuned with in-domain data.

All our sentence-level models for the English–German language pair are trained on filtered versions of Europarl, NewsCommentary, Rapid, CommonCrawl, ParaCrawl, Wikititles, and back-translations. For in-domain fine-tuning, we use newstest 2011–2016. Our submission is composed of transformer-big models implemented in OpenNMT-py with 6 layers of hidden size 4096, 16 attention heads, and a dropout of 0.1. The differences in development performance between the best single model, an ensemble of save-points of a single training run and our final submission are reported in Table 4. We gain 2 BLEU points with the ensemble of save-points, and an additional 0.8 points by adding in-domain fine-tuned models into the ensemble. This highlights the well-known effectiveness of ensembling and domain adaptation for translation quality.

Furthermore, we trained additional models using MarianNMT with the same training data and fine-tuning method. In this case, we also included right-to-left decoders that are used as a complement in the standard left-to-right decoders in rescoring approaches. In total, we also end up with 9 models including:

	BLEU news2018
Single model	44.61
5 save-points	46.65
5 save-points + 4 fine-tuned	**47.45**

Table 4: English–German development results comparing the best single model, an ensemble of 5 save-points, and an ensemble of 5 save-points and 4 independent runs fine-tuned on in-domain data.

- 3 independent models trained for left-to-right decoding,
- 3 independent models trained for right-to-left decoding,
- 3 save-points based on continued training of one of the left-to-right decoding models.

The save-points were added later as we found out that models kept on improving when using larger mini-batches and less frequent validation in early stopping. Table 5 lists the results of various models on the development test data from 2018.

	BLEU news2018	
Model	Basic	Fine-tuned
L2R run 1	43.63	45.31
L2R run 2	43.52	45.14
L2R run 3	43.33	44.93
L2R run3 cont'd 1	43.65	45.11
L2R run3 cont'd 2	43.76	45.43
L2R run3 cont'd 3	43.53	45.67
Ensemble all L2R	44.61	46.34
Rescore all L2R		46.49
R2L run 1	42.14	43.80
R2L run 2	41.96	43.67
R2L run 3	42.17	43.91
Ensemble all R2L	43.03	44.70
Rescore all R2L		44.73
Rescore all L2R+R2L		**46.98**

Table 5: English–German results from individual MarianNMT transformer models and their combinations (cased BLEU).

There are various trends that are interesting to point out. First of all, fine-tuning gives a consistent boost of 1.5 or more BLEU points. Our initial runs were using a validation frequency of 5 000 steps and a single GPU with dynamic mini-batches

that fit in 13G of memory. The stopping criterion was set to 10 validation steps without improving cross-entropy on heldout data (newstest 2015 + 2016). Later on, we switched to multi-GPU training with two GPUs and early stopping of 20 validation steps. The dynamic batching method of MarianNMT produces larger minibatches once there is more memory available, and multi-GPU settings simply multiply the working memory for that purpose. We realized that this change enabled the system to continue training substantially, and Table 5 illustrates the gains of that process for the third L2R model.

Another observation is that right-to-left decoding models in general work less well compared to the corresponding left-to-right models. This is also apparent with the fine-tuned and ensemble models that combine independent runs. The difference is significant with about 1.5 BLEU points or more. Nevertheless, they still contribute to the overall best score when re-scoring n-best lists from all models in both decoding directions. In this example, re-scoring is done by simply summing individual scores. Table 5 also shows that re-scoring is better than ensembles for model combinations with the same decoding direction because they effectively increase the beam size as the hypotheses from different models are merged before re-ranking the combined and re-scored n-best lists.

The positive effect of beam search is further illustrated in Figure 1. All previous models were run with a beam size of 12. As we can see, the general trend is that larger beams lead to improved performance, at least until the limit of 64 in our experiments. Beam size 4 is an exception in the left-to-right models.

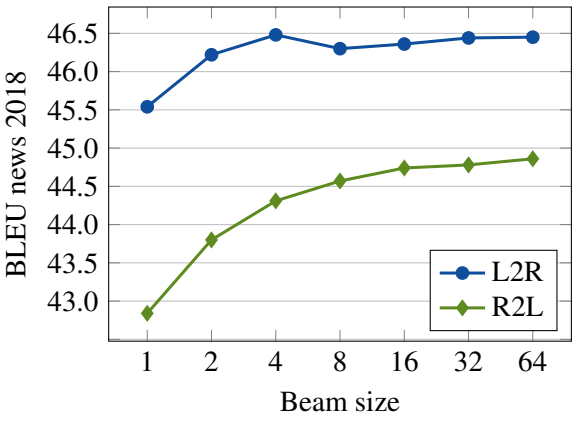

Figure 1: The effect of beam size on translation performance. All results use model ensembles and the scores are case-sensitive.

3.3 English–Finnish and Finnish–English

The problem of open-vocabulary translation is particularly acute for morphologically rich languages like Finnish. In recent NMT research, the standard approach consists of applying a word segmentation algorithm such as BPE (Sennrich et al., 2016b) or SentencePiece (Kudo and Richardson, 2018) during pre-processing. In recent WMT editions, various alternative segmentation approaches were examined for Finnish: hybrid models that back off to character-level representations (Östling et al., 2017), and variants of the Morfessor unsupervised morphology algorithm (Grönroos et al., 2018). This year, we experimented with rule-based word segmentation based on Omorfi (Pirinen, 2015). Omorfi is a morphological analyzer for Finnish with a large-coverage lexicon. Its segmentation tool[5] splits a word form into morphemes as defined by the morphological rules. In particular, it distinguishes prefixes, infixes and suffixes through different segmentation markers:

Intia→	*←n*	*ja*	*Japani→*	*←n*	*pää→*	*←ministeri→*
India	GEN	and	Japan	GEN	prime	minister

←t	*tapaa→*	*←vat*	*Tokio→*	*←ssa*
PL	meet	3PL	Tokyo	INE

While Omorfi provides word segmentation based on morphological principles, it does not rely on any frequency cues. Therefore, the standard BPE algorithm is run over the Omorfi-segmented text in order to split low-frequency morphemes.

In this experiment, we compare two models for each translation direction:

- One model segmented with the standard BPE algorithm (joint vocabulary size of 50 000, vocabulary frequency threshold of 50).
- One model where the Finnish side is pre-segmented with Omorfi, and both the Omorfi-segmented Finnish side and the English side are segmented with BPE (same parameters as above).

All models are trained on filtered versions of Europarl, ParaCrawl, Rapid, Wikititles, newsdev2015 and newstest2015 as well as back-translations. Following our experiments at WMT

[5] https://flammie.github.io/omorfi/pages/usage-examples.html#morphological-segmentation

2018 (Raganato et al., 2018), we also use domain labels (⟨*EP*⟩ for Europarl, ⟨*Web*⟩ for ParaCrawl, Rapid and Wikititles, and ⟨*NEWS*⟩ for newsdev, newstest and the back-translations). We use newstest2016 for validation. All models are trained with MarianNMT, using the standard Transformer architecture.

Figures 2 and 3 show the evolution of BLEU scores on news2016 during training. For English–Finnish, the Omorfi-segmented system shows slightly higher results during the first 40 000 training steps, but is then outperformed by the plain BPE-segmented system. For Finnish–English, the Omorfi-segmented system obtains higher BLEU scores much longer, until both systems converge after about 300 000 training steps.

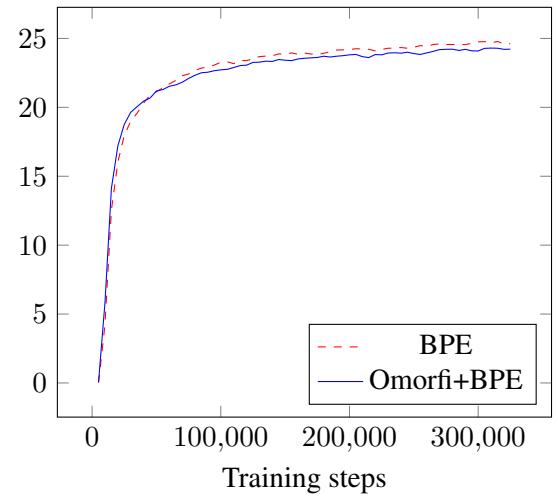

Figure 2: Evolution of English–Finnish BLEU scores (on y-axis) during training.

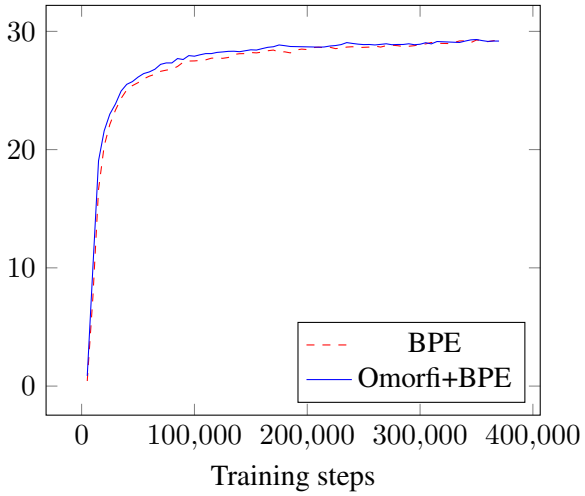

Figure 3: Evolution of Finnish–English BLEU scores (on y-axis) during training.

Table 6 compares BLEU scores for the 2017 to 2019 test sets. The Omorfi-based system shows consistent improvements when used on the source side, *i.e.* from Finnish to English. However, due to timing constraints, we were not able to integrate the Omorfi-based segmentation into our final submission systems. In any case, the difference observed in the news2019 set after submission deadline is within the bounds of random variation.

Data set	Δ BLEU EN-FI	Δ BLEU FI-EN
news2017	−0.47	+0.36
news2018	−0.61	+0.38
news2019	+0.19	+0.04

Table 6: BLEU score differences between Omorfi-segmented and BPE-segmented models. Positive values indicate that the Omorfi+BPE model is better, negative values indicate that the BPE model is better.

We tested additional transformer models segmented with the SentencePiece toolkit, using a shared vocabulary of 40k tokens trained only on the parallel corpora. We do this with the purpose of comparing the use of a software tailored specifically for Finnish language (Omorfi) with a more general segmentation one. These models were trained with the same specifications as the previous ones, including the transformer hyperparameters, the train and development data and the domain-labeling. Since we used OpenNMT-py to train these models, it is difficult to know whether the differences come from the segmentation or the toolkit. We, however, find it informative to present these results. Table 7 presents the obtained BLEU scores with both systems.

We notice that both systems yield similar scores for both translation directions. SentencePiece models are consistently ahead of Omorfi+BPE, but this difference is so small that it cannot be considered convincing nor significant.

Our final models for English-to-Finnish are standard transformer models with BPE-based segmentation, trained using MarianNMT with the same settings and hyper-parameters as the other experiments. We used the filtered training data using the relaxed settings of the language model filter to obtain better coverage for this language pair. The provided training data is much smaller and we also have less back-translated data at our disposal, which motivated us to lower the threshold

Model		news 2017	news 2019
SentencePiece	EN-FI	25.60	20.60
Omorfi+BPE	EN-FI	25.50	20.13
SentencePiece	EN-FI	31.50	25.00
Omorfi+BPE	FI-EN	31.21	24.06

Table 7: BLEU scores comparison between SentencePiece and Omorfi+BPE-segmented models.

of taking examples from web-crawled data. Domain fine-tuning is done as well using news test sets from 2015, 2016 and 2018. The results on development test data from 2017 are listed in Table 8.

Model	BLEU news2017 L2R	R2L
Run 1	27.68	28.01
Run 2	28.64	28.77
Run 3	28.64	28.41
Ensemble	29.54	29.76
Rescored	29.60	29.72
– L2R+R2L	**30.66**	
Top matrix	21.7	

Table 8: Results from individual MarianNMT transformer models and their combinations for English to Finnish (cased BLEU). The *top matrix* result refers to the best system reported in the on-line evaluation matrix (accessed on May 16, 2019).

A striking difference to English–German is that right-to-left decoding models are on par with the other direction. The scores are substantially higher than the currently best (post-WMT 2017) system reported in the on-line evaluation matrix for this test set, even though this also refers to a transformer with a similar architecture and back-translated monolingual data. This system does not contain data derived from ParaCrawl, which was not available at the time, and the improvements we achieve demonstrate the effectiveness of our data filtering techniques from the noisy on-line data.

For Finnish-to-English, we trained MarianNMT models using the same transformer architecture as for the other language pairs. Table 9 shows the scores of individual models and their combinations on the development test set of news from WMT 2017. All models are trained on the same filtered training data using the strict settings of the language model filter including the back-translations produced for English monolingual news.

Model	BLEU news2017 L2R	R2L
Run 1	32.26	31.70
Run 2	31.91	31.83
Run 3	32.68	31.81
Ensemble	33.23	33.03
Rescored	33.34	32.98
– L2R+R2L	**33.95**	
Top (with ParaCrawl)	34.6	
Top (without ParaCrawl)	25.9	

Table 9: Results from individual MarianNMT transformer models and their combinations for Finnish to English (cased BLEU). Results denoted as top refer to the top systems reported at the on-line evaluation matrix (accessed on May 16, 2019), one trained with the 2019 data sets and one with 2017 data.

In contrast to English-to-German, models in the two decoding directions are quite similar again and the difference between left-to-right and right-to-left models is rather small. The importance of the new data sets from 2019 are visible again and our system performs similarly, but still slightly below the best system that has been submitted this year to the on-line evaluation matrix on the 2017 test set.

3.4 The English–Finnish rule-based system

Since the WMT 2018 challenge, there has been development in four areas of translation process in the rule-based system for English–Finnish:

1. The standard method in handling English noun compounds was to treat them as multiword expressions (MWE). This method allows many kinds of translations, even multiple translation, which can be handled in semantic disambiguation. However, because noun compounding is a common phenomenon, also a default handling method was developed for such cases, where two or more consecutive nouns are individually translated and glued together as a single word. The system works so that if the noun combination is not handled as MWE, the second strategy is applied (Hurskainen, 2018a).

2. The translation of various types of questions has been improved. Especially the translation of indirect questions was defective, because the use of*if* in the role of initiating the indirect question was not implemented. The conjunction*if* is ambiguous, because it is used also for initiating the conditional clause (Hurskainen, 2018b).

3. Substantial rule optimizing was carried out. When rules are added in development process, the result is often not optimal. There are obsolete rules and the rules may need new ordering. As a result, a substantial number of rules (30%) were removed and others were reordered. This has effect on translation speed but not on translation result (Hurskainen, 2018c).

4. Temporal subordinate clauses, which start with the conjunction *when* or *while*, can be translated with corresponding subordinate clauses in Finnish. However, such clauses are often translated with participial phrase constructions. Translation with such constructions was tested. The results show that although they can be implemented, they are prone to mistakes (Hurskainen, 2018d).

These improvements to the translation system contribute to fluency and accuracy of translations.

4 Document-level approaches

To evaluate the effectiveness of various document-level translation approaches for the English–German language pair, we experimented with a number of different approaches which are described below. In order to test the ability of the system to pick up document-level information, we also created a shuffled version of the news data from 2018. We then test our systems on both the original test set with coherent test data divided into short news documents and the shuffled test set with broken coherence.

4.1 Concatenation models

Some of the previously published approaches use concatenation of multiple source-side sentences in order to extend the context of the currently translated sentence (Tiedemann and Scherrer, 2017). In addition to the source-side concatenation model, we also tested an approach where we concatenate the previously translated sentence with the current source sentence. The concatenation approaches we tested are listed below.

- MT-concat-source: (2+1) Concatenating previous source sentence with the current source sentence (Tiedemann and Scherrer, 2017). (3+1a) Concatenating the previous two sentences with the current source sentence. (3+1b) Concatenating the previous, the current and the next sentence in the source languages.

- MT-concat-target: (1t+1s+1) Concatenating the previously translated (target) sentence with the current source sentence.

- MT-concat-source-target: (2+2) Concatenating the previous with the current source sentence and translate into the previous and the current target sentence (Tiedemann and Scherrer, 2017). Only the second sentence in the translation will be kept for evaluation of the translation quality.

Extended context models only make sense with coherent training data. Therefore, we ran experiments only with the training data that contain translated documents, *i.e.* Europarl, NewsCommentary, Rapid and the back-translations of the German news from 2018. Hence, the baseline is lower than a sentence-level model on the complete data sets provided by WMT. Table 10 summarizes the results on the development test data (news 2018).

	BLEU news2018	
System	Shuffled	Coherent
Baseline	38.96	38.96
2+1	36.62	37.17
3+1a	33.90	34.30
3+1b	34.14	34.39
1t+1s+1	36.82	37.24
2+2	38.53	**39.08**

Table 10: Comparison of concatenation approaches for English–German document-level translation.

The results overall are rather disappointing. All but one of the concatenation models underperform and cannot beat the sentence-level baseline. Note that the concat-target model (1t+1s+1) even refers to an oracle experiment in which the reference

translation of the previous sentence is fed into the translation model for translating the current source sentence. As this is not very successful, we did not even try to run a proper evaluation with system output provided as target context during testing. Besides the shortcomings, we can nevertheless see a consistent pattern that the extended context models indeed pick up information from discourse. For all models we observe a gain of about half a BLEU point when comparing the shuffled to the non-shuffled versions of the test set. This is interesting and encourages us to study these models further in future work, possibly with different data sets, training procedures and slightly different architectures.

4.2 Hierarchical attention models

A number of approaches have been developed to utilize the attention mechanism to capture extended context for document-level translation. We experimented with the two following models:

- NMT-HAN: Sentence-level transformer model with a hierarchical attention network to capture the document-level context (Miculicich et al., 2018).
- selectAttn: Selective attention model for context-aware neural machine translation (Maruf et al., 2019).

For testing the selectAttn model, we used the same data with document-level information as we applied in the concatenation models. For NMT-HAN we had to use a smaller training set due to lack of resources and due to the implementation not supporting data shards. For NMT-HAN we used only Europarl, NewsCommentary and Rapid for training. Table 11 summarizes the results on the development test data. Both of the tested models need to be trained on sentence-level first, before tuning the document-level components.

Model	Sentence-level	Document-level
NMT-HAN	35.03	31.73
selectAttn	35.26	34.75

Table 11: Results (case-sensitive BLEU) of the hierarchical attention models on the coherent newstest 2018 dataset.

The architecture of the selective attention model is based on the general transformer model but with quite a different setup in terms of hyperparameters and dimensions of layer components etc. We applied the basic settings following the documentation of the software. In particular, the model includes 4 layers and 8 attention heads, and the dimensionality of the hidden layers is 512. We applied a sublayer and attention dropout of 0.1 and trained the sentence-level model for about 3.5 epochs. We selected monolingual source-side context for our experiments and hierarchical document attention with sparse softmax. Otherwise, we also apply the default parameters suggested in the documentation with respect to optimizers, learning rates and dropout. Unfortunately, the results do not look very promising as we can see in Table 11. The document-level model does not even reach the performance of the sentence-level model even though we trained until convergence on development data with patience of 10 reporting steps, which is quite disappointing. Overall, the scores are below the standard transformer models of the other experiments, and hence, we did not try to further optimize the results using that model.

For the NMT-HAN model we used the implementation of Miculicich et al. (2018) with the recommended hyperparameter values and settings. The system is based on the OpenNMT-py implementation of the transformer. The model includes 6 hidden layers on both the encoder and decoder side with a dimensionality of 512 and the multi-head attention has 8 attention heads. We applied a sublayer and attention dropout of 0.1. The target and source vocabulary size is 30K. We trained the sentence-level model for 20 epochs after which we further fine-tuned the encoder side hierarchical attention for 1 epoch and the joint encoder-decoder hierarchical attention for 1 epoch. The results for the NMT-HAN model are disappointing. The document-level model performs significantly worse than the sentence-level model.

5 Results from WMT 2019

Table 12 summarizes our results from the WMT 2019 news task. We list the official score from the submitted systems and post-WMT scores that come from models described above. For Finnish–English and English–Finnish, the submitted systems correspond to premature single models that did not converge yet. Our submitted English–German model is the ensemble of 9 models described in Section 3.2.

Language pair	Model	BLEU
English–German	submitted	41.4
	L2R+R2L	42.95
Finnish–English	submitted	26.7
	L2R+R2L	27.80
English–Finnish	submitted	20.8
	rule-based	8.9
	L2R+R2L	23.4

Table 12: Final results (case-sensitive BLEU scores) on the 2019 news test set; partially obtained after the deadline.

The ensemble results clearly outperform those results but were not ready in time. We are still below the best performing system from the official participants of this year's campaign but the final models perform in the top-range of all the three tasks. For English–Finnish, our final score would end up on a third place (12 submissions from 8 participants), for Finnish–English it would be the fourth-best participant (out of 9), and English–German fifth-best participant (out of 19 with 28 submissions).

6 Conclusions

In this paper, we presented our submission for the WMT 2019 news translation task in three language pairs: English–German, English–Finnish and Finnish–English.

For all the language pairs we spent considerable time on cleaning and filtering the training data, which resulted in a significant reduction of training examples without a negative impact on translation quality.

For English–German we focused both on sentence-level neural machine translation models as well as document-level models. For English–Finnish, our submissions consists of an NMT system as well as a rule-based system whereas the Finnish–English system is an NMT system. For the English–Finnish and Finnish–English language pairs, we compared the impact of different segmentation approaches. Our results show that the different segmentation approaches do not significantly impact BLEU scores. However, our experiments highlight the well-known fact that ensembling and domain adaptation have a significant positive impact on translation quality.

One surprising finding was that none of the document-level approaches really worked, with some even having a negative effect on translation quality.

Acknowledgments

The work in this paper was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113), and the MeMAD project funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 780069.

The authors gratefully acknowledge the support of the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence and project 270354/273457.

References

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393, Belgium, Brussels. Association for Computational Linguistics.

Arvi Hurskainen. 2018a. Compound nouns in English to Finnish machine translation. Technical Reports in Language Technology 32, University of Helsinki.

Arvi Hurskainen. 2018b. Direct and indirect questions in English to Finnish machine translation. Technical Reports in Language Technology 33, University of Helsinki.

Arvi Hurskainen. 2018c. Optimizing rules in English to Finnish machine translation. Technical Reports in Language Technology 34, University of Helsinki.

Arvi Hurskainen. 2018d. Participial phrases in English to Finnish machine translation. Technical Reports in Language Technology 35, University of Helsinki.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. 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. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki neural machine translation system. In *Proceedings of the Second Conference on Machine Translation*, pages 338–347, Copenhagen, Denmark. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Alexandra Birch Chris Callison-Burch Marcello Federico Nicola Bertoldi Brooke Cowan Wade Shen Christine Moran Richard Zens Chris Dyer Ondrej Bojar Alexandra Constantin Evan Herbst Philipp Koehn, Hieu Hoang. 2007. Moses: Open source toolkit for statistical machine translation. *CoRR*, Annual Meeting of the Association for Computational Linguistics (ACL).

Tommi A. Pirinen. 2015. Omorfi — free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Alessandro Raganato, Yves Scherrer, Tommi Nieminen, Arvi Hurskainen, and Jörg Tiedemann. 2018. The University of Helsinki submissions to the WMT18 news task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 488–495, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. Association for Computational Linguistics.

Vesa Siivola, Mathias Creutz, and Mikko Kurimo. 2007a. Morfessor and VariKN machine learning tools for speech and language technology. In *8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerp, Belgium, August 27-31, 2007*, pages 1549–1552. ISCA.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007b. On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Trans. Audio, Speech & Language Processing*, 15(5):1617–1624.

Felix Stahlberg, Adriá de Gispert, and Bill Byrne. 2018. The University of Cambridge's machine translation systems for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 508–516, Belgium, Brussels. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. 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.

B.2 MeMAD WMT corpus filtering task paper

The University of Helsinki Submission to the WMT19 Parallel Corpus Filtering Task

Raúl Vázquez, Umut Sulubacak and **Jörg Tiedemann**

University of Helsinki
{name.surname}@helsinki.fi

Abstract

This paper describes the University of Helsinki Language Technology group's participation in the WMT 2019 parallel corpus filtering task. Our scores were produced using a two-step strategy. First, we individually applied a series of filters to remove the *'bad'* quality sentences. Then, we produced scores for each sentence by weighting these features with a classification model. This methodology allowed us to build a simple and reliable system that is easily adaptable to other language pairs.

1 Introduction

Data-driven methodologies define the state of the art in a wide variety of language processing tasks. The availability of well-formed, clean data varies from language to language, and finding such data in sufficient amounts can prove challenging for some of the lower-resourced languages. In particular, the increasingly common neural machine translation systems are highly sensitive to the quality as well as the quantity of training data (Khayrallah and Koehn, 2018), which creates an impediment to achieving good-quality translations in a low-resource scenario.

The web is a massive resource for text data in a wide array of languages. However, it is costly to manually extract high-quality parallel samples from the web, and automatically-crawled datasets such as the ParaCrawl Corpus[1] are typically quite noisy. Designing automatic methods to select high-quality aligned samples from noisy parallel corpora can therefore make crawling the web a more viable option for compiling useful training data.

To emphasize this untapped potential, Koehn et al. (2018) proposed the Shared Task on Parallel Corpus Filtering as part of WMT in 2018. We participated in this year's task with three sets of quality scores. Each score is a different aggregation of a shared set of features, with each feature representing a local quality estimate focusing on a different aspect. Section 2 contains a brief discussion of this year's shared task. We present our scoring system in Section 3, discussing the filters we used for feature extraction in Section 3.2, and the aggregate scorers in Section 3.3. Finally, we report our contrastive results in Section 4.

2 Task Description

This year, the corpus filtering task organizers decided to pose the problem under more challenging conditions by focusing on low-resource scenarios, as opposed to previous year German–English (Koehn et al., 2018). In particular, two parallel corpora are to be scored for filtering: Nepali–English and Sinhala–English. The task for each participating team is to provide a quality score for each sentence pair in either or both of the corpora. The scores do not have to be meaningful, except that higher scores indicate better quality. The computed scores are then evaluated under four scenarios: training SMT and NMT systems, on samples of 5 million and 1 million words each, where the samples are obtained from the corresponding corpus using the quality scores.

Participants are provided with raw corpora to score, which were crawled using the ParaCrawl pipeline, and consist of 40.6 million (English) words for Nepali–English, and 59.6 million for Sinhala–English. Additionally, some parallel and monolingual corpora were provided for each language pair. We used the parallel datasets to train some of our scoring systems[2]. Some descriptive

[1] ParaCrawl can be downloaded from https://paracrawl.eu/

[2] En–Si: OpenSubtitles and GNOME/KDE/Ubuntu; En–Ne: Bible (two translations), Global Voices, Penn Treebank, GNOME/KDE/Ubuntu, and Nepali Dictionary.

corpus	lang. pair	sent. pairs	EN words
ParaCrawl	EN–NE	2.2M	40.6M
additional	EN–NE	543K	2.9M
ParaCrawl	EN–SI	3.4M	45.5M
additional	EN–SI	647K	3.7M

Table 1: Statistics on the ParaCrawl data and the used parallel data. Only English word counts reported.

statistics of the data we have used can be found in Table 1.

3 Scoring system

We first independently applied a series of filters to the data and computed relevant numerical features with them. We have previously corroborated the filters' effectiveness, since we have used them to clean the noisy datasets provided for this year's news translation task at WMT with satisfactory results. Then, we selected a cut-off value for each filter and trained a classifier over the features to compute a global score for each sentence pair, which we used to rank them.

3.1 Cleaning up the clean training data

Some of our filters require clean data for training. We observed that the provided parallel data still contained quite a lot of noise, and therefore, we applied some additional heuristic filters to clean it further. In particular, we used the following heuristics to remove pairs with characteristics that indicate likely problems in the data:

- Removing all sentence pairs with a length ratio above 3 between the source and the target.
- Removing pairs with very long sentences containing more than 100 words.
- Removing sentences with extremely long words, *i.e.* excluding all sentence pairs with words of 40 or more characters.
- Removing sentence pairs that include HTML or XML tags.
- Removing sentence pairs that include characters outside of the decoding table of Devanagari (for Nepalese) and Sinhala characters besides punctuation and whitespace.
- Removing sentence pairs that include Devanagari or Sinhala characters in English.

The procedure above discarded around 23% of the data for Nepali–English, and we kept around 440k parallel sentences from the original data. For Sinhala–English, we removed about 19% of the data and kept 522k sentence pairs for training.

3.2 Filters

Word alignment. Our first filter applies statistical word alignment models to rank sentence pairs. Word alignment models implement a straightforward way of estimating the likelihood of parallel sentences. In particular, IBM-style alignment models estimate the probability $p(f|a, e)$ of a foreign sentence f given an "emitted" sentence e and an alignment a between them.

We used *eflomal*[3] (Östling and Tiedemann, 2016) for word-level alignment, as it provides significant benefits. First, it is an efficient algorithm based on Gibbs sampling, as opposed to the slower expectation maximization methods commonly used for training. This method is thus able to train and align large quantities of data in a small amount of time. Second, this software allows us to load model priors, a feature we use to initialize the aligner with previously stored model parameters. This is handy for our filtering needs, as we can now train a model on clean parallel data and apply that model to estimate alignment probabilities of noisy data sets.

For obtaining model priors, we use the cleaned training data described above, tokenized with the generic tokenizer from the Moses toolkit (Koehn et al., 2007). We cut all words at 10 characters to improve statistics and training efficiency. With this, we train for both language pairs a Bayesian HMM alignment model with fertilities in both directions, and estimate the model priors from the symmetrized alignment. We then use those priors to run the alignment of the noisy datasets using only a single iteration of the final model to avoid a strong influence of the noisy data on alignment parameters. As it is intractable to estimate a fully normalized conditional probability of a sentence pair under the given higher-level word alignment model, *eflomal* estimates a score based on the maximum unnormalized log-probability of links in the last sampling iteration. In practice, this seems to work well, and we take that value to rank sentence pairs by their alignment quality.

[3]Software available from https://github.com/robertostling/eflomal

Language model filter. The second filter applies language models for source and target languages. In our approach, we opt for a combination of source and target language models, and focus on the comparison between scores coming from both models. The idea with this filter is to prefer sentence pairs for which the cross-entropy with the clean monolingual language models is low for both languages, and that the absolute difference between the cross-entropy of aligned sentences is low as well. The intuition is that both models should be roughly similarly surprised when observing sentences that are translations of each other. In order to make the values comparable, we trained our language models on parallel data sets.

As both training data sets are rather small, and as we aim for an efficient and cheap filter, we chose a traditional n-gram language model. To further avoid data sparseness and to improve comparability between source and target languages, we also base our language models on BPE-segmented texts (Sennrich et al., 2016) using a BPE model trained on the cleaned parallel data set with 37k merge operations per language. *VariKN* [4] (Siivola et al., 2007b,a) is the perfect toolkit for the purpose of estimating n-gram language models with subword units. It implements Kneser-Ney growing and revised Kneser-Ney pruning methods with the support of n-grams of varying size and the estimation of word likelihoods from text segmented into subword units. In our case, we set the maximum n-gram size to 20, and a pruning threshold of 0.002. Finally, we compute cross-entropies for each sentence in the noisy parallel training data, and store five values as potential features for filtering: the source and target language cross-entropy, $H(S, q_s)$ and $H(T, q_t)$, as well ad the average, max and absolute difference between them, i.e., $avg(H(S, q_s), H(T, q_t))$, $abs(H(S, q_s) - (T, q_t))$ and $max(H(S, q_s), H(T, q_t))$.

Language identifiers. A third filter applies off-the-shelf language identifiers. In particular, we use the Python interface of the Compact Language Detector[5] version 2 (CLD2) from the Google Chrome project, and the widely used `langid.py` package (Lui and Baldwin, 2012), to classify each sentence in the datasets.

We generate 4 features from these classifiers. For each language, we use the reliability score by CLD2 only if the predicted language was correct, and zero otherwise; and we use the detection probability of `langid.py` only if the language was classified correctly, and zero otherwise.

Character scores. Another simple filter computes the proportion of Devanagari, Sinhala and Latin–1 characters in Nepali, Sinhala and English sentences, respectively. For this computation, we ignore all whitespace and punctuation characters using common Unicode character classes.

Terminal punctuation. This heuristic filter generates a penalty score with respect to the co-occurrence of terminal punctuation marks ('.', '...', '?', '!') in a pair of sentences. In order to have a finer granularity than $\{0, 1\}$, we penalize both asymmetry (to catch many-to-one alignments) and large numbers of terminal punctuation (to cover very long sentences, URLs and code). For a given source and target sentence pair, we initialize a score as the absolute difference between source and target terminal punctuation counts. Then, we increment this score by the number of terminal punctuation beyond the first occurrence in both source and target sentences.

The intended effect is for the ideal sentence pair to contain either no terminal punctuation or a single terminal punctuation on either side ($score = 0$). In practice, many sentences are very far from the ideal ($score \gg 100$), and it is counter-intuitive to use a larger positive value to represent a higher penalty. To address both problems, we finally make the following update:

$$score = -log(score + 1)$$

Non-zero numerals. This filter assumes that numerals used to represent quantities and dates will be typically translated in the same format, and penalizes sentence pairs where numerals do not have a one-to-one correspondence or do not occur in the same sequence.

Sinhala uses the same Western Arabic numerals used in the Latin alphabet. Nepali uses Devanagari numerals, following the same decimal system as Western Arabic numerals. This filter takes that into account, and first converts those to digits between $[0, 9]$. After numeric normalization, the filter extracts sequences of numerals from each

[4] VariKN is available from https://vsiivola.github.io/variKN/

[5] The Python implementation of CLD2 is available at https://github.com/aboSamoor/pycld2

pair of sentences, preserving their relative order. Considering that a leading zero can be omitted in some numeric sequences such as in dates and numbered lists, the digit '0' is ignored. Finally, the score is calculated as a similarity measure between the extracted sequences in the range $[0, 1]$ using `SequenceMatcher.ratio()` from Python's `difflib`.

Clean-corpus filter Finally, we use the well-proven *clean-corpus-n* script from Moses to produce a binary feature augmented by a feature that marks sentences including HTML or XML tags.

All in all, we obtain 15 potential features from these filters. However, some of them are to be considered redundant and the information they provide is already encoded in some other variable. For instance, using the reliability score produced by CLD2 together with the prediction probability from `langid.py` would not provide crucial additional information to a model. Table 2 summarizes the filters we used to train our scoring models.

№	Feature	Definition
1	word-align	$\sim p(f\|a, e)$
2	lang-model	$H(S, q_s)$
3		$H(T, q_t)$
4	lang-id	src reliability score
5		tgt reliability score
6	char-score	English chars %
7		Ne/Si chars %
8	term-punct	penalty for asymmetric & excessive term. punct.
9	non-zero	similarity between non-zero digit seq.
10	clean-corpus	1, if kept 0, otherwise

Table 2: List of features extracted from the filters.

3.3 Scorers

We trained a logistic regression classifier and a random forest classifier to score each sentence pair using the features presented in Section 3.2. We trained three independent binary classifiers under the following settings:

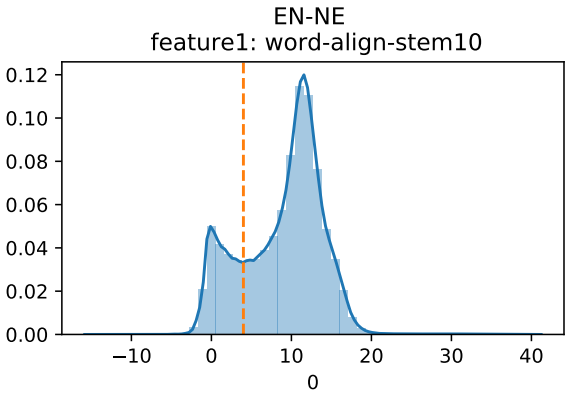

Figure 1: Distribution and cutoff value of feature 1 (word alignment) in the English–Nepali ParaCrawl corpus.

1. Applying all filters to the additional parallel corpora, and using filtered data as positive examples, and filtered-out data as negative examples.

2. Applying all filters to the corresponding ParaCrawl corpus, and using filtered data as positive examples, and a sample of 600k filtered-out examples as negative examples.

3. Applying all filters to both the ParaCrawl and the additional parallel corpora, and using these as positive examples, and a sample of 1M filtered-out examples as negative examples.

	lang. pair	RF AIC	RF BIC	LR AIC	LR BIC
PC	en-ne	17.8	-1.0e+7	-1.3	-2.5e+6
PC+BIC	en-ne	16.8	-1.1e+7	-0.9	-2.9e+6
PC	en-si	15.4	-9.4e+6	-1.5	-2.3e+6
PC+BIC	en-si	15.6	-1.1e+7	-1.4	-2.9e+6

Table 3: AIC and BIC obtained with random forest (RF) and logistic regression (LR) models. Comparison between the first chosen thresholds for ParaCrawl (PC) data and the model that optimizes the information criteria (PC+BIC).

For each filter under the first two scenarios, we adjusted thresholds based on score distributions, attempting to keep a balance between having restrictive thresholds that limited the amount of positive examples, and having lax thresholds

data	langpair	word-align	lang-model (*src*)	lang-model (*tgt*)	lang-id (*src*)	lang-id (*tgt*)	char-score (*%En*)	char-score (*%Ne/Si*)	term-punct	non-zero	clean-corpus
additional clean	ne-en	1	5	0	—	0	0	0	–2	0.5	0
ParaCrawl	ne-en	4	10	9	0	0	0	0	–2	0.5	0
ParaCrawl bestBIC	ne-en	—	—	—	0	0	0	0	–2	0.5	0
additional clean	ne-si	2	6	5	0	0	0	0	–1.5	0.5	0
ParaCrawl	ne-si	3	10	10	0	0	0	0	–1	0.5	0
ParaCrawl bestBIC	ne-si	—	10	10	0	0	0	0	–2	0.5	0

Table 4: Selected threshold value for each feature.

that classified many low-quality examples as positive. In some cases the score distributions were clearly bi-modal, making it easy to determine cutoff values (*e.g.* see Figure 1); while in other cases, we had to opt for a more empirical approach. For this reason, we have a second model that optimizes the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayes Information Criterion (BIC) (Schwarz et al., 1978) under scenario 2. This model was chosen from among 7 models trained with different reasonable combinations of the features. In Table 3, we compare the information criteria for both models. Finally, under the third scenario we chose to combine the data using the defined cutoff values from the previous two to include a significant amount of examples from both data sets.

Table 4 summarizes the threshold values used for each feature. After applying the filters, we kept 240k sentences (≈ 11% of the total) from the ParaCrawl EN–NE, 230k sentences (≈ 7%) from ParaCrawl EN–SI; 239k (≈ 44%) from the additional clean EN–NE data, and 231k (≈ 36%) from the additional clean EN–SI data. This means that, when combining them for scenario 3, we get 419k sentences (≈ 15%) for EN–NE, and 537k for EN–SI (≈ 14%). In order to avoid overfitting to the negative examples in scenarios 2 and 3, which vastly outnumber the positive ones, we performed stratified sampling of the negative examples where we selected 600K and 1M negative examples, respectively. We then randomly split the data into train (70%) and test (30%) sets.

4 Results

We report the accuracy on the test set achieved by the aforementioned models in Table 5. We do not report the accuracy of the random forest classifiers since they are all ≈ 99.99%. This is likely because the algorithm "cuts" through the variables in a similar way to how we chose the threshold values. For the same reason, they are unsuitable for the scoring task at hand. The output produced is a sharp classification that does not help rank the sentences. In contrast, the logarithmic regression model softens the output probabilities, emulating the creation of a composite index when used in combination without the threshold selection procedure.

	lang. pair	accuracy
additional	en-ne	78.21%
ParaCrawl	en-ne	96.09%
ParaCrawl+BIC	en-ne	96.46%
All data	en-ne	86.55%
additional	en-si	78.82%
ParaCrawl	en-si	95.26%
ParaCrawl+BIC	en-si	95.26%
All data	en-si	91.14%

Table 5: Accuracy values on the test data for the trained logistic regression models. Additional refers to the additional parallel clean data provided, ParaCrawl+BIC to the model that optimized the BIC, and All data to scenario 3.

In a final step, we also combined the score given by the regression model with two heuristic features that we deemed to be important for the ranking. One of them is the character score that we introduced earlier, which computes the proportion of language-specific characters in the string ignoring punctuation and whitespace. With this factor, we heavily penalize sentence pairs that contain large

portions of foreign text. The second factor is based on the heuristics that translated sentences should exhibit similar lengths in terms of characters. This feature is proven to be efficient for common sentence alignment algorithms, and hence, we add the character length ratio as another factor in the final score. For simplicity, we just multiply the three values without any extra weights to obtain the final ranking score. The system that applies those additional factors is marked with *char-length* in Table 6 with the SMT results on the development test set.

model	NE–EN	SI–EN
baseline	4.22	4.77
logreg	4.91	5.06
+char-length	4.82	5.32
bestBIC	4.63	4.91

Table 6: BLEU scores using SMT on 5 million sampled training examples. The *baseline* refers to the Zipporah model reported by the organizers of the shared task.

We only ran experiments with the provided SMT model. We do not present results from the NMT model, since we encountered complications while running the pre-processing script in the provided development pack for the task. We believe it might be due to character encoding and noise in the data. However, we did not further investigate the source of said problem. The SMT scores are listed in Table 6. We can see that we indeed outperform the baseline model, but the scores are still so low that we deem the resulting models to be essentially useless. The performance for our three attempts are rather similar, with the plain logistic regression model having a slight advantage, and a small improvement provided by the char-length filter for the case of Sinhala–English. For that reason, we selected that model as our final submission, with the plain logreg model as a contrastive run to be evaluated.

By inspecting the provided data we draw the conclusion that the low quality of the final MT models is mainly due to the overall poor quality of the data, rather than solely an issue of the scoring algorithms. The final results of the shared task suggest that it has not been possible to squeeze much more out of the data. As seen in Table 7, submissions for this year demonstrate a narrow range of scores, and our primary submissions rank above average despite their poor performance.

	model	1M	5M	10M
EN–NE	best	4.21	4.62	4.74
	UHel (1)	3.19	3.87	4.31
	average	3.03 ± 1.22	3.60 ± 1.12	3.96 ± 0.89
	UHel (2)	1.29	2.05	3.83
EN–SI	best	4.27	4.76	4.94
	UHel (1)	3.26	3.84	4.12
	average	3.00 ± 1.13	3.43 ± 1.09	3.92 ± 0.87
	UHel (2)	2.28	3.24	3.96

Table 7: An overview of the relative performance (in BLEU scores) of our (1) primary and (2) contrastive SMT models trained on 1, 5, and 10 million samples. The *best* and *average* rows represent the highest score and the mean ± standard deviation among this year's submissions, respectively.

5 Conclusions

In this paper, we presented our rescoring system for the WMT 2019 Shared Task on Parallel Corpus Filtering. Our system is based on contrastive scoring models using features extracted from different kinds of data-driven and heuristic filters. We used these models to assign quality scores to each sentence pair. This methodology allowed us to build a simple and reliable system that is easily adapted to other language pairs. The machine translation quality indeed improves, however, BLEU scores remain particularly low. This raises questions about the general quality of the data. More detailed analyses of the data sets seem to be necessary to draw further conclusions.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113).

as well as the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme (grant № 780069).

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages

74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Gideon Schwarz et al. 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. Association for Computational Linguistics.

Vesa Siivola, Mathias Creutz, and Mikko Kurimo. 2007a. Morfessor and VariKN machine learning tools for speech and language technology. In *8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Antwerp, Belgium, August 27-31, 2007*, pages 1549–1552. ISCA.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007b. On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Trans. Audio, Speech & Language Processing*, 15(5):1617–1624.

B.3 MeMAD DiscoMT paper

Analysing concatenation approaches to document-level NMT in two different domains

Yves Scherrer[1] **Jörg Tiedemann**[1] **Sharid Loáiciga**[2]
[1]Department of Digital Humanities, University of Helsinki
[2]CLASP, Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg
yves.scherrer@helsinki.fi jorg.tiedemann@helsinki.fi
sharid.loaiciga@gu.se

Abstract

In this paper, we investigate how different aspects of discourse context affect the performance of recent neural MT systems. We describe two popular datasets covering news and movie subtitles and we provide a thorough analysis of the distribution of various document-level features in their domains. Furthermore, we train a set of context-aware MT models on both datasets and propose a comparative evaluation scheme that contrasts coherent context with artificially scrambled documents and absent context, arguing that the impact of discourse-aware MT models will become visible in this way. Our results show that the models are indeed affected by the manipulation of the test data, providing a different view on document-level translation quality than absolute sentence-level scores.

1 Introduction

Shortly after the change of paradigm in Machine Translation (MT) from statistical to neural architectures, the interest in discourse phenomena flourished again. This is not by chance, as neural models can embed larger text spans into contextual representations and can be set up to learn relevant features from the raw data to produce better translations.

It is still unclear though how the impact of discourse on MT quality should be evaluated and analyzed. On one side, it is difficult to pinpoint particular contextual features that neural MT (NMT) models are picking up. On the other, it is difficult to judge good translations purely in terms of discourse features. In this paper, we investigate the discourse-related biases in data. Our contributions are twofold:

- we provide a thorough analysis of two popular machine translation datasets in terms of document-level features,
- we train different context-aware MT models (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Maruf et al., 2019; Junczys-Dowmunt, 2019) on the two datasets and evaluate them using a comparative setup with artificially scrambled data.

As discourse properties of the data, we consider pronouns and coreference chains, connectives, and negation. For the evaluation of translation quality and the influence of document-level context, we contrast context-aware models at test time with (1) clean coherent text, (2) incoherent input and (3) zero-context input.[1] For the second type, we scramble sentences and insert document boundaries at arbitrary positions in the test data. For the third approach, we add document boundaries after each test instance. This setup provides a cheap way of testing the influence of contextual information on translation performance that can be measured in common ways, for example, facilitating automatic evaluation metrics such as BLEU or METEOR.

2 Related work

2.1 Discourse

Research about discourse and MT has shifted from explicitly enhancing systems with discourse knowledge to evaluating how much the systems have learned specific discourse features through different resources, test suites being a popular one (cf. Sim Smith, 2017; Popescu-Belis, 2019). Throughout, however, particular discourse phenomena are consistently targeted, as they are indeed indicators of globally good, cohesive and coherent texts. Pronouns (Hardmeier and Federico, 2010; Guillou, 2012; Hardmeier et al., 2013;

[1] *Context* here refers to text outside of the sentence to be translated.

Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 51–61
Hong Kong, China, November 3, 2019.

Guillou and Hardmeier, 2016; Müller et al., 2018; Guillou et al., 2018) have been largely at the center of attention, and more recently the translation of pronouns in the context of their coreferential chains has been looked at (Lapshinova-Koltunski and Hardmeier, 2017; Voita et al., 2018; Lapshinova-Koltunski et al., 2019). Other devices studied are verbal tenses (Gong et al., 2012; Loáiciga et al., 2014; Ramm and Fraser, 2016) and connectives (Meyer et al., 2012; Meyer and Popescu-Belis, 2012), although not using neural models. Motivated by approximating the ability of systems to grasp more abstract properties related to coherence, ambiguous words have also been targeted (Rios Gonzales et al., 2017; Bawden et al., 2018; Rios et al., 2018), as well as ellipsis (Voita et al., 2019). Last, negation (Fancellu and Webber, 2015) is a rather understudied phenomenon, but like pronouns and their antecedents, the scope of the negation can be in a different sentence.

In this paper we investigate these features in the training data and assess translation using standard automatic metrics and a data scrambling strategy.

2.2 Context-aware NMT

Tiedemann and Scherrer (2017) present a simple approach to context-aware NMT: instead of training the model on pairs of single source and target sentences, they add sentences from the left context to the sentence to be translated, either only on the source side or both on source and target sides. These models are evaluated on a German–English corpus extracted from OpenSubtitles, and the best results are obtained with two source sentences and one target sentence. Agrawal et al. (2018) extend these experiments by considering additional contexts. They evaluate their work on the IWSLT 2017 dataset for English–Italian, which consists of transcripts of TED talks.

In 2019, the WMT conference featured for the first time a document-level translation task for English–German (Barrault et al., 2019). One of the best-performing systems (Junczys-Dowmunt, 2019) is based on a similar idea: all sentences of a document are concatenated and translated as a whole. Documents whose length exceeds the maximum sequence length defined by the model are simply split.

The approaches outlined above, which we refer to as "concatenation models", do not require any change to the NMT model architecture. Other recent work explores the feasability of extending NMT models to make them context-aware. A common approach is to use additional encoders for the context sentence(s) with a modified attention mechanism (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018). Another technique (Miculicich et al., 2018; Maruf et al., 2019) explores the integration of context through a hierarchical architecture which models the contextual information in a structured manner using word-level and sentence-level abstractions.

The different models have been evaluated on different language pairs and different datasets. In this paper, we focus on a single language pair, English–German (in both directions), and on two textual domains: news translation and movie subtitles translation. For the news translation task (denoted as *WMT*) we rely on the established setup of WMT 2019[2] with the Newstest2018 data as our dedicated test set. For the movie subtitles (referred to as *OST*), we use data from the OpenSubtitles corpus released on OPUS[3] with our own split into training, development and test data. More details about the data and our setup will be given in the following section.

3 Two datasets for English–German document-level translation

Different text genres and types exhibit different types of discourse-level properties. The choice of training corpus therefore determines what features a NMT model can potentially learn, and the choice of test corpus determines which features can be reliably evaluated. Our experiments are based on two datasets that cover the same language pair, but very different textual characteristics.

The **OST** dataset is built from the English–German part of the publicly available OpenSubtitles2016 corpus (Lison and Tiedemann, 2016). Of the 16,910 movies and TV series in the collection, 16,510 are used for training, and 4 each are held out for development and testing purposes. Each movie is considered a single document. It corresponds to the dataset used in Tiedemann and Scherrer (2017). General properties of this dataset can be found in Table 1.

The **WMT** dataset comprises the subset of corpora allowed at the WMT 2019 news translation

[2]See http://www.statmt.org/wmt19/translation-task.html.

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

Corpus	Documents	Sentences	Sents/Doc	Tokens DE	Tokens EN	Tokens/Sent
OST Train	16,510	13,544k	820	104,447k	111,729k	8.0
OST Valid	4	5k	1249	41k	43k	8.4
OST Test	4	5k	1249	38k	47k	8.4
WMT Train	583,358	12,690k	22	259,384k	276,401k	21.1
WMT Valid	236	5k	22	106k	111k	21.1
WMT Test	122	3k	25	64k	68k	21.9

Table 1: General characteristics of the two datasets. Tokens/Sent values are averaged over the DE and EN tokens.

task which contains document boundaries. The training set includes parallel data from the Europarl v9, NewsCommentary v14, and Rapid2019 collections. We select the Newstest2015 and Newstest2016 corpora as our validation set and the Newstest2018 corpus as our test set. General properties of this dataset can be found in Table 1.

Table 1 shows that the two datasets are comparable in terms of sentence numbers.[4] However, the documents in OST are up to 50 times larger than those in WMT (cf. column *Sents/Doc*). On the other hand, WMT sentences are more than twice as long than OST sentences (cf. column *Tokens/Sent*), which is in line with our expectations.

A third dataset based on transcripts of TED talks (Cettolo et al., 2012), has also been used for document-level translation (Agrawal et al., 2018). We do not consider this dataset for training due to its smaller size, but use the PROTEST test suite, which is based on this corpus, for evaluation (Guillou and Hardmeier, 2016; Guillou et al., 2018).

3.1 Discourse-level properties

In recent literature, various linguistic features have been identified to contribute to document-level coherence and cohesion. In this section, we assess the two datasets in order to estimate their suitability and difficulty for document-level translation. We investigate the following phenomena:

Pronouns: We first extract a list of pronouns per language by tagging the training corpora with SpaCy[5], extracting the tokens labeled as PRON and manually cleaning the resulting list (cf. Table 7). Then, the frequency of pronouns is computed independently for English and German.

The results in Table 2 show that about every 10th word of the OST corpus is a pronoun, whereas pronouns are three to four times rarer in the WMT corpus.[6] This divergence is to be expected, as OST consists mainly of dialogues.

Not all pronouns are intrinsically hard to translate. Therefore, we also examine how many **ambiguous pronouns** occur in the corpora. To this end, the English and German corpora are word-aligned using Eflomal (Östling and Tiedemann, 2016) and for each source pronoun (as defined in the list extracted previously), the target pronouns are retrieved. If this list contains at least two words totalling each at least 10% of occurrences, we consider the source pronoun as ambiguous (cf. Table 7). This feature is computed separately for both translation directions.

On average, about half of the pronoun occurrences are ambiguous, with most ambiguities concerning case (e.g. *me* translating both to accusative *mich* and dative *mir*). The English pronouns in the OST dataset deviate from this tendency, mainly because of the prevalence of *you*: this pronoun is ambiguous both in terms of number and politeness and can be translated as *du*, *ihr*, or *Sie* (see also Sennrich et al., 2016).

Connectives: As part of their *Accuracy of Connective Translation* metric, Hajlaoui and Popescu-Belis (2013) provide a list of eight ambiguous English connectives and their German translations. We count the number of sentence pairs that contain both an English connective and one of its German translations, regardless of its associated sense.

Ambiguous connectives show an inverse frequency distribution compared to pronouns: they are about ten times as frequent in WMT than in OST. This divergence can again be attributed to genre differences.

[4] By sentences, we mean the lines obtained by the sentence alignment process.

[5] spacy.io

[6] The numbers for German are higher because the pronoun list contains more relative and demonstrative pronouns than the English one, as a result of annotation differences in the SpaCy training corpora.

Corpus	Pronouns		Ambiguous pronouns		Ambiguous connectives	Negations		Negation discrep.	Coreference chains		Cross-sent. pron. coref.	
	DE	EN	DE	EN	DE–EN	DE	EN	DE–EN	DE	EN	DE	EN
OST Train	106.0	97.0	44.1	71.1	5.0	151.6	162.8	57.1	290.5	148.3	67.2	44.5
OST Valid	104.7	92.7	49.9	73.0	6.2	165.5	171.5	65.6	346.1	167.5	70.2	46.4
OST Test	101.1	99.3	53.0	69.7	5.8	148.9	191.9	75.0	292.5	178.8	66.8	46.9
WMT Train	36.1	20.0	20.1	13.5	60.2	176.1	176.2	19.6	670.3	495.3	91.9	80.6
WMT Valid	44.2	29.6	24.6	20.8	62.5	182.1	177.2	23.8	693.5	544.2	111.5	97.6
WMT Test	44.0	25.8	25.9	20.0	58.3	167.4	169.1	18.3	726.8	535.0	115.4	99.7
	per thousand tokens				per thousand lines							

Table 2: Discourse-level features in the OST and WMT datasets. Coreference values were computed on a subset of the training corpora.

Negations: We establish a list of sentential and nominal negation words for both languages (cf. Table 7) and count the number of sentences that contain at least one negation word. We also count **negation discrepancies**, i.e. aligned sentence pairs where a negation was identified in one language but not in the other.

While the overall frequencies of negations are similar in both corpora, there are significantly more discrepancies in the OST dataset. These can be ascribed to two factors: free translation (a negation can be paraphrased with expressions such as *fail to*, *doubt if*, etc.), and sentence alignment errors.

Coreference chains: We assume that a large amount of pronouns, connectives and negations do not require access to large contexts for their correct translation, either because they are unambiguous or because the current sentence is sufficient for their disambiguation. To corroborate this assumption, we annotate the English corpora with the Stanford CoreNLP coreference resolver (Manning et al., 2014; Clark and Manning, 2016) and the German corpora with the CorZu coreference resolver (Tuggener, 2016).[7]

We first report the numbers of coreference chains identified by the resolvers. These numbers are hard to compare across languages due to different performance levels of the two resolvers, and translationese factors such as explicitation. However, they confirm the intuition that news text contains more referring entities than movie dialogues.[8]

Second, we count **cross-sentential pronominal coreference chains**, i.e. chains that span at least two sentences, contain at least one third-person pronoun and at least two different mention strings. The results suggest that about every 10th line of the WMT dataset and about every 20th line of the OST dataset contains a pronoun that requires access to the context for its correct translation. Given the overall training data sizes, NMT models should thus be able to pick up this signal.

Overall, the examined discourse-level features show consistent patterns across the training, validation and test sets. This was not necessarily expected for the WMT corpus, whose training set stems from a wide variety of sources.[9]

Three other discourse-level features could have been analyzed as well: We did not include verbal tenses, as we do not expect them to be particularly problematic for the German–English language pair. Likewise, we did not include measures for lexical consistency (Carpuat and Simard, 2012), as this was already reported to be handled well in SMT. Finally, we did not include ellipsis (Voita et al., 2019) as we found it difficult to detect and not very relevant for German.

4 Context-aware MT models

In this paper, our main focus lies on concatenation models as one of the most straightforward and successful approaches to document-level NMT. We train various concatenation models on both datasets and for both translation directions in order to perform a systematic study on this setup.

[7]Due to slow performance, we could only analyze 13% of the English OST, 5% of the English WMT and 5% of the German WMT training sets. We nevertheless believe that the reported proportions are representative of the entire dataset.

[8]Note also that the WMT dataset may benefit from higher recall as the coreference resolution pipelines are typically trained on newswire data.

[9]For the MT training, we shuffle the datasets keeping documents and document boundaries intact.

Inspired by Agrawal et al. (2018), we name the configurations according to the following schema:

$$i\textbf{Prev + Curr + } j\textbf{Next} \rightarrow k\textbf{Prev + Curr}$$

where i denotes the number of previous sentences on the source side, j the number of following sentences on the source side, and k the number of previous sentences on the target side. In all models, only the current sentence is evaluated. The following configurations are tested:

- Curr → Curr (baseline)
- 1Prev + Curr → Curr
- 1Prev + Curr + 1Next → Curr
- 2Prev + Curr → Curr
- 1Prev + Curr → 1Prev + Curr
- 1Prev + Curr + 1Next → 1Prev + Curr

Several discourse-level properties, among which most prominently pronoun gender, also depend on the previously generated output in the target language. Therefore, we also include an oracle variant where we the reference translation of the previous sentence (instead of its source) is fed to the system:

- 1PrevTarget + Curr → Curr

Furthermore, we also train fixed window models as in Junczys-Dowmunt (2019):

- 100T → 100T: A model that sees chunks of at most 100 tokens (after subword encoding) on either source and target side.
- 250T → 250T: A model that sees chunks of at most 250 tokens (after subword encoding) on either source and target side.

Note that these chunks are not produced using a sliding window but rather break documents at arbitrary positions unless they are less than the maximum size in length. We adopt the same annotation scheme as proposed in the original approach, marking segment and document boundaries with special symbols for document-internal breaks and continuations. We never break sentences from the original alignment into pieces, which would negatively affect the model and complicate the alignment of training examples.

The chosen chunk lengths seem very small, especially when considering subword units. Table 3 lists some basic statistics that demonstrate the effect of the chunking approach. We can see that even 100-token windows create reasonably large units that combine context beyond sentence boundaries. For the WMT dataset with larger sentences, we observe an average of almost 3 joined segments per chunk. For the subtitle data, the situation is much more extreme: most segments are very short and a 100-token window corresponds to about 10 segments. Hence, this approach yields a substantial increase of contextual information compared to the baseline.

Window size	Chunks	Sents/chunk
OST training data:		
100 tokens	1 282 985	10.6
250 tokens	496 207	27.3
WMT training data:		
100 tokens	4 286 535	3.0
250 tokens	1 729 601	7.3

Table 3: Basic statistics of fixed-size windows data.

Junczys-Dowmunt (2019) suggested to use even larger chunks, but that did not seem to work well in our current settings. Already the second model with a maximum of 250 tokens did not converge to any reasonable result when trained from scratch. We tried to address this problem by initialising the larger model with a pre-trained 100-token model but this approach did not lead to satisfactory results either. Therefore, we exclude all models larger than 100 tokens from our discussions below.

All models are based on the standard Transformer architecture and were trained with MarianNMT (Junczys-Dowmunt et al., 2018). For the WMT EN→DE models, we added 10.3M lines of backtranslations. These backtranslations consisted of German news documents (News2018) translated to English with a sentence-level model; document boundaries were kept intact. We did not include backtranslations for the opposite translation direction to investigate their impact on discourse-level translation.

Our experiments with recently proposed hierarchical attention networks for document-level NMT, in particular Miculicich et al. (2018) and Maruf et al. (2019), either underperformed or could not cope with the data sizes and document lengths of our training sets. For comparison, we nevertheless report results of a selective attention (Maruf et al., 2019) model for the WMT

EN→DE task. This model has to be trained in a two-step procedure: (1) a standard sentence-level model is trained on all the training data and, (2) a document-level model is trained on top of the sentence-level model that adds the inter-sentential information from the surrounding context using the attentive connections of the extended network. We focused on source-side attention for the wider context and did not explore further setups due to computational costs and unsatisfactory baseline results. Otherwise, we use the standard settings recommended in the released software.

5 Evaluation

Each system is evaluated on the respective test set using the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) metrics. In particular, we evaluate each of them on three variants of the test set:

Consistent context: the context sentences of the test set are appended in their natural order, as they appear in the data.

Inconsistent context: the test set is shuffled such that the context sentences are random.

No context: each sentence of the test set is considered its own document, so no contextual information is made available.

This setup allows us to check whether observed improvements are due to the additional context or to other factors.[10] A good context-aware system should perform best with consistent context and worst with inconsistent context.

Note that the concatenation models need some special treatment at test time. The sliding window approaches need to be post-processed in order to remove non-relevant parts of the translation in all cases where we train models with extended target language content. For simplicity, we rely on the segment separation tokens that are produced in translation similar to the ones seen during training. We have found this approach to be very robust, in the sense that the models reliably learn to place them at appropriate positions.

For the non-sliding window approaches with fixed maximum size, sentence splitting is not as straightforward and requires some additional treatment. Segments are also separated by separation tokens but we realized that they do not necessarily match with the segment boundaries in the reference data even though the original paper suggests that this should be rather stable (Junczys-Dowmunt, 2019). This is especially fatal if the number of segments does not match. Therefore, we apply standard sentence alignment based on length-correlation and lexical matches using hunalign (Varga et al., 2005) to link the system output to the reference translations. The reported results from the fixed-size models are based on this approach.

5.1 Generic translation metrics

We report BLEU and METEOR scores for all our experiments in Tables 4 and 5. The results and significance tests were computed using *MultEval* (Clark et al., 2011).

By and large, the concatenation models are able to exploit contextual information: BLEU as well as METEOR scores decrease by statistically significant amounts if the context is inconsistent or absent. However, it is difficult to distinguish a winning configuration. In particular, the system that obtains the highest absolute scores is not necessarily the one that learns most from contextual information. The *1Prev+Curr → 1Prev+Curr* system obtains the highest absolute scores among sliding window systems in all four tasks, but is not particularly affected by context inconsistencies. On the other hand, the system using target-language data is most perturbed when context is inconsistent or absent, at least for the OST dataset.[11] It seems therefore that target-language context is as least as important as source-language context. Comparative numbers on the WMT dataset are all very similar, making it hard to draw conclusions.

The 100T fixed-window models perform competitively in terms of absolute scores, compared to the sliding window approaches, despite the alignment problems mentioned above.[12] The compar-

[10] For example, the *1Prev + Curr → Curr* system sees each source sentence twice as often as the *Curr → Curr* system, which might affect general model performance without necessarily improving context awareness.

[11] Note however that we feed the reference instead of the system output at test time for efficiency reasons. Therefore, the numbers cannot be directly compared directly with the other systems, which do not have access to this oracle-type information.

[12] Due to realignment, the number of sentences in the test set varies slightly, which prevents us from computing significance scores. Therefore, the absence of the significance marker * on the *100T → 100T* result lines does not mean that

Dataset:	OST EN → DE						WMT EN → DE					
Context:	Consistent		Incons. (Δ)		None (Δ)		Consistent		Incons. (Δ)		None (Δ)	
System	B	M	B	M	B	M	B	M	B	M	B	M
Curr → Curr (baseline)	21.7	42.6	0.0	0.0	0.0	0.0	39.3	56.9	0.0	0.0	0.0	0.0
1Prev+Curr → Curr	20.9	41.6	-0.3*	-0.5*	-0.2	-0.2	37.6	55.3	-0.5*	-0.3*	-0.2	-0.4*
1Prev+Curr+1Next → Curr	20.1	40.8	-1.0*	-1.2*	-0.6*	-0.5*	34.7	52.3	-0.4*	-0.4*	-0.5*	-0.4*
2Prev+Curr → Curr	20.3	40.4	-0.6*	-0.8*	-0.8*	-0.4*	34.9	53.1	-0.3*	-0.3*	-0.4*	-0.4*
1Prev+Curr → 1Prev+Curr	22.5	43.2	-0.7*	-0.7*	-0.3*	-0.5*	**39.6**	**57.3**	-0.5*	-0.4*	-0.2	-0.3*
1Prev+Curr+1Next → 1Prev+Curr	21.5	42.8	-0.5*	-1.0*	-0.1	-0.6*	38.5	56.0	**-0.8***	**-0.6***	-0.6*	-0.6*
1PrevTarget+Curr → Curr	22.0	42.5	-1.4*	-1.5*	**-1.3***	-1.3*	37.7	55.6	-0.4*	-0.3*	**-0.7***	**-0.7***
100T → 100T	**22.9**	**44.4**	**-1.9**	**-1.9**	-0.5	**-1.8**	39.0	57.2	-0.4	-0.5	0.0	**-0.7**
Selective attention	–	–	–	–	–	–	34.8	53.0	0.0	0.0	-0.2	-0.2

Table 4: BLEU (B) and METEOR (M) scores for EN → DE translation. Absolute scores are reported for the Consistent setting, whereas differences (relative to Consistent) are reported for the Inconsistent and None settings. Statistical significance at $p < 0.05$, obtained by bootstrap resampling, is marked with *.

Dataset:	OST DE → EN						WMT DE → EN					
Context:	Consistent		Incons. (Δ)		None (Δ)		Consistent		Incons. (Δ)		None (Δ)	
System	B	M	B	M	B	M	B	M	B	M	B	M
Curr → Curr (baseline)	27.4	27.6	0.0	0.0	0.0	0.0	34.9	**34.9**	0.0	0.0	0.0	0.0
1Prev+Curr → Curr	26.7	26.8	-0.4*	-0.3*	-0.3*	-0.1*	31.6	32.3	-0.3	0.0	-0.8*	-0.5*
1Prev+Curr+1Next → Curr	24.7	25.5	-0.1	-0.1	-0.3*	0.0	23.0	26.5	-0.1	0.0	**-2.2***	-0.3*
2Prev+Curr → Curr	26.0	26.3	-0.7*	-0.3*	-0.6*	-0.1*	22.0	26.1	-0.1	0.0	-1.3*	**-0.8***
1Prev+Curr → 1Prev+Curr	27.5	27.7	-0.3*	-0.2*	-0.4*	-0.2*	**35.0**	**34.9**	**-0.4***	0.0	-0.9*	-0.5*
1Prev+Curr+1Next → 1Prev+Curr	20.7	24.3	-0.1	0.0	+3.3*	+0.6*	31.2	32.4	-0.3*	**-0.2***	-1.5*	-0.6*
1PrevTarget+Curr → Curr	26.9	27.0	-1.0*	-0.7*	-1.0*	-0.6*	32.7	33.2	-0.3	0.0	-1.1*	-0.5*
100T → 100T	**29.3**	**28.8**	**-1.6**	**-1.0**	**-2.2**	**-1.3**	34.7	**34.9**	+0.1	+0.1	-0.7	-0.3

Table 5: BLEU (B) and METEOR (M) scores for DE → EN translation.

	anaphoric								event	pleonastic	
	it					they		it/they	it	it	
	intra		inter		intra	inter	sing.	group			
	subj.	non-subj.	subj.	non-subj.							**Total**
Examples:	*25*	*25*	*25*	*25*	*10*	*10*	*5*	*15*	*30*	*30*	***200***
OST Curr → Curr	9	7	6	7	5	3	1	5	20	28	91
OST 1Prev + Curr → 1Prev + Curr	10	6	12	9	5	6	1	2	24	25	100
WMT Curr → Curr	14	12	9	10	5	4	0	8	20	26	108
WMT 1Prev + Curr → 1Prev + Curr	9	11	13	12	5	5	1	5	19	28	108

Table 6: Absolute numbers of PROTEST EN → DE pronoun translations evaluated semi-automatically as correct.

DE Pronouns:	ich, es, das, wir, sich, Sie, er, du, sie, die, was, mir, mich, uns, der, man, dich, ihn, dir, dies, ihm, ihr, wer, 's, Ihnen, dem, denen, euch, ihnen, den, Ihr, diese, dessen, deren, einen, dieser, wen, welche, einem, wem, dieses, jene, diesen, dasselbe, welches, einander
Ambiguous:	Sie, den, denen, der, die, diese, dieser, ihm, ihn, ihnen, ihr, man, mich, mir, sich, sie, uns
EN Pronouns:	I, you, it, we, he, what, me, they, who, she, him, them, us, her, himself, itself, themselves, one, yourself, myself, whom, ourselves, i, 'em, herself, mine, yours, ya
Ambiguous:	her, him, it, me, myself, one, she, them, they, us, who, whom, you, yourself
EN Connectives:	although, even though, since, though, meanwhile, while, yet, however
DE Negations:	nicht, nie, niemand, nichts, nirgends, nirgendwo, kein, weder
EN Negations:	no, not, never, nobody, noone, no-one, nothing, nowhere, none, neither, nor

Table 7: List of words and lemmas used to detect discourse-level properties.

ison between consistent, inconsistent and absent context reveals a clear difference between the two datasets: For WMT, the results are almost the same for the three scenarios. This can be attributed to the longer sentences in the WMT test set, which makes the 100 token window performing similar to the one without extended context, as discussed in section 4. In contrast, for the subtitle data, we see notable performance drops when disturbing the model with random or absent context. In this dataset, segments are shorter and 100-token windows substantially increase the context that is available for translation (there are 9.68 sentences per chunk on average).

The selective attention model yields absolute scores with consistent context that are not competitive and barely beat the baseline. It also seems to fail to pick up relevant information from the wider document context, as it obtains almost identical results with inconsistent and absent context.

The WMT EN → DE models have seen back-translations during training but the DE → EN models have not. The results suggest that the additional data helps the models distinguishing consistent from inconsistent input, but further tests will be required to corroborate this hypothesis.

The WMT dataset has shorter documents and longer sentences with more complex discourse-level features. Although this may indicate that it is a more challenging dataset for our models, the performances seem very similar across systems, and it is hard to discriminate informative patterns. However, the inconsistent setting appears to be affected by genre, with none or very small differences with the WMT data, suggesting that the longer sentences are more self-contained in terms of discourse features and that systems effectively pick this signal up. In this same sense (and counterintuitively), the differences between inconsistent and none seem to suggest that as long as the system has access to big enough window, the order in which the document is fed is less important.

5.2 Test suite metrics

Discourse-specific metrics such as Guzmán et al. (2014) would be welcome to assess the translation quality on specific discourse-level features such as those discussed in Section 3.1. However, they have the disadvantage of relying on a discourse parser, which we do not have for German. At least, we are able to evaluate the quality of pronoun translation thanks to the existence of two test suites for English–German pronoun translation: **PROTEST** (Guillou and Hardmeier, 2016; Guillou et al., 2018) is based on TED talks transcripts. These consist of planned speech documents, therefore the genre is somewhere in the middle between news text and dialog. **ContraPro** (Müller et al., 2018) uses material from OpenSubtitles. Due to the overlap of the ContraPro data and our OST training set, we do not use this test suite.

Table 6 reports PROTEST results for two selected systems, the *Curr* → *Curr* baseline and the best-performing variable-window concatenation model *1Prev+Curr* → *1Prev+Curr*. The results draw on a semi-automatic evaluation scheme, where pronouns are accepted as correct if they match the reference and the remaining pronouns are evaluated by hand. The manual evaluation was done by one of the authors.[13] Overall recall of all systems is around 50%, and the differences between systems are quite small.

It can be seen that the models trained on the news dataset obtain higher recall. This confirms our observation in Section 3.1 that the WMT dataset contains higher numbers of coreference chains and cross-sentence pronominal coreference. The context-aware models show small improvements only in the OST dataset. Crucially, the context-aware models show consistently higher numbers in the category of inter-sentential anaphoric pronouns, one of the categories where the previous sentence context is indeed expected to help most. However, most observed differences may not be statistically significant.

The PROTEST evaluation confirms the findings of the WMT18 evaluation (Guillou et al., 2018). In both of these evaluations the *pleonastic* and *event* categories are the least problematic. *Intra-* and *inter-sentential* pronouns are somewhat in the middle but remain difficult, while cases where the anaphor and the antecedent mismatch in features (*they-singular, it/they group*) are very poorly handled.

6 Conclusion

We have presented two English–German document-level translation datasets and shown that they represent different text genres with

the differences are not significant.

[13] We used the provided tool described in Hardmeier and Guillou (2016).

different distributions of discourse-level features. The context-aware NMT models on these datasets show performance differences that are to some extent indicative of the underlying textual characteristics: the longer sentences in the news dataset make it harder to find differences between training configurations or evaluation setups. Fixed-window approaches show surprisingly good results on the movie subtitles dataset, but the impact of the realignment process remains to be investigated further.

The general performance of a document-level MT system can be assessed by testing translation quality with consistent and artificially scrambled context. Models that are able to learn relevant discourse features will be affected if the context is incoherent or absent. Our results show that this test provides a complementary view on the systems' performances.

Our study further suggests that the connections between discourse features and MT results should be analyzed more thoroughly. The detailed breakdown of the distribution of discourse-level properties could be a first step towards the compilation of property-specific test sets.

Automatic measures can be complemented with manual assessment of the outcome from the different test scenarios, which further reveals the effect of discourse features available to the system. We show that pronoun test suites such as PROTEST are a good start for this assessment, although multilingual coverage remains a problem for a systematic evaluation of this kind.

Acknowledgements

The work in this paper was supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113), and the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 780069.

The authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. 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.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussá, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*, pages 128–188, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. 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. Association for Computational Linguistics.

Marine Carpuat and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada. Association for Computational Linguistics.

Mauro Cettolo, Girardi Christian, and Federico Marcello. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of the European Association for Machine Translation*, page 261–268.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: Induction, search and model errors. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 21–29, Denver, Colorado, USA. Association for Computational Linguistics.

Zhengxian Gong, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012. 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*, EMNLP-CoNLL 2012, pages 276–285, Jeju Island, Korea. Association for Computational Linguistics.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, University of the Aegean, Samos, Greece.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, IWSLT 2010, pages 283–289, Paris, France.

Christian Hardmeier and Liane Guillou. 2016. A graphical pronoun analysis tool for the PROTEST pronoun evaluation test suite. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 318–330.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. 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. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? In *arXiv preprint, arXiv:1704.05135*.

Marcin Junczys-Dowmunt. 2019. Microsoft Translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 424–432, Florence, Italy. Association for Computational Linguistics.

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

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in English-German translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. Cross-lingual incongruences in the annotation of coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 26–34, Minneapolis, USA. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Sharid Loáiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. 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. Association for Computational Linguistics.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Workshop on Hybrid Approaches to Machine Translation at EACL 2012*, HyTra, pages 129—138, Avignon, France.

Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*, AMTA 2012.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. 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. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. 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, Belgium, Brussels. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

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

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations.

Anita Ramm and Alexander Fraser. 2016. Modeling verbal inflection for English to German SMT. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 21–31, Berlin, Germany. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. 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. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. 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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

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

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

Don Tuggener. 2016. *Incremental coreference resolution for German*. Ph.D. thesis, University of Zürich.

Daniel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. 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. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. 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. Association for Computational Linguistics.