A Review of Discourse-level Machine Translation

Xiaojun Zhang
Xi'an Jiaotong-Liverpool University
Suzhou, China
xiaojun.zhang01@xjtlu.edu.cn

Abstract

Machine translation (MT) models usually translate a text at sentence level by considering isolated sentences, which is based on a strict assumption that the sentences in a text are independent of one another. However, the fact is that the texts at discourse level have properties going beyond individual sentences. These properties reveal texts in the frequency and distribution of words, word senses, referential forms and syntactic structures. Disregarding dependencies across sentences will harm translation quality especially in terms of coherence, cohesion, and consistency. To solve these problems, several approaches have previously been investigated for conventional statistical machine translation (SMT). With the fast growth of neural machine translation (NMT), discourse-level NMT has drawn increasing attention from researchers. In this work, we review major works on addressing discourse related problems for both SMT and NMT models with a survey of recent trends in the fields.

1 Introduction

In the last several decades, the field of machine translation (MT) has experienced three main historical periods including rule-based MT (RMT) (Nirenburg et al., 1986), statistical MT (SMT) (Kong and Zhou, 2010) and neural MT (NMT) (Kalchbrenner and Blunsom, 2013) (Sutskever et al., 2014), unremittingly improving performances of MT systems. As an active research field in natural language processing (NLP), MT is regarded as a sequence-to-sequence prediction task, which aims to find the most probable target language text for a source language one that shares the most similar meaning. It is challenging to generate the high-quality translations, because MT models need to not only thoroughly understand the source text but also have good knowledge of the target text (Hasler et al., 2014).

Natural languages, from bottom to top, are composed of several linguistic units including word, phrase, clause, sentence, paragraph, and discourse(Asher and Lascarides, 2003)(Longacre, 1996). The sentences in a discourse, like words in a sentence, are closely related to one another. However, MT systems usually translate a text in a sentence-by-sentence fashion, and neglect inter-sentence dependencies. Loss of discourse information leads to a series of problems in understanding the semantic meaning of inputs and generating coherent and consistent translations of outputs for an MT system.

Along with the development of MT, a large amount of effort has been going into addressing discourse explicitly in translation models. The 1990s saw an intensification of research efforts aimed at endowing RMT-translated texts with the same document and discourse properties as their source texts (Webber, 2014). This included work on stylistics (Dimarco and Mah, 1994), discourse relations (Mitkov, 1993) and referring forms (Wada, 1990)(Bond and Ogura, 1998) and pronoun translation (Chan and T'sou, 1999)(Ferrández et al., 1999)(Nakaiwa, 1999). Discourse was widely investigated and demonstrated promising results in different aspects including language modelling (Foster et al., 2010), discourse connectives (Meyer and Poláková, 2013)(Meyer and Webber, 2013), lexical cohesion (Xiong et al., 2013), anaphora resolution (Le Nagard and Koehn, 2010)(Taira et al., 2012) and topic adaption (Su et al., 2012)(Hasler et al., 2014) in SMT development and researches. In recent years, NMT has made significant progress towards to constructing and utilizing a single large neural network to handle the entire translation task. The performance of NMT models has surpassed that of traditional SMT in various language pairs (Luong et al., 2015). As discourse information have been shown useful for translation models,

Proceedings of the 2nd International Workshop on Discourse Processing, pages 4–12
Suzhou, China, December 7, 2020. ©2020 Association for Computational Linguistics

a growing number of work investigated a variety of neural architectures to model document structure (Wang et al., 2017a)(Jean et al., 2017)(Miculicich et al., 2018)(Zhang et al., 2018), discourse structure (Chen et al., 2020) and cache memory (Tu et al., 2017)(Kuang et al., 2017)(Maruf and Haffari, 2018). Similar to research line in SMT, several work focusing on explicitly modelling different discourse phenomena especially anaphora (Wang et al., 2018b)(Wang et al., 2018a). Some researchers further studied the discourse-aware evaluation and analysis (Läubli et al., 2018)(Kim et al., 2019).

The aim of this survey is to highlight the major works that have been undertaken in the space of discourse-level MT in SMT and NMT. In the term of discourse-level MT, we mean works which utilise inter-sentential context information comprising discourse aspects of a text or surrounding sentences in the text. In addition to this, we also cover the evaluation strategies introduced to account for improvements in this domain and conclude by presenting avenues for future research. Before moving on with the main agenda, we briefly describe the basics of statistical and neural MT models and their evaluation in the following section.

2 Preliminaries

2.1 Statistical Machine Translation

Statistical machine translation starts with a very large data set of good translations, that is, a corpus of texts which have already been translated into multiple languages, and then uses those texts to automatically infer a statistical model of translation. That statistical model is then applied to new texts to make a guess as to a reasonable translation.

SMT splits translation into three problems: (1) build a language model (LM); (2) build a translation model (TM); and (3) search for maximizing the product. Each of these problems is itself a rich problem which can be solved in many different ways.

The challenge in building a good LM is that there are so many distinct conditional probabilities that need to be estimated (Koehn, 2009). However, the training procedure is likely to underestimate the probability of bigrams which don't appear in the training set, and overestimate the probability of those which do. The problem is even worse for trigrams. Two basic approaches of linear interpolation and discount factor are always utilized to release the difficulties of probability estimation (Clarkson, 1999).

Two notions of fertility and distortion derived from alignments are particularly useful in building up the TM (Brown et al., 1990) and some simple parameters are related to the both notions as a) fertility probability, the probability that the source word has fertility; b) distortion probability, which is the probability that an source segment at position corresponds to a target segment at position in a target sentence; and c) translation probability, one for each target segment and source segment. This should not be confused with the case when source and target segments are sentences.

As to search for maximizing the product, for example, how translation from French to English can be viewed as the problem of finding an English sentence (e) which maximizes the probability $pr(e|f)$, and that this was equivalent to maximizing $pr(f|e)pr(e)$, where f is a French sentence. Using Bayes' theorem this is equivalent to finding which maximizes

$$pr(e,a|f) = \frac{pr(f,a|e)pr(e)}{p(f)} (1)$$

Here, a refers to alignment.

Furthermore, although the basic ideas remain the same, many improvements to these ideas have been made since 1990. The biggest single advance seems to have been a movement away from words as the unit of language, and towards phrase-, tree- and forest-based models, which give greatly improved performance.

2.2 Neural Machine Translation

A standard NMT model directly optimizes the conditional probability of a target sentence $Y = y_1, ..., y_j$ given its corresponding source sentence $X = x_1, ..., x_j$:

$$P(Y|X;\theta) = \prod_{j=1}^{J} P(y_i|Y < j, X;\theta) (2)$$

where θ is a set of model parameters and y¡j denotes the partial translation. The probability $P(Y|X;\theta)$ is defined on the neural network based encoder-decoder framework (Sutskever et al., 2014)(Cho et al., 2014), where the encoder summarizes the source sentence into a sequence of representations $H = H_1, ..., H_I$ with $H \in R^{I*d}$, and the decoder generates target words based on the representations. Typically, this framework can be implemented as recurrent neural network (RNN) (Bahdanau et al., 2016), convolutional neural network (CNN) (Gehring et al., 2017) and Transformer (Vaswani et al., 2017). The parameters of the NMT model are trained to maximize the likelihood of a set of training examples $D = [x^m, y^m]_{m=1}^{M}$:

$$L(\theta) = \arg\max_{\theta} \sum_{m=1}^{M} \log P(y^m|x^m;\theta)(3)$$

As seen, the standard NMT model is computed in a sentence-by-sentence manner, neglecting discourse functionalities across sentences. However, discourse-aware NMT aims to improve the performance by incorporating discourse information into models. Besides, BLEU (Papineni et al., 2002) is the most commonly used automatic metric to evaluate translation qualities of MT outputs. Its range is 0-100% indicating the similarity between the MT outputs and the reference translations (i.e. the higher the better).

3 Discourse

Discourse contains seven fundamental properties including cohesion, coherence, intentionality, acceptability, informatively, situationality and intertextuality (Beaugrande and Dressler, 1981). Cohesion and coherence are two most basic properties that establish “connectedness” in a text (Sanders and Maat, 2006).

Cohesion is a surface property of the text that is realised by explicit clues. It occurs whenever “the interpretation of some element in the discourse is dependent on that of another” (Bernhardt, 1980). Cohesion is mainly realised by the way of pronominal reference including anaphora and coreference. Anaphora is the use of an expression whose interpretation depends specifically upon antecedent expression. The anaphoric (referring) term is called an anaphor. Sometimes anaphor may rely on the postcedent expression, and this phenomenon is called cataphora. Taking Sentence (a) in Figure 1 for example, the pronoun It is an anaphor, which points to the left toward its antecedent Audi. Zero Anaphora (pronoun-dropping) is a more complex case of anaphora. In some languages such as Chinese and Japanese, certain classes of words can be omitted to make the sentence compact yet comprehensible when the identity of the pronouns can be inferred from the context. Coreference means two or more expressions (e.g. nouns) in a text refer to the same referent. As the referents point to persons or things in the real world, the coreference relation can exist independently of the context. Taking Sentence (b) in Figure 1 for instance, the noun phrases HK Chief Executive and Mr. Tung Chee-hwa point to the same person, although their surfaces are totally different.

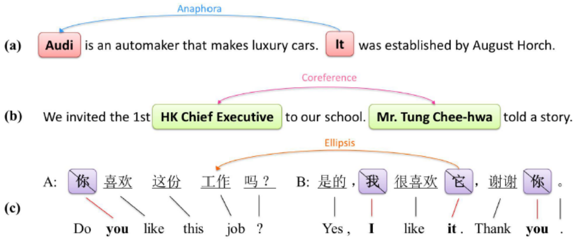
Anaphora
(a) Audi is an automaker that makes luxury cars. It was established by August Horch.
Coreference
(b) We invited the 1st HK Chief Executive to our school. Mr. Tung Chee-hwa told a story.
Ellipsis
(c)
A: 你 喜欢 这份 工作 吗？ B: 是的，我 很喜欢 它，谢谢 你 。
Do you like this job ? Yes , I like it . Thank you .

Figure 1: Examples of cohesion

Coherence is related to the connectedness of the “mental representation of the text rather than of the text itself” (Sanders and Maat, 2006) to make a text semantically meaningful. It is created referentially, when different parts of a text refer to the same entities, and relationally, by means of coherence relations such as “Cause-Consequence” between different discourse segments. Researchers studied discourse structure mainly based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Discourse Tree Bank (PDTB) annotation methodology (Marcu et al., 2000). RST relations are applied recursively in a text, until all units in that text are constituents in a predefined relation. The result of such analysis is that RST structure are typically represented as trees, with one top level relation that encompasses other relations at lower levels. Besides, the tree also contains a number of predefined relations such as “Attribution” (Cause-Consequence), “Elaboration” etc. The PDTB annotation methodology is proposed based on RST, but highlights the role of the connectives. According to whether containing a connective or not, discourse relations can be divided into two categories: explicit and implicit.

4 Discourse Structure

Foster et al. (Foster et al., 2010) try the first attempt to incorporate structural information into SMT. They tagged each sentence with features such as kind of session, identity of the speaker, time period, and then used domain adaptation methods to balance between a LM trained from similar data and a background LM. Marcu et al. (Marcu, 2000) found that there are significant differences in discourse structure of between Japanese and English. Thus, they propose an "analysis–transfer–translate" pipeline: firstly, Japanese text is parsed into RST tree; and then it is transferred into English style RST tree; finally process translation based on the RST tree. Although all the training data are manually annotated (high cost), the method really improves the translation in term of coherence. Besides, Tu et al. (Tu et al., 2017) propose a novel translation framework, which mainly includes three steps: 1) Source RST tree acquisition: a source sentence is parsed into an RST tree; 2) Rule extraction: translation rules are extracted from the source tree and the target string via bilingual word alignment; 3) RST-based translation: the source RST- tree is translated with translation rules. Experiments show that their approach achieves improvements of about +2 BLEU points than the baseline system on Chinese–English.

Because of the superior ability to preserve sequence information over time, Long Short Term Memory (LSTM) has obtained strong results on a variety of sequence modeling tasks. Sequence models construct sentence representations as an order-sensitive function of the sequence of tokens. In contrast, tree-structured models compose each phrase and sentence representation from its constituent sub-phrases according to a given syntactic structure over the sentence. Tai et al. (Tai et al., 2015) introduce a Tree-LSTM, a generalization of LSTMs to tree- structured network topologies. The difference between Tree-LSTM and LSTM is that the Tree-LSTM composes its state from an input vector and the hidden states of arbitrarily many child units. Thus, the standard LSTM can then be considered a special case of the Tree-LSTM where each internal node has exactly one child. They show its superiority for representing sentence meaning over a sequential LSTM in two tasks: predicting the semantic relatedness of two sentences and sentiment classification.

Although their works show promising improvements, there are several underlying drawbacks: 1) some models are trained on small-scale or manually-created data sets, it is not reliable when adopting these approaches to large-scale MT task; and 2) the performance of discourse parser is still not reliable, thus incorporating the structure information into NMT will result in error propagation problems.

5 Document Structure

One direction is cache-based methods, which employ cache to retain bilingual phrase pairs from the best hypothesis of previously translated sentences and then use it as an additional feature in log-linear model of SMT. Tiedemann (Tiedemann, 2010) integrated cache-based language and translation models within a PBSMT decoder and used an exponential decay factor to carry over word preferences from one sentence to the next. When a source phrase is considered for translation, its cache translation score is computed using the phrase probabilities of matching phrases found in the cache and the decay factor. Their examples illustrate better translation especially in repetition and consistency, however, the experimental score show modest improvements. Gong et al. (Gong et al., 2011) extended this work by using three caches: dynamic cache, static cache, and topic cache. They show a better improvement when all three caches are used in combination.

Other efforts are in document-level decoding. Focusing on translation consistency, Xiao et al. (Xiao et al., 2011) employed a forced-decoding method: identify ambiguous words in the output of baseline system, and then obtain a set of consistent translations based on frequencies and finally re-decode input using the filtered set of translation options. Hardmeier et al. (Hardmeier et al., 2012) approach translation as an optimization task. He proposed a stochastic local search decoding method for PBSMT, which permits free document-wide dependencies in the models. Their work on decoding try to reduce the searching space but it is difficult to incorporate new knowledge.

Recently, researchers began to explore neural-network (NN)-based discourse-level approaches for sequence modeling. Conversational models need to predict the next sentence by considering the historical utterances in a conversation. Vinyals and Le (Vinyals and Le, 2015) built an end-to-end conversational system using a sequence-to-sequence framework. In order to capture the lager-context information, they simply concatenate previous utterances together as the input. Their preliminary results show that the method is able to converse well and extract knowledge from lager-context. Li et al. (Li et al., 2016) argue that simply incorporating context information into context independent message will increase the workload of a generation system and has the risk of bringing in noise to the generation process. To better preserve the original search intent, Sordoni et al. (Sordoni et al., 2015) proposed a novel Hierarchical Recurrent Encoder-Decoder (HRED) model to summarize these historical queries. Besides, Serban et al. (Serban et al., 2016) adopt the framework to the task of dialogue response generation. They use HRED to summarize a single representation from both the current and previous sentences. Experiments demonstrated that availing of the historical representation helps to maintain the dialogue context.

The continuous vector representation of a symbol encodes multiple dimensions of similarity, equivalent to encoding more than one meaning of a word. Consequently, NMT needs to spend a substantial amount of its capacity in disambiguating source and target words based on the context defined by a source sentence. Without additional information, standard NMT models are facing inconsistency and ambiguity problems. Calixto and Liu (Calixto and Liu, 2017) utilize global image features extracted using a pre-trained convolutional neural network and incorporate them in NMT. The work related to multi-source and multi-target NMT are investigated one-to-many or many-to-one languages translation tasks by integrating additional encoders or decoders into encoder-decoder framework, and their experiments show promising results. More recently, some researchers propose to use an additional set of an encoder and attention to model more information. For example, Jean et al. (Jean et al., 2017) use it to encode and select part of the previous source sentence for generating each target word.

6 Evaluation

More recently, there are some new publications on discourse-level NMT evaluation. In order to evaluate discourse phenomena in NMT, Bawden et al. (Bawden et al., 2018) conducted experiments from three aspects: 1) comparing multi-encoder models (Zoph and Knight, 2016)(Jean et al., 2017) with different strategies; 2) investigating the impacts of source- and target-side history information on NMT; 3) presenting a novel evaluation through the use of two discourse test sets targeted at coreference and lexical coherence/cohesion. Voita et al. (Voita et al., 2018) introduced a context-aware model and demonstrated its usefulness for anaphora resolution as well as translation. Besides, Xiong (Xiong et al., 2013) proposed modeling method to use discourse context and reward to refine the translation quality from the perspective of coherence. Some researchers proposed to extend the Transformer model to take advantage of discourse-level context (Miculicich et al., 2018). Following the work from Tu et al. (Tu et al., 2013), Kuang et al. (2017) and Maruf & Haffari (Maruf and Haffari, 2018) continue to exploit cache memory for improving the performance of discourse-level NMT. Through human evaluation, Läubli (Läubli et al., 2018) has found that discourse-level evaluation for MT can improve to discriminate the errors which are hard or impossible to spot at the sentence level.

7 Dropped Pronoun Issue

As discussed in Section 2.3, one issue of cohesion is pronominal anaphora. Targeting cohesion phenomena, some researchers investigated approaches of incorporating anaphora information to improve the performance of MT. For instance, Le Nagard and Koehn (Le Nagard and Koehn, 2010) presented a method to aid English pronoun translation into French for SMT by integrating an anaphora resolution system. In the thesis, we mainly focus on the more complicated phenomenon: dropped pronoun (DP), which can be regarded as a special case of pronominal anaphora. Thus, in the following contents, we mainly review related work on DP.

7.1 Dropped Pronoun Recovery

There are two research strands related to DP recovery. One is called Zero-pronoun (ZP) resolution. ZP resolution contains three steps: ZP detection, anaphoricity determination and reference linking. Researchers (Zhao and Ng, 2007)(Kong and Zhou, 2010)(Chen and Ng, 2013) propose rich features using different machine learning models. For example, Chen and Ng (Chen and Ng, 2013) propose a support vector machine (SVM) classifier using 32 features including lexical, syntax and grammatical roles and show significant improvement on this task. Another research direction is related to a wider range of empty category (EC) phenomena, which aims to recover long-distance dependencies, discontinuous constituents and certain dropped elements in phrase structure treebanks (Yang et al., 2006). However, their work mainly focuses on intra-sentential characteristics as opposed to the discourse level. More recently, Yang et al. (Yang et al., 2015) explored DP recovery for Chinese text messages based on both ZP and EC.

Most of their work either applies manual annotation (Yang et al., 2015) recovering or uses existing but small-scale resources (OntoNotes corpus contains 144K coreference instances, but only 15% of them are dropped subjects). There are two drawbacks on current work: 1) performance is not reliable when directly using the results of these systems in translation process; 2) the data is not big enough to drive a large neural model. Therefore, the primary challenge of this work is how to automatically build a large-scale high-quality DP training corpus.

7.2 Dropped Pronoun Translation

Some work has been done on DP translation for SMT models (Chung and Gildea, 2010)(Le Nagard and Koehn, 2010)(Taira et al., 2012). Le Nagard and Koehn (Le Nagard and Koehn, 2010) presented a method to aid English pronoun translation into French by using the results of a co-reference (CR) system, unfortunately, their results are not convincing due to the poor performance of the CR system. Chung and Gildea (Chung and Gildea, 2010) systematically examine the effects of EC on MT with three methods: pattern, conditional random field (CRF) and parsing. The results show that this work can really improve the end translation, even though the automatic prediction of EC is not highly accurate. Furthermore, Taira et al. (Taira et al., 2012) propose both simple rule-based and manual methods to add DPs on the source side for Japanese–English translation. However, the BLEU scores of both methods are nearly identical, which indicates that only considering the single source sentence and forcing the insertion of pronouns may be less principled than tackling the problem head on by integrating them into the SMT model itself.

Their work regards the task of DP/EC recovering as a pre-processing stage for MT. Although these parameters are tuned independently, this direct idea is still worth trying. DP neural translation received relatively little attention from the MT community, thus we are encouraged to explore DP translation for NMT models (Wang et al., 2017b).

Acknowledgments

This work is supported by the XJTLU KSF project (Grant Number: KSF-E-24) and GDUFS open project (Grant Number: CTS201501). The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. ArXiv: 1409.0473.

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.

Robert-Alain de Beaugrande and Wolfgang Ulrich Dressler. 1981. *Einführung in die Textlinguistik*. De Gruyter. Publication Title: Einführung in die Textlinguistik.

Stephen A. Bernhardt. 1980. Review of COHESION IN ENGLISH. *Style*, 14(1):47–50. Publisher: Penn State University Press.

Francis Bond and Kentaro Ogura. 1998. Reference in Japanese–English Machine Translation. *Machine Translation*, 13(2):107–134.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Iacer Calixto and Qun Liu. 2017. Incorporating Global Visual Features into Attention-based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Samuel W. K. Chan and Benjamin K. T'sou. 1999. Semantic Inference for Anaphora Resolution: Toward a Framework in Machine Translation. *Machine Translation*, 14(3):163–190.

Chen Chen and Vincent Ng. 2013. Chinese Zero Pronoun Resolution: Some Recent Advances. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1360–1365, Seattle, Washington, USA. Association for Computational Linguistics.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling Discourse Structure for Document-level Neural Machine Translation. *arXiv:2006.04721 [cs]*. ArXiv: 2006.04721.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2010. Effects of Empty Categories on Machine Translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 636–645, Cambridge, MA. Association for Computational Linguistics.

P. R. Clarkson. 1999. *Adaptation of statistical language models for automatic speech recognition*. Ph.D., University of Cambridge. Accepted: 1999.

Chrysanne Dimarco and Keith Mah. 1994. A Model of Comparative Stylistics for Machine Translation. *Machine Translation*, 9(1):21–59. Publisher: Springer.

Antonio Ferrández, Manuel Palomar, and Lidia Moreno. 1999. An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation*, 14(3):191–216.

George Foster, Pierre Isabelle, and Roland Kuhn. 2010. Translating structured documents. In *In Proceedings of AMTA*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR. ISSN: 2640-3498.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 909–919, USA. 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.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic Topic Adaptation for Phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–337, Gothenburg, Sweden. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? *arXiv:1704.05135 [cs, stat]*. ArXiv: 1704.05135.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and Why is Document-level Context Useful in Neural Machine Translation? *arXiv:1910.00294 [cs]*. ArXiv: 1910.00294.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press. Google-Books-ID: kKYgAwAAQBAJ.

Fang Kong and Guodong Zhou. 2010. A Tree Kernel-Based Unified Framework for Chinese Zero Anaphora Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA. Association for Computational Linguistics.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Cache-based Document-level Neural Machine Translation. *ArXiv*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

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.

Chaozhuo Li, Yu Wu, Wei Wu, Chen Xing, Zhoujun Li, and Ming Zhou. 2016. Detecting Context Dependent Messages in a Conversational Environment. *arXiv:1611.00483 [cs]*. ArXiv: 1611.00483.

Robert E. Longacre. 1996. *The Grammar of Discourse*. Springer Science & Business Media. Google-Books-ID: GHRkak8j8x8C.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text & Talk*, 8(3):243–281. Publisher: De Gruyter Mouton Section: Text & Talk.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press. Google-Books-ID: VyjED9VOn5MC.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The Automatic Translation of Discourse Structures. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Sameen Maruf and Gholamreza Haffari. 2018. Document Context Neural Machine Translation with Memory Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Thomas Meyer and Lucie Poláková. 2013. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of Discourse Connectives in (Machine) Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. 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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Ruslan Mitkov. 1993. How Could Rhetorical Relations Be Used in Machine Translation? In *Intentionality and Structure in Discourse Relations*.

Hiromi Nakaiwa. 1999. Automatic Extraction of Rules for Anaphora Resolution of Japanese Zero Pronouns in Japanese–English Machine Translation from Aligned Sentence Pairs. *Machine Translation*, 14(3):247–279.

Sergei Nirenburg, Victor Raskin, and Allen Tucker. 1986. On Knowledge-Based Machine Translation. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Ted Sanders and H.L.W. Maat. 2006. Cohesion and Coherence: Linguistic Approaches. In *Encyclopedia of Language & Linguistics*, pages 591–595. Journal Abbreviation: Encyclopedia of Language & Linguistics.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783, Phoenix, Arizona. AAAI Press.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 553–562, New York, NY, USA. Association for Computing Machinery.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 459–468, Jeju Island, Korea. Association for Computational Linguistics.

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 - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero Pronoun Resolution can Improve the Quality of J-E Translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118, Jeju, Republic of Korea. Association for Computational Linguistics.

Jörg Tiedemann. 2010. Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A Novel Translation Framework Based on Rhetorical Structure Theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–374, Sofia, Bulgaria. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context Gates for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

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.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv:1506.05869 [cs]*. ArXiv: 1506.05869.

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.

Hajime Wada. 1990. Discourse Processing in MT: Problems in Pronominal Translation. In *COLING 1990 Volume 1: Papers presented to the 13th International Conference on Computational Linguistics*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating Pro-Drop Languages with Reconstruction Models. *arXiv:1801.03257 [cs]*. ArXiv: 1801.03257.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to Jointly Translate and Predict Dropped Pronouns with a Shared Reconstruction Mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2997–3002, Brussels, Belgium. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Siyou Liu, Hang Li, Andy Way, and Qun Liu. 2017b. A novel and robust approach for pro-drop language translation. *Machine Translation*, 31(1):65–87.

Bonnie Webber. 2014. Discourse for Machine Translation. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 27–27, Phuket,Thailand. Department of Linguistics, Chulalongkorn University.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. *Proceedings of the 2011 MT Summit*, 13.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü, and Qun Liu. 2013. Modeling lexical cohesion for document-level machine translation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, IJCAI '13, pages 2183–2189, Beijing, China. AAAI Press.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2006. Kernel-Based Pronoun Resolution with Structured Syntactic Knowledge. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 41–48, Sydney, Australia. Association for Computational Linguistics.

Yaqin Yang, Yalin Liu, and Nianwen Xue. 2015. Recovering dropped pronouns from Chinese text messages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 309–313, Beijing, China. Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 541–550, Prague, Czech Republic. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-Source Neural Translation. *arXiv:1601.00710 [cs]*. ArXiv: 1601.00710.