On Integrating Discourse in Machine Translation

Karin Sim Smith
Department of Computer Science, University of Sheffield, UK
{kmsimsmith1}@sheffield.ac.uk

Abstract

As the quality of Machine Translation (MT) improves, research on improving discourse in automatic translations becomes more viable. This has resulted in an increase in the amount of work on discourse in MT. However many of the existing models and metrics have yet to integrate these insights. Part of this is due to the evaluation methodology, based as it is largely on matching to a single reference. At a time when MT is increasingly being used in a pipeline for other tasks, the semantic element of the translation process needs to be properly integrated into the task. Moreover, in order to take MT to another level, it will need to judge output not based on a single reference translation, but based on notions of fluency and of adequacy – ideally with reference to the source text.

1 Introduction

Despite the fact that discourse has long been recognised as a crucial part of translation (Hatim and Mason, 1990), when it comes to Statistical Machine Translation (SMT), discourse information has been mostly neglected. Recently increasing amounts of effort have been going into addressing discourse explicitly in MT, with research covering lexical cohesion (Wong and Kit, 2012; Xiong et al., 2013b,a; Gong et al., 2015; Mascarell et al., 2015), discourse connectives (Cartoni et al., 2012, 2013; Meyer and Popescu-Belis, 2012; Meyer, 2011; Meyer et al., 2011; Steele, 2015; Steele and Specia, 2016), discourse relations (Guzmán et al., 2014), pronoun prediction (Guillou, 2012; Hardmeier et al., 2013b; Guillou, 2016) and negation (Fancellu and Webber, 2014; Wetzel and Bond, 2012).

Considerable progress was made in the field of SMT over the past two decades, culminating in models which give surprisingly good output given the limited amount of crosslingual information they have. Neural Machine Translation (NMT) models are now the most performant, to the extent that in the past year they have been the best performing at WMT (Bojar et al., 2016), and although deeper than the linguistically superficial SMT, to evaluate progress we need to be able to measure the extent to which these models successfully integrate discourse. Besides the difficulty of the task, one of the issues preventing progress is a lack of understanding regarding the problem: what is the purpose of translation. In order to fulfil its role, MT needs to capture and transfer the communicative message of the Source Text (ST) into the Target Text (TT). While MT cannot be expected to assess the pragmatics, in terms of the intended effect on the target audience of the Source Language (SL) and ensuring a corresponding effect on the target audience of the Target Language (TL), there is a basic communicative intent in terms of the semantics which has to surely be taken into account in evaluation, if we are to move beyond stringing together phrase matches.

Despite agreement on the shortcomings of BLEU (Papineni et al., 2002), for example (Smith et al., 2016), the standard metrics are still based on comparison to a single reference translation, which is inflexible (requiring a professional translation for every text automatically translated), and is also unrealistic as a text can be translated many ways, all of them valid. We would also argue that it does not incentivise the integration of deeper linguistic elements.

In the next section (Section 2) we give a brief survey of recent work on Discourse in MT. We

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 110–121,
Copenhagen, Denmark, September 8, 2017.

then describe the constraints of SMT architecture (Section 3), followed by a brief description of the translation process from the human translator's perspective (Section 4) and a review of the limitations of the current evaluation paradigm (Section 6).

2 Discourse in MT

While the survey by Hardmeier (2012) provides a good overview of Discourse in SMT at the time, his survey has been superceded by a flurry of work, much of it in association with the Workshop on DiscoMT (Webber et al., 2013, 2015). We give a brief survey of more recent research in the field of discourse, since his survey, specifically as it relates to discourse phenomena in the MT context.

Reference resolution and pronoun prediction Anaphora resolution, as reference resolution to something or someone previously mentioned, is a very challenging issue in MT which has been studied by several researchers over the past few years (Hardmeier, 2012). It is something that SMT currently handles poorly, again due to the lack of intersentential references. Anaphoric references are affected in several ways. The context of the preceding sentences is absent, meaning that the reference is undetermined. Even once it is correctly resolved (by additional pre-training or a second-pass), reference resolution is directly impacted by linguistic differences. For example, the target language may have multiple genders for nouns while the source only has one. The result is that references can be missing or wrong.

Novák and Žabokrtský (2014) developed a crosslingual coreference resolution between Czech and English, with mixed results, indicating the complexity of the problem. Subsequently Hardmeier et al. (2013b) have attempted a new approach to anaphora resolution by using neural networks which independently achieve comparable results to a standard anaphora resolutions system, but without the annotated data.

Luong and Popescu-Belis (2016) focus on improving the translation of pronouns from English to French by developing a target language model which determines the pronoun based on the preceding nouns of correct number and gender in the surrounding context. They integrate by means of reranking the translation hypotheses and improving over the baseline of the DiscoMT 2015 shared task.

Luong and Popescu-Belis (2017) develop a probabilistic anaphora resolution model which they integrate in a Spanish-English MT system, to improve the translation of Spanish personal and possessive pronouns into English using morphological and semantic features. They evaluate the Accuracy of Pronoun Translation (APT) using the translated pronouns of the reference translation and report an additional 41 correctly translated pronouns from a base line of 1055.

More recently, pronoun prediction in general has been the focus of increased attention, resulting in the creation of a specific WMT Shared Task on *Cross-lingual Pronoun Prediction* (Guillou et al., 2016), and to the development of resources such as test suites (Guillou and Hardmeier, 2016) for the automatic evaluation of pronoun translation. This has led to varied submissions on the subject, predicting third person subject pronouns translated from French into English; (Novák, 2016; Loáiciga, 2015; Wetzel et al., 2015). Most recently, we have seen an entire thesis on incorporating pronoun function into MT (Guillou, 2016), the main point being that pronouns should be handled according to their function– both in terms of handling within SMT and in terms of evaluation.

However, progress has been hard and Hardmeier (2014) suggests that besides evaluation problems, this is due to a failure to fully grasp the extent of the pronoun resolution problem in a crosslingual setting, and that anaphoric pronouns in the ST cannot categorically be mapped onto target pronouns. If these issues can be successfully addressed, it will mark significant progress for MT output in general.

In her thesis Loaiciga Sanchez (2017) focuses on pronominal anaphora and verbal tenses in the context of machine translation, on the basis that a pronoun and its antecedent (the token which gives meaning to it), or a verbal tense and its referent, can be in different sentences and result in errors in MT output, directly impacting cohesion. She reports direct improvements in terms of BLEU scores for both elements. Again one cannot help wondering whether the improvement in terms of quality of the text as a whole is actually much higher than reflected in the improvements over BLEU score.

Verb tense In specific work on verbs, (Loaiciga et al., 2014) researches improving alignment for non-contiguous components of verb phrases by

POS tags and heuristics. They then annotated Europarl and trained a tense predictor which they integrate in an MT system using factored translation models, predicting which English tense is appropriate translation for a particular French verb. This results in a better handling of tense, with the added benefit of an increased BLEU score.

Again on verbs, but this time with a focus on the problems that arise in MT from the verb-particle split constructions in English and German, Loáiciga and Gulordava (2016) construct test suites and compare how syntax and phrase-based SMT systems handle these constructs. They show that often there are alignment issues (with particles aligning to null) which lead to mistranslations, and that the syntax-based systems performed better in translating them.

Lexical Cohesion There has been work in the area of lexical cohesion in MT assessing the linguistic elements which hold a text together, and how well these are rendered in MT.

Wong and Kit (2012) study lexical cohesion as a means of evaluating the quality of MT output at document level, but in their case the focus is on it as an evaluation metric. While human translators intuitively ensure cohesion, their research indicated that MT output is often represented as direct translations of ST items that may be inappropriate in the target context. They conclude that MT needs to learn to use lexical cohesion devices appropriately.

These findings are echoed by Beigman Klebanov and Flor (2013) in their research on word associations within a text, who consider pairs of words and define a metric for calculating the *lexical tightness* of MT versus Human Translation (HT). The fact that they had to first improve on the raw MT output before the experiment, indicates that it was of insufficient quality in the first place, however this is perhaps due to the age of data (dating to 2008 evaluation campaign), as MT has progressed considerably since then.

Xiong and Zhang (2013) attempt to improve lexical coherence via a topic-based model, using a Hidden Topic Markov Model (HTMM) to determine the topic in the source sentence. They extract a coherence chain for the source sentence, and project it onto the target sentence to make lexical choices during decoding more coherent. They report very marginal improvement with respect to a baseline system in terms of automatic evaluation. This could indicate that current evaluation metrics are limited in their ability to account for improvements related to discourse.

Xiong et al. (2013a) focus on ensuring lexical cohesion by reinforcing the choice of lexical items during decoding. They subsequently compute lexical chains in the ST, project these onto the TT, and integrate these into the decoding process with different strategies. This is to try and ensure that the lexical cohesion, as represented through the choice of lexical items, is transferred from the ST to TT. Gong et al. (2015) attempt to integrate their lexical chain and topic based metrics into traditional BLEU and METEOR scores, showing greater correlation with human judgements on MT output.

In their work on comparative crosslingual discourse phenomena, Lapshinova-Koltunski (2015) find that use of various lexical cohesive devices can vary from language to language, and also depend on genre. In a different context, Mascarell et al. (2014) experiment with enforcing lexical consistency at document level for coreferencing compounds. They illustrate that for languages with heavy compounding such as German, translations of coreferencing constituents in subsequent sentences are sometimes incorrect, due to the lack of context in SMT systems. They experiment with two SMT phrase based systems, applying a compound splitter in one of them, caching constituents in both systems, and find that besides improving translations the latter also results in fewer out-of-vocabulary nouns.

Guillou (2013) investigates lexical cohesion across a variety of genres in HT, in an attempt to determine standard practice amoung professional translators, and compare it to output from SMT systems. She uses a metric (Herfindahl Hirschman Index (HHI)) to determine the terminological consistency of a single term in a single document, investigating consistency across words of different POS category. She finds that in SMT consistency occurs by chance, and that inconsistencies can be detrimental to the understanding of a document.

One of the problems with repetition is indeed automatically recognising where it results in consistency, and where it works to the detriment of lexical variation. Most recently, Martìnez Garcia et al. (2017) use word embeddings to promote lexical consistency at document level, by implementing a new feature for their document-level de-

coder. In particular, they try to encourage consistency for the same word to be translated in a similar manner throughout the document. They deploy a cosine similarity metric between word embeddings for the current translation hypothesis and the context to check if they are semantically similar. Despite the fact that a bilingual annotator judging at document level found the improved output to be better than the baseline 60% of the time, and equal 20% of the time (i.e. the improved output is better or the same for 80% of the documents), there was *no statistical significance* in the automatic evaluation scores (Martìnez Garcia et al., 2017).

Word Sense Disambiguation The very nature of languages is such that one word in a particular language has no one-to-one mapping in another; a particular word in the source could be semantically equivalent to several in the target, and there is a need to disambiguate.

In their work, Mascarell et al. (2015) use trigger words from the ST to try to disambiguate translations of ambiguous terms, where a word in the source language can have different meanings and should be rendered with a different lexical item in the TT depending on the context it occurs in.

Xiong and Zhang (2014)'s sense-based SMT model tries to integrate and reformulate the Word Sense Disambiguation (WSD) task in the translation context, predicting possible target translations. Zhang and Ittycheriah (2015) experiment with three types of document level features, using context to try and improve WSD. They use context on both target and source side, and establish whether the particular alignments had already occurred in the document, to help in disambiguating the current hypothesis. Experimenting with the Arabic-English language pair, they show an increased BLEU (Papineni et al., 2002) score and a decreased error rate.

Discourse relations and discourse connectives Discourse relations have long been recognised as crucial to the proper understanding of a text (Knott and Dale, 1993), as they provide the structure between units of discourse (Webber et al., 2012). Discourse relations can be implicit or explicit. If explicit, they are generally signalled by the discourse connectives.

While Marcu et al. (2000) and Mitkov (1993) previously investigated coherence relations as a means of improving translation output and ensuring it was closer to the target language this was taken no further at the time. Taking inspiration from Rhetorical Structure Theory (RST), Tu et al. (2013) proposed an RST-based translation framework on basis of elementary discourse units (EDU)s, in an attempt to better segment the ST in a meaningful manner, and ensure a better ordering for the translation. This approach is more sensitive to discourse structure, and introduces more semantics into the SMT process. Their research is effected using a Chinese RST parser, and they aim to ensure a better ordering of EDUs, although the framework still has a limited sentence-based window.

There have been a few previous experiments specifically assessing discourse relations in an MT context. Guzmán et al. (2014) used discourse structures to evaluate MT output. They hypothesize that the discourse structure of good translations will have similar discourse relations. They parse both MT output and the reference translation for discourse relations and use tree kernels to compare HT and MT discourse tree structures. They improve current evaluation metrics by incorporating discourse structure on the basis that 'good translations should tend to preserve the discourse relations' (Guzmán et al., 2014).

Discourse connectives, also known as discourse markers, are cues which signal the existence of a particular discourse relation, and are vital for the correct understanding of discourse. Yet current MT systems often fail to properly handle discourse connectives for various reasons, such as incorrect word alignments, the presence of multiword expressions as discourse markers, and the prevalence of ambiguous discourse markers. These can be incorrect or missing (Meyer and Poláková, 2013; Steele, 2015; Yung et al., 2015).

In particular, where discourse connectives are ambiguous, e.g. some can be temporal or causal in nature, the MT system may choose the wrong connective translation, which distorts the meaning of the text. It is also possible that the discourse connective is implicit in the source, and thus needs to be inferred for the target. While a human translator can detect this, an MT system cannot.

In their work, Zufferey and Popescu-Belis (2017) automatically labelling the meaning of discourse connectives in parallel corpora to improve MT output. In separate work on discourse connectives, Li et al. (2014b) also find that some con-

nectives are ambiguous in English, and in their research on the Chinese-English language pair subsequently report on a corpus study into discourse relations and an attempt to project these from one language which has a PDTB resource, to another which lacks it (Li et al., 2014a). They again mention that there are mismatches, between implicit and explicit discourse connectives. For the same language pair, Yung et al. (2015) research how discourse connectives which are implicit in one language (Chinese), may need to be made explicit in another (English). This is similar to work by Steele (2015) who use placeholder tokens for the implicit items in the source side of the training data, and trains a binary classifier to predict whether or not to insert a marker in the TT. This notion of *explicitation*, and the opposite *implicitation*, is the subject of research by Hoek et al. (2015), who find that implicitation and explicitation of discourse relations occurs frequently in human translations. There seems to be a degree to which the implicitation and explicitation of discourse relations depends on the discourse relation they signal, and the language pair in question.

Negation Recently work has begun on negation in MT, decomposing the semantics of negation and with an error analysis on what MT systems get wrong with negation (Fancellu and Webber, 2015a). For the language pair which they considered (Chinese-English) the conclusion was that determining the scope of negation was the biggest problem, with reordering the most frequent cause. Subsequently, Fancellu and Webber (2015b) show that the translation model scoring is the cause of the errors in translating negation. In general, MT systems often miss the focus of the negation, which results in incorrectly transferred negations that affect coherence.

Coherence Sim Smith et al. (2015) illustrate the type of discourse errors that often arise in MT, which affect coherence in particular. They then illustrate how assessing coherence in an MT context is very different from previous monolingual coherence tasks (Sim Smith et al., 2016), which are often performed on a summarized or shuffled version of a coherent document and where the task is to reorder the sentences correctly. In the latter, the sentences in question are themselves coherent, unlike in MT. They reimplement existing entity models, in addition to a syntax model, which is extended to improve on the state-of-the-art for the shuffling task (Sim Smith et al., 2016).

Trends While there has been much solid research on discourse in MT, the results that are reported are surprisingly limited. In considering why this is the case, we believe while the constraints in the SMT decoder have provided a ceiling on progress, we cannot help wondering whether the accepted current methods of evaluation are at fault, failing to recognise progress in discourse.

3 Constraints

The dominance of SMT a couple of decades ago was detrimental to the inclusion of many linguistic elements. As reported by Hardmeier (2015), "the development of new methods in SMT is usually driven by considerations of technical feasibility rather than linguistic theory". Most decoders work on a sentence by sentence basis, isolated from context, due to both modelling and computational complexity. This directly impacts the extent to which discourse can be integrated.

Docent (Hardmeier et al., 2013a) is a document level decoder, which has a representation of a complete TT translation, to which changes can be made to improve the translation. It uses a multi-pass decoding approach, where the output of a baseline decoder is modified by a small set of extensible operations (e.g. replacement of phrases), which can take into account document-wide information, while making the decoding process computationally feasible. To date, attempts to influence document level discourse in SMT in this manner have been limited. Stymne et al. (2013) attempted to incorporate readability constraints into Docent, in effect jointly achieving the translation and simplification. A similar document level framework was recently developed by Martìnez Garcia et al. (2017), who developed a new operation to ensure that changes could be made to the entire document in one step, making (see Section 2).

As Hardmeier (2015) points out, training on *domain* has traditionally been seen as a way of making the output more relevant. But this is insufficient– it may well capture translation probabilities appropriate to a specific kind of text at training time, but SMT does not capture the full context of the lexical items during decoding and hence sometimes fails to correctly disam-

biguate. So while Hardmeier (2015) suggests that the "crosslinguistic relation defined by word alignments is a sort of translational equivalence relation", we would claim that equivalence in a translation context traditionally includes an element of semantics which is totally absent in SMT, which is the paradigm he was referring to. While SMT is a complex and finely tuned system, which brought about considerable progress in the MT domain, it is linguistically impoverished, superficially concatenating phrases which have previously been found to align with those of another language when training, with no reference to the intended meaning in context. NMT has been proven to capture elements of context (syntactic and semantic), which are already helping to make NMT output better than that of SMT.

All of these constraints in SMT have restricted integration of linguistic elements and prevented progress to another level. With the success of NMT and the changed architecture it brings, embrace this opportunity to advance to a deeper level of translation. As illustrated by recent comparative research into output from Phrase Based Machine Translation (PBMT) and NMT systems (Popović, 2017; Burchardt et al., 2017), the latter is capable of producing output which is far more linguistically informed.

It would seem a good time to revisit the basics of translation theory, with a view to taking MT to a deeper level.

4 Translation as communication

The popularity of SMT in the past couple of decades has largely been to the exclusion of deeper linguistic elements (besides the linguistically-informed element of syntax-based systems). Performance of SMT systems surpassed previous rules-based systems, and progress was characterised by the famous quote by Frederick Jelinek :"Every time I fire a linguist, the performance of the speech recognizer goes up".

Translation theory has evolved over the years, from the functional and dynamic equivalence of Nida and Taber (1969), to Baker (1992)'s view of equivalence (word, grammatical, textual, pragmatic equivalence), Hatim and Mason (1990)'s view of the translator as a communicator and mediator and the Relevance theory of Sperber and Wilson (1986).[1] Essentially nowadays there is broad agreement on the importance of discourse analysis: on the need to extract the communicative intent and transfer it to the target language- in an appropriate manner, taking account of the cultural context, and the genre.

While there is now a great need for translation, which cannot be met by humans (in terms of the cost or number of human translators), MT can be usefully deployed for gisting, and for some language pairs even as a good quality first draft. However, if it is to be more, for example to be used as part of a pipeline for a series of tasks, then it needs to embrace its role in terms of semantics. Used in pipelines such as voice translators, where Speech Acts are relevant, or as vital components of a multimodal framework, we cannot ignore the fact that semantics are currently not a core building block in MT.

As has been said by others previously (Becher, 2011), MT could benefit from mimicking the way a human translator works. Translators makes several passes on a text. They begin by reading the ST, and extracting the communicative intent– establishing what the author of the text is trying to say. They identify any cultural references, and any acronyms or terminology relevant to the domain. For the former, they need to be aware of the significance of the references and their connotations. They then attempt to transfer these *in an appropriate manner* to the TT, taking account of their TT audience. While MT is far from this it has to at least begin to grapple with semantics, if it is to perform a meaningful role.

5 Semantics

In terms of proposing how this might look for evaluation purposes, we would suggest that semantic parsing may offer one way forward. While this is not available in many languages, and may start off as a limited evaluation method, there are ways in which this can be done.

Progress in the field of semantics has been considerable recently, and in particular work based on Univeral Dependencies (UD)[2] would seem to offer new opportunities which MT evaluation could benefit from: UD are annotations of multilingual treebanks which have been built to ensure crosslingual compatibility. The latest version (2.0) covers 50 languages. Recent work by (Reddy et al., 2016)

[1] Cognitive Linguistics is a further development which is beyond the scope of this paper

[2] http://universaldependencies.org/

to build on this and transform dependency parses into logical forms (for English) opens up opportunities for crosslingual semantic parsing. While still a field in development, it is one option to be explored if we want to evaluate the semantic transfer in MT. We could foresee that initially at least it could be achieved by developing text cases (see Section 6) on the back of annotations, ensuring that the basic semantics of a sentence in one language (the ST) match that of another (the TT). While ultimately this requires the MT to be of a good standard for parsing, for NMT with a good language pair this is now the case, and indeed has to be for any meaningful attempt to integrate discourse. In the short term, test cases can be devised that do not involve a parser, merely test the ability of a system to effect semantic transfer. In Reddy et al. (2017), they give a concrete example using their semantic interface based on UD for a multilingual question-answering experiment, where they generate ungrounded logical forms for several languages in parallel and map these to Freebase parses which they use for answering a set of standard questions (translated for the German and Spanish). They simplify to ensure crosslingual compatibility, but essentially illustrate how semantic parsing can work crosslingually. For an indepth explanation of the process, see Reddy et al. (2017).

Using these as a test bed and running against WMT systems as additional evaluation could be very useful, perhaps indicating which systems are more capable of capturing and translating the *meaning* of the source. In the long run, ideally the aim is to capture the meaning of the ST, and then based on that generate the TT (a kind of concept-to-text-generation). That would of course involve a shift in paradigm for MT.

6 Evaluation of MT output

Current evaluation methods Hardmeier (2012) already touches on the problem of current evaluation methods. In particular, he mentions the shortcomings of ngram-based metrics and the issue of sentence level evaluation, where much of discourse is document level: "However, it could be argued that the metric evaluation in the shared task itself was biased since the document-level human scores evaluated against were approximated by averaging human judgments of sentences seen out of context, so it is unclear to what extent the evaluation of a document-level score can be trusted." It has to be pointed out that that human evaluation is also not at document level. The problems with BLEU are well illustrated in research by (Smith et al., 2016), proving that optimizing by BLEU scores can actually lead to a drop in quality. However, another major problem is the fact that the evaluation of MT output is still largely based on comparison to a single reference or gold standard translation. A reference, or gold standard translation, is *one* version. A text can be translated in *many* ways, all of which will reflect the translator's interpretation of what the ST is saying. To constrain the measure of correctness to a single reference is only consulting *one* interpretation of the ST. There could be equally good (or better) examples of MT output which are not being scored as highly as they should, simply because they employ a different lexical choice.

Recently, there has also been a trend towards totally ignoring the ST during evaluation of WMT submissions, where 'human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation' (Bojar et al., 2016). So human assessors are asked to rate a given translation by how close it is to the *reference* translation, with no regard to the *source* text. The process is treated as a monolingual direct assessment of translation fluency and adequacy. We would argue that surely adequacy should be based on how well the meaning of the ST has been transferred to the TT, and that to ignore the ST (simply relying on the one rendering of it) is to lose that direct dependency. Whereas a proper measure of adequacy is whether the translation captures and transfers the semantics from ST to TT.

Moreover, the human assessment of the output has recently become 'researcher based judgments only'- which is also problematic, in that the researchers in question are not generally trained in translation, and some are monolingual. This means that they will not necessarily capture discourse information, such as the implicit discourse relations of the reference translation, for example, and know to look for them in the MT output. Not knowing the source language means that you cannot assess the correctness of the output if it alters from the reference.

Moving forward As mentioned by Guzmán et al. (2014), 'there is a consensus in the MT com-

munity that more discourse-aware metrics need to be proposed for this area to move forward'. In terms of evaluation in training, one novel idea is the use of post edits in evaluation (Popović et al., 2016)- this can be seen as more informative and reliable feedback, if done by a human translator, and can be directly used to improve the system. Post edits could also form the basis of test items.

Both Popović (2017); Burchardt et al. (2017) directly or indirectly touch on the issue of evaluation. As part of her analysis Popović (2017) attempts to classify the type of errors made by each system. A most constructive development, Burchardt et al. (2017) introduces a test suite which while it is common and invaluable in software engineering, is not widespread for this domain. With the suite of tests they aim to cover different phenomena, and how the systems handle them, saying they aim to focus on new insights not on how well the systems match the reference (Burchardt et al., 2017).

In the past there have been examples of unit testing for evaluation of MT quality, in particular (King and Falkedal, 1990) who developed theirs for evaluation of different MT systems before financial outlay. Nevertheless, a substantial amount of the logic is still valid: evaluating the strengths and weaknesses of output from various MT systems, with tests focussing on specific aspects (syntactic, lexical ambiguity etc) for particular language pairs.

In a more general vein, Lehmann et al. (1996) develop test suites for NLP in their Test Suites For Natural Language Processing work, for the general evaluation of NLP systems. Their test suites aimed to be reusable, focused on particular phenomena and consisted of a database which could identify test items covering specific phenomena. Similarly, the MT community could potentially develop relevant tests in github, with agreement on format and peer reviews.

This type of method could easily be adopted as a means of evaluation in the context of WMT tasks, and besides being much more informative, would help to pinpoint strengths and weaknesses, leading to more focussed progress. Existing test suites, such as the ones developed by Guillou and Hardmeier (2016) and Loáiciga and Gulordava (2016), could be integrated and added to, giving a more comprehensive and linguistically-based evaluation of system submissions. Unit tests can be added to by interested parties, with peer reviewing if appropriate. The resulting suite could eventually cover a whole host of discourse aspects, and an indication therefore of how different systems perform, and where there is work to be done. The concept is not new, and could build on previous initiatives and experience, such as (Hovy et al., 2002) to ensure it is adaptable yet robust, providing a baseline for progress in particular aspects of discourse.

7 Conclusions

As is clear from the amount of work in Section 2, there has recently been a wealth of research on discourse in MT, which now needs to be integrated, but the incentive to integrate much of it into an MT system is not there while evaluation remains reference-based.

The fact that Martìnez Garcia et al. (2017) found in their recent substantial and innovative research that automatic metrics "are mostly insensitive to the changes introduced by our document-based MT system", is a clear illustration that something is not working. MT is progressing, and evaluation needs to do the same.

There are numerous difficulties with evaluation of discourse phenomena, particularly if it is automatic. But the potential advantages of progressing beyond single reference-based evaluation are considerable– not least the ability to evaluate without first commissioning a reference translation each time. At a time when MT is being used in a pipeline where dialogue acts play an important role, it is vital that evaluation of MT be based on something more substantial than string matching to a single reference, or judgements made without regard for ST. Once MT begins to integrate an element of semantics, it no longer makes sense to evaluate on a single reference. While the translator's role as mediator will not easily be replaced by machines– as yet it cannot capture the pragmatics or recreate the contextual richness for the target audience– nevertheless we must ensure we assess MT output based on a measure of adequacy compared to the *source*, if it is to fulfil its purpose in terms of communication.

Acknowledgments

Many thanks to the anonymous reviewers for their insightful comments and pointers to any omissions.

References

Mona Baker. 1992. In Other Words: A Coursebook on Translation. Routledge.

Viktor Becher. 2011. When and why do Translators add connectives? A corpus-based study. *Target*, volume 23, pages 26–47.

Beata Beigman Klebanov and Michael Flor. 2013. Associative Texture is Lost in Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 27–32. http://www.aclweb.org/anthology/W13-3304.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W/W16/W16-2301.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Charles University, Prague, Czech Republic.

Bruno Cartoni, Andrea Gesmundo, James Henderson, Cristina Grisot, Paola Merlo, Thomas Meyer, Jacques Moeschler, Andrei Popescu-Belis, and Sandrine Zufferey. 2012. Improving MT Coherence Through Text-Level Processing of Input Texts: the COMTIS Project. http://lodel.irevues.inist.fr/tralogy/index.php.

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(2):65–86.

Federico Fancellu and Bonnie Webber. 2015a. Translating Negation: A Manual Error Analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 2–11. http://www.aclweb.org/anthology/W15-1301.

Federico Fancellu and Bonnie Webber. 2015b. Translating Negation: Induction, Search And Model Errors In *Syntax, Semantics and Structure in Statistical Translation*, pages 21–29.

Federico Fancellu and Bonnie Webber. 2014. Applying the semantics of negation to SMT through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, EACL, pages 598–606. http://aclweb.org/anthology/E/E14/E14-1063.pdf.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 52–58. http://www.aclweb.org/anthology/W/W15/W15-2504.pdf.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 1–10. http://dl.acm.org/citation.cfm?id=2380943.2380944.

Liane Guillou. 2013. Analysing Lexical Consistency in Translation. In *Proceedings of the Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 10–18. http://www.aclweb.org/anthology/W13-3302.

Liane Guillou. 2016. Incorporating Pronoun Function into Statistical Machine Translation. Ph.D. thesis, University of Edinburgh.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany. Association for Computational Linguistics*. Berlin, Germany.

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, 2014, June 22-27, 2014, Baltimore, MD,*

USA, Volume 1: Long Papers. The Association for Computer Linguistics, pages 687–698. http://aclweb.org/anthology/P/P14/P14-1065.pdf.

Christian Hardmeier. 2012. Discourse in Statistical Machine Translation. *Discours 11-2012* (11).

Christian Hardmeier. 2014. Discourse in Statistical Machine Translation. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.

Christian Hardmeier. 2015. On Statistical Machine Translation and Translation Theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 168–172. http://aclweb.org/anthology/W15-2522.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013a. Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *51st Annual Meeting of the Association for Computational Linguistics, 2013, Proceedings of the Conference System Demonstrations, 4-9 August 2013, Sofia, Bulgaria*. pages 193–198. http://aclweb.org/anthology/P/P13/P13-4033.pdf.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013b. Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 380–391. http://www.aclweb.org/anthology/D13-1037.

Basil Hatim and Ian Mason. 1990. *Discourse and the translator*. Longman.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2015. The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 41–46. http://aclweb.org/anthology/W15-2505.

Eduard Hovy, Margaret King, and Andrei Popescu-Belis. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation* 17(1):43–75. http://www.jstor.org/stable/40008209.

Margaret King and Kirsten Falkedal. 1990. Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '90, pages 211–216. https://doi.org/10.3115/997939.997976.

Alistair Knott and Robert Dale. 1993. Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations.

Ekaterina Lapshinova-Koltunski. 2015. Exploration of Inter-and Intralingual Variation of Discourse Phenomena. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 158–167. http://aclweb.org/anthology/W15-2521.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervè Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP: Test Suites for Natural Language Processing. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '96, pages 711–716. https://doi.org/10.3115/993268.993292.

Jessy Junyi Li, Marine Carpuat, and Ani Nenkova. 2014a. Cross-lingual Discourse Relation Analysis: A corpus study and a semi-supervised classification system. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, pages 577–587. http://aclweb.org/anthology/C14-1055.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014b. Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,pages 283–288. http://aclweb.org/anthology/P14-2047.

Sharid Loáiciga. 2015. Predicting Pronoun Translation Using Syntactic, Morphological and Contextual Features from Parallel Data. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 78–85. http://aclweb.org/anthology/W15-2511.

Sharid Loáiciga and Kristina Gulordava. 2016. Discontinuous Verb Phrases in Parsing and Machine Translation of English and German. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*

Sharid Loáiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *The Ninth Language Resources and Evaluation Conference*. EPFL-CONF-198442.

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

Ngoc-Quang Luong and Andrei Popescu-Belis. 2016. A Contextual Language Model to Improve Machine Translation of Pronouns by Re-ranking Translation Hypotheses. *Baltic Journal of Modern Computing* 4(2):292.

Ngoc-Quang Luong and Andrei Popescu-Belis. 2017. Machine Translation of Spanish Personal and Possessive Pronouns Using Anaphora Probabilities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, EPFL-CONF-225949.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The Automatic Translation of Discourse Structures. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. Stroudsburg, PA, USA, NAACL 2000, pages 9–17. http://dl.acm.org/citation.cfm?id=974305.974307.

Eva Martìnez Garcia, Carles Creus, Cristina España Bonet, and Lluís Màrquez. 2017. Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Charles University, Prague, Czech Republic.

Laura Mascarell, Mark Fishel, Natalia Korchagina, and Martin Volk. 2014. Enforcing Consistent Translation of German Compound Coreferences. In *KONVENS*. pages 58–65.

Laura Mascarell, Mark Fishel, and Martin Volk. 2015. Detecting Document-level Context Triggers to Resolve Translation Ambiguity. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 47–51. http://aclweb.org/anthology/W15-2506.

Thomas Meyer. 2011. Disambiguating TemporalContrastive Discourse Connectives for Machine Translation. In *Proceedings of the ACL 2011 Student Session*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-SS '11, pages 46–51. http://dl.acm.org/citation.cfm?id=2000976.2000985.

Thomas Meyer and Lucie Poláková. 2013. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*. Sofia, Bulgaria, page 8.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL 2012, pages 129–138.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *SIGDIAL Conference*. The Association for Computer Linguistics, pages 194–203.

Ruslan Mitkov. 1993. How Could Rhetorical Relations be used in Machine Translation. In *Proceedings of the ACL Workshop on Intentionality and Structure in Discourse Relations*.

Eugene A. Nida and C.R. Taber. 1969. *The Theory and Practice of Translation*. E. J. Brill, Leiden.

Michal Novák. 2016. Pronoun Prediction with Linguistic Features and Example Weighing. In *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 602–608.

Michal Novák and Zdeněk Žabokrtský. 2014. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 14–24.

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*. Stroudsburg, PA, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Maja Popović, Mihael Arčan, and Arle Lommel. 2016. Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*. Riga, Latvia.

Maya Popović. 2017. Comparing Language Related Issues for NMT and PBMT between German and English. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Charles University, Prague, Czech Republic.

Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics* 4.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal Semantic Parsing. *arXiv preprint arXiv:1702.03196* .

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2015. A Proposal for a Coherence Corpus in Machine Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 52–58. https://aclweb.org/anthology/W/W15/W15-2507.pdf.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. The Trouble with Machine Translation Coherence. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*.

Aaron Smith, Christian Hardmeier, and Jörg Tiedemann. 2016. Climbing Mount BLEU: The Strange World of Reachable High-BLEU Translations. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA, USA.

David Steele. 2015. Improving the Translation of Discourse Markers for Chinese into English. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Denver, Colorado, NAACL, pages 110–117. http://aclweb.org/anthology/N/N15/N15-2015.pdf.

David Steele and Lucia Specia. 2016. Predicting and Using Implicit Discourse Elements in Chinese-English Translation. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*. Riga, Latvia, pages 305–317.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical Machine Translation with Readability Constraints. In *Proceedings of NODALIDA*. pages 375–386.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A Novel Translation Framework Based on Rhetorical Structure Theory. In *The Association for Computer Linguistics*, pages 370–374.

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. http://aclweb.org/anthology/W15-25.

Bonnie Webber, Markus Egg, and Vali Kordoni. 2012. Discourse Structure and Language Technology. *Natural Language Engineering* 18(4):437–490.

Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jorg Tiedemann. 2013. *Proceedings of the ACL Workshop on Discourse in Machine Translation (DiscoMT 2013)*. Association for Computational Linguistics. http://www.aclweb.org/anthology-new/W/W13/#3300.

Dominikus Wetzel and Francis Bond. 2012. Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Jeju, Republic of Korea, SSST-6, pages 20–29. http://dl.acm.org/citation.cfm?id=2392936.2392940.

Dominikus Wetzel, Adam Lopez, and Bonnie Webber. 2015. *A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction*, Association for Computational Linguistics, pages 115–121.

Billy Tak-Ming Wong and Chunyu Kit. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion To Document Level. In *Proceedings of EMNLP-CoNLL*. pages 1060–1068.

Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lv, and Qun Liu. 2013a. Modeling Lexical Cohesion for Document-Level Machine Translation. In *Proceedings of IJCAI*.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. 2013b. Lexical Chain Based Cohesion Models for. Document-Level Statistical Machine Translation. In *Proceedings of EMNLP*. pages 1563–1573.

Deyi Xiong and Min Zhang. 2013. A Topic-Based Coherence Model for Statistical Machine Translation. In *Proceedings of AAAI*. pages 977–983.

Deyi Xiong and Min Zhang. 2014. A Sense-Based Translation Model for Statistical Machine Translation. In *Association for Computational Linguistics*. pages 1459–1469.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015. Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 142–152. http://aclweb.org/anthology/W15-2519.

Rong Zhang and Abraham Ittycheriah. 2015. Novel Document Level Features for Statistical Machine Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal.

Sandrine Zufferey and Andrei Popescu-Belis. 2017. Discourse connectives: theoretical models and empirical validations in humans and computers. In *Formal Models in the Study of Language*, Springer International Publishing, pages 375–390.