A Proposal for a Coherence Corpus in Machine Translation

Karin Sim Smith[§], **Wilker Aziz**[†] and **Lucia Specia**[§]
§Department of Computer Science, University of Sheffield, UK
{kmsimsmith1,l.specia}@sheffield.ac.uk

†Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
w.aziz@uva.nl

Abstract

Coherence in Machine Translation (MT) has received little attention to date. One of the main issues we face in work in this area is the lack of labelled data. While coherent (human authored) texts are abundant and incoherent texts could be taken from MT output, the latter also contains other errors which are not specifically related to coherence. This makes it difficult to identify and quantify issues of coherence in those texts. We introduce an initiative to create a corpus consisting of data artificially manipulated to contain errors of coherence common in MT output. Such a corpus could then be used as a benchmark for coherence models in MT, and potentially as training data for coherence models in supervised settings.

1 Introduction

Discourse information has only recently started to attract attention in MT, particularly in Statistical Machine Translation (SMT), the focus of this paper. Most decoders work on a sentence by sentence basis, isolated from context, due to both modelling and computational complexity. An exception are approaches to multi-pass decoding, such as Docent (Hardmeier et al., 2013a). Our work focuses on an issue which has not yet been much explored in MT, that of coherence.

Coherence is undeniably a cognitive process, and we will limit our remit to the extent that this process is guided by linguistic elements discernible in the discourse. While it does include cohesion, it is wider in terms of describing how a text becomes semantically meaningful overall, and additionally spans the entire document. We are interested in capturing aspects of coherence as defined by Grosz and Sidner (1986), based on the attentional state, intentional structure and linguistic structure of discourse. As a result, we believe that a coherent discourse should have a context and a focus, be characterised by appropriate coherence relations, and structured in a logical manner.

Previous computational models for assessing coherence in a monolingual context have covered entity transitions (Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Burstein et al., 2010; Guinaudeau and Strube, 2013), syntactic patterns (Louis and Nenkova, 2012), discourse relations (Lin et al., 2011), distributed sentence representations (Li and Hovy, 2014) and lexical chains (Somasundaran et al., 2014). For evaluation, these studies in coherence have typically used automatically summarized texts, or texts with sentences artificially shuffled as their 'incoherent' data. The latter is an example of artificially created labelled data, distorting the ordered logic of the text and thus affecting some aspects of coherence. However, it is inadequate for our task, as MT preserves the sentence ordering, but suffers from other aspects of incoherence. Moreover, while the MT output can potentially be considered 'incoherent', it contains a multitude of problems, which are not all due to lack of coherence.

For the evaluation of coherence models in the MT context, as well as for supervised learning of coherence models it is necessary to have data annotated with issues of incoherence. In particular, we are interested in coherence issues which are deemed to occur specifically in MT output. The purpose of this initiative is to ensure that we can assess coherence models by isolating other issues that are not related to coherence.

In the remainder of this paper, we start by presenting previous work (Section 2). We then describe how problems related to lack of coherence are manifested in MT output (Section 3). In Section 4 we detail how we plan to manipulate the data in systematic ways to create a corpus of artificially generated incoherent data.

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 52–58,
Lisbon, Portugal, 17 September 2015.

2 Existing work

There has been previous work in the area of lexical cohesion in MT (Wong and Kit, 2012; Xiong et al., 2013a; Xiong et al., 2013b; Tiedemann, 2010; Hardmeier, 2012; Carpuat and Simard, 2012). Lexical cohesion is part of coherence, as it looks at the linguistic elements which hold a text together. However, there has been very little work in the wider area of coherence as a whole.

Besides lexical cohesion, another discourse related phenomenon that has been addressed in MT is reference resolution. As detailed in greater depth by Hardmeier (2012), the results for earlier attempts to address this issue were not very successful (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010). More recent work includes that of Guillou (2012), which highlights the differences of coreference depending on the language pair. Since then Hardmeier et al. (2013b) have used a new approach for anaphora resolution via neural networks which achieves comparable results to a standard anaphora resolutions system, but without annotated data. Recently work has begun on negation in MT, particularly by Wetzel and Bond (2012; Fancellu and Webber (2014). There is also work focusing on evaluation against reference translations (Guzmán et al., 2014) based on the comparison between discourse trees in MT versus reference. This information was found to improve evaluation of MT output.

Drawing from research on topic modelling (Eidelman et al., 2012), where lexical probabilities conditioned on topics are computed, Xiong and Zhang (2013) attempt to improve coherence based using topic information. They determine the topic of the source sentence and project it onto the target as a feature to ensure the decoder selects the appropriate words. They observed slight improvements in terms of general standard metrics, indicating perhaps that these metrics fail to account for discourse improvements.

As far as we aware, no attempts have been made to create a corpus exhibiting incoherence, other than by shuffling ordered sentences. There has been work in other areas to introduce errors in correct texts. For example, Felice and Yuan (2014) and Brockett et al. (2006) inject grammatical errors common to non-native speakers of English in good quality texts. Felice and Yuan (2014) use existing corrected corpora to derive the error distribution, while Brockett et al. (2006) adopt a deterministic approach based on hand-crafted rules. Logacheva and Specia (2015) inject various types of errors to generate negative data for quality estimation purposes, but these are at the word level, and the process was guided by post-editing data. They derived an error distribution of MT output by inspecting post editing data. We do not have a similar way of inducing a distribution of errors for coherence. A large amount of post editings of entire documents would be needed, and it still be difficult to isolate which of the edits relate to coherence errors.

3 Issues of incoherence in MT systems

Current MT approaches suffer from a lack of linguistic information at various stages (modelling, decoding, pruning) causing the lack of coherence in the output. Below we describe a number of issues that are generally viewed as coherence issues which MT approaches deal poorly with and which have also been the subject of previous work. The examples given have been identified in error analysis done by ourselves in either of the following corpora:

- the **newstest** data (source and output) from the WMT corpus,[1] focusing on French and German source, and English as output.
- the **LIG** corpus (Potet et al., 2012) of French-English translations: 361 parallel documents comprising source, reference translation, machine translated output and post-edited output, drawn from various WMT editions.

The following are issues of incoherence which have been identified by ourselves (below) and others (Section 2) as particularly common in MT systems.

Lexical cohesion MT has been shown to be be consistent in its use of terminology (Carpuat and Simard, 2012), which can be an advantage for narrow texts domains with significant training data. However, MT systems may output direct translations of source text items that may be inappropriate in the target context. Moreover, while a specific target text word may correctly translate a source text word in one context, it may require a totally different word in another context. In our data *'boucher'* occur

[1] http://www.statmt.org/wmt13/

more often as a French noun, corresponding to *'butcher'*. This increases the probability of the translation equivalence *'butcher'*, yet in the translated text it is used as a noun indicating to *'block'* (for example, *'road block'*).

src: *'Cette anne, c'était **au tour de** l'Afrique de nommer le président et elle a nommé la Libye.'*

mt: *'This year, it was **at the tour of** Africa to appoint the president and has appointed Libya.'*

ref: *'This year it was Africa**'s turn** to nominate the chairman, and they nominated Libya.'*

Here the wrong meaning of *tour* was used, and renders the sentence incoherent. As Wong and Kit (2012) note, the lexical cohesion devices have to not only be recognised, but used appropriately. And this may differ from the source text to the target text.

Referencing Anaphora resolution is a very challenging issue in current MT approaches (Michal, 2011; Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Hardmeier et al., 2013b; Guillou, 2012). This is again due to the fact that inter-sentential references are lost in most decoders as they translate one sentence at a time. Reference resolution is 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.

src: *'L'extrême droite européenne est caractérisée par **son** racisme...'*

mt: *'The extreme right is characterised by **his** racism...'*

ref: *'A common feature of Europe's extreme right is its racism...'* (Potet et al., 2012).

Here the pronoun *'son'*, referring to the racism of the extreme right, is wrongly rendered as *'his'*.

Discourse connectives Discourse connectives are vital for the correct understanding of discourse. Yet in MT systems these can be incorrect or missing (Meyer and Poláková, 2013; Meyer and Popescu-Belis, 2012; Meyer et al., 2011; Steele, 2015). In particular, where discourse connectives are ambiguous, e.g. those which 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 need to be inferred. While a human translator can detect this, an MT system cannot.

src: *'Die Rechtsanwlte der Republikaner haben in 10 Jahren in den USA **übrigens** nur 300 Flle von Wahlbetrug verzeichnet.'*

mt: *'The Republican lawyers have listed over 10 years in the United States, only 300 cases of electoral fraud.'*

ref: ***Indeed**, Republican lawyers identified only 300 cases of electoral fraud in the United States in a decade.*

The discourse marker is missing altogether in the MT output above (in addition to the ordering error). While small, cue words guide the reader and help create the logic in the text. Here the discourse marker was for emphasis, illustrating the writer's claim.

Syntax structure Different languages have different syntactic structures. In MT system the syntax of the target language may get distorted, often too close to the syntax of the source language, leading to an incoherent sentence formation.

src: *'Ce ne sera pas le cas, comme le démontre clairement l'histoire raciale de l'Amérique.'*

mt: *'This is not the case, as clearly demonstrates the history of race in America.'*

ref: *It will not, as America's racial history clearly shows.* (Potet et al., 2012)

Here the natural logic of the sentence is distorted, with the subject coming after the verb, directly affecting the coherence.

Clauses ordering Particularly in hierarchical or tree-based MT systems, the order of clauses within sentences may have become reversed, or may be unnatural for the target language.

src: *'Das Opfer war später an den Folgen der schweren Verletzungen gestorben.'*

mt: *'The victim was later at the consequences of the serious injuries died.'*

ref: *'The victim later died as a result of the serious injuries.'* (Bojar et al., 2014).

This can affect the understanding of the sentence, the overall logic of it in the context of the surrounding sentences, or simply require a reread which itself is indicative of impaired coherence.

src: *'Bereits im Jahr 1925 wurde in Polen eine Eisenbahn-Draisine gebaut, für die ein Raketenantrieb geplant war. Der Autor des Entwurfs und die Details dieses Vorhabens blieben leider unbekannt.'*

mt: *'Already in 1925 a railway trolley was built in Poland, for which a rocket was planned. The author of the design and the details of the project remained unfortunately unknown.'*

ref: *In 1925, Poland had already built a handcar which was supposed to be fitted with a rocket engine. Unfortunately, both the project's designer, and the project's details, are unknown.* (Bojar et al., 2013)

The reference translation has a clausal pattern which is more cohesive to the English reader.

Negation MT systems often miss the focus of the negation. This results in incorrectly transferred negations that affect coherence (Wetzel and Bond, 2012; Fancellu and Webber, 2014).

src: *'**Aucun** dirigeant serbe **n**'acceptera l'indépendance du Kosovo'*

mt: *'**No** leader of Serbia will **not** accept the independence of Kosovo..'*

ref: *'No leader of Serbia will accept the independence of Kosovo'*.(Potet et al., 2012)

In this case the negation is distorted, influenced by the structure of the source text.

4 Artificially generating coherence errors

Significant work has already been done in the areas of coreference resolution (Michal, 2011; Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Hardmeier et al., 2013b; Guillou, 2012) and negation (Wetzel and Bond, 2012; Fancellu and Webber, 2015; Fancellu and Webber, 2014) in MT. In our corpus we will focus on less studied issues and limit ourselves to targeting coherence more specifically than cohesion.

The proposed framework will take as input well-formed documents that are determined 'coherent' (i.e. grammatically correct and coherent) and then artificially distort them in ways (detailed below) that directly affect coherence in the manner that an MT system would. The resulting texts will make a corpus of 'incoherent' texts for assessing the ability of models to discriminate between coherent and incoherent texts.

This will be done in a flexible manner, such that the incoherent documents can be created for a variety of (coherent) input texts. Moreover they can be created for specific types of errors. The quality of MT output varies greatly from one language pair and MT system to another. For example, the output from a French-English MT system trained in very large collections is superior to that of, for example, an English-Finnish system trained on smaller quantities of data (Koehn and Monz, 2005; Bojar et al., 2015).The errors encountered also vary, depending on the language pair, in particular for aspects such as discourse markers and syntax. Some of these error patterns are more relevant for particular language pairs, e.g. negation for French-English, which is otherwise a well-performing language pair.

We propose to inject errors programmatically in a systematic manner, as detailed below.

4.1 Error distribution

While ideally we would establish the distribution of errors from their occurrences in MT output, determining an appropriate error distribution based on observations is very problematic. The distributions would be specific to given language pairs and MT systems. More important, detecting coherence automatically to count errors is difficult: if we could do that, than we would be able to directly solve the problem we are attempting to, i.e. measure coherence. This is exactly why we need this corpus. Additionally, manual inspection and annotation for coherence is very hard to formalise as a task, time consuming and costly. Therefore, the distribution of errors in our corpus will be based on linguistic insights, and on findings from previous work, where available. Where this is not the case, for instance for distorting discourse patterns, versions of the corpus with different proportions of errors will be created. We will inject errors systematically and incrementally to vary the degree and location of the errors.

The errors will be introduced systematically via pattern-matching, and as highlighted by Brockett et al. (2006), may not be distributed in a natural way.

4.2 Error Injection

We will inject errors of the types below via the four basic edit operations, as appropriate for each type of error: replace, delete, add, shift.

Sentence level discourse structure We will inject errors related to discourse elements, in terms of cue words, and their organisation. A comparison of the discourse connectives in the MT and the Human Translation (HT) will be established, and where these differ, a syntactic check is made automatically (Pitler and Nenkova,) to establish if the connective is a synonym or incorrect. We can also refer to the discourse connectives in the original source text, and automatically check, for example, if the correct sense of the connective has been transferred. These can be identified from a list compiled from appropriate resources (e.g. DiMLex for German, LexConn for French)(Stede and Umbach, 1998; Roze and Danlos, 2012) and a list of problematic ones derived e.g. from work by (Meyer and Popescu-Belis, 2012; Meyer and Poláková, 2013) for French.

We can parse the discourse tree structure and extract grammatical information using the Stanford parser[2] and POS tagger[3], before distorting the parse tree by swapping nodes at the relevant level.

Lexical cohesion We propose replacing entities with alternatives (which will directly affect lexical coherence), using phrase tables from an MT system to generate likely entity variations. This has to be tailored to ensure that the result reflects realistic error levels, so need to verify correct parameter to gauge the amount of substitutions. We can also investigate pre-trained word embeddings, such as word2vec representations (Mikolov et al., 2013), and using word intrusion detection (Chang et al., 2009).

Clausal patterns Coherent syntax patterns can be derived from coherent text, for example using patters established in (Louis and Nenkova, 2012). We can determine the clausal patterns from training data, establishing frequent patterns which are indicative of specific coherence relations. Then the order of sibling nodes in the syntax tree can be modified (e.g. reversed) at the appropriate level in order to alter the order of clauses. The exact level of the distortion will be determined according to pre-defined criteria – e.g. every 8th clause, to depth 5 in the parse tree or, where possible, derived from the MT output.

[2]http://nlp.stanford.edu/software/lex-parser.shtml
[3]http://nlp.stanford.edu/software/tagger.shtml

5 Conclusion

We have introduced our initiative for artificially generating a corpus with coherence errors from well-formed data that specifically simulate coherence issues in MT.

Other possible direction could be to use an n-best list, taking sentences from different positions in that list for each source sentence to form a possibly incoherent document. Similarly, we could extract sentences from multiple MT systems for the same text, alternating their origin and concatenating to form one single document. In both cases, a difficulty that remains is that of isolating coherence issues from other errors and from stylistic issues, as well as quantifying the degree of incoherence in the generated texts.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Comput. Linguist.*, 34(1):1–34.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of WMT*, pages 12–58, Baltimore, Maryland.

Ondrej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, , Philipp Koehn, , Christof Monz, Matteo Negri, Pavel Pecina, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, WMT, Lisbon, Portugal.

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting esl errors using phrasal smt techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 249–256.

Jill Burstein, Joel R. Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Proceedings of the Con-*

ference of the North American Chapter of the Association for Computational Linguistics, pages 681–684.

Marine Carpuat and Michel Simard. 2012. The trouble with smt consistency. In *Proceedings of WMT*, pages 442–449, Montreal, Canada.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 115–119.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of ACL*, pages 125–129.

Federico Fancellu and Bonnie L. 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*, EACL, pages 598–606, Gothenburg, Sweden.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado, June. Association for Computational Linguistics.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Comput. Linguist.*, 12(3):175–204, July.

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*, EACL '12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of ACL*, pages 93–103.

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, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 687–698. The Association for Computer Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT*, pages 283–289.

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

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*, pages 380–391, Seattle, Washington.

Christian Hardmeier. 2012. Discourse in statistical machine translation. *Discours 11-2012*, (11).

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124.

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.

Jiwei Li and Eduard H. Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2048, Doha, Qatar.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of ACL*, pages 997–1006.

Varvara Logacheva and Lucia Specia. 2015. The role of artificially generated negative data for quality estimation of machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 51–58, Antalya, Turkey.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of EMNLP-CoNLL*, pages 1157–1168, Jeju Island, Korea.

Thomas Meyer and Lucie Poláková. 2013. Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013*, Sofia, Bulgaria.

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)*, pages 129–138, Avignon, France.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *SIGDIAL Conference*, pages 194–203. The Association for Computer Linguistics.

Novák Michal. 2011. Utilization of anaphora in machine translation. In *WDS Week of Doctoral Students*, June.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore.

Marion Potet, Emmanuelle Esperança-rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a large database of french-english smt output corrections.

Charlotte Roze and Laurence Danlos. 2012. Lexconn: a french lexicon of discourse connectives. *Discours*.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING*.

Manfred Stede and Carla Umbach. 1998. Dimlex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL, pages 1238–1242, Montreal, Quebec.

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*, NAACL, pages 110–117, Denver, Colorado.

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.

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*, SSST-6, pages 20–29, Jeju, Republic of Korea.

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 and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *Proceedings of AAAI*, pages 977–983.

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.