Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 178–189

The Trouble with Machine Translation Coherence

Karin SIM SMITH[§], Wilker AZIZ[†], Lucia SPECIA[§]

§Department of Computer Science, University of Sheffield, UK
†Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

{kmsimsmith1,l.specia}@sheffield.ac.uk , w.aziz@uva.nl

Abstract. This paper introduces the problem of measuring coherence in Machine Translation. Previously, local coherence has been assessed in a monolingual context using essentially coherent texts. These are then artificially shuffled to create an incoherent one. We investigate existing models for the task of measuring the coherence of machine translation output. This is a much more challenging case where coherent source documents are machine translated into a target language and the task is to distinguish them from their human translated counterparts. We benchmark state-of-the-art coherence models, and propose a new model which explores syntax following a more principled method to learn the syntactic patterns. This extension outperforms existing ones in the monolingual shuffling task on news data, and performs well in our new, more challenging task. Additionally, we show that breaches in coherence in the translation task are much more difficult to capture by any model.

Keywords: Machine Translation, Discourse, Coherence Modelling

1 Introduction

A coherent discourse is said to be one that has meaningful connections between its utterances (Jurafsky and Martin, 2009). The task of automatically evaluating text coherence has been addressed within applications such as text summarisation and ordering (Lapata, 2005; Barzilay and Lapata, 2008), where shuffled sentences or inadequate summaries can lead to less coherent documents. Coherence has then been measured with entity grids, discourse relations and syntax patterns, with experiments run on the original and the artificially modified texts to distinguish coherent from incoherent texts.

We introduce the more challenging problem of evaluating the coherence of documents generated by Machine Translation (MT) systems. This is a very different scenario. Firstly, it is more subtle as the sudden breaks in transitions or shifts of focus which result from shuffling in the traditional monolingual test scenario are absent, as sentences are translated in their original (source) order. Secondly, the machine translated output may contain other textual issues, such as ungrammatical fragments, which

can affect the application of such models in various ways, making it harder to pinpoint coherence-related problems. Finally, judgements on the coherence of the translations may be dependent on the source text. Nevertheless, measuring coherence in MT is important: given the way translations are generated by standard MT systems, on a sentence-by-sentence basis, several phenomena spanning sentence boundaries can lead to incoherent document translations, such as incorrect co-referencing, inadequate discourse markers, and lack of lexical cohesion, as established by previous corpus analyses (Sim Smith et al., 2015).

We apply three existing coherence models to original, shuffled and machine translated texts in an attempt to evaluate their ability to discriminate between coherent and incoherent documents: an entity-grid model (Barzilay and Lapata, 2008), an entity graph similarity metric (Guinaudeau and Strube, 2013), and a model based on syntactic patterns (Louis and Nenkova, 2012). In addition, we propose a fully generative extension of the syntax-based coherence model. We illustrate the difference between assessing the output from MT systems and assessing the coherence of shuffled texts in a highly consistent, structured corpus.

The remainder of this paper, is organised as follows: in Section 2 we review related work on coherence and cohesion in the context of MT. Section 3 covers the background of the coherence models used in this paper. Experiments and results are discussed in Section 4.

2 Related Work

There has been recent 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), as a sub-category of coherence, looking at the linguistic elements which hold a text together. However, there seems to be little work in the wider area of coherence as a whole. Coherence is indeed a more complex discourse element to define in the first place. While it does include cohesion, it also describes how a text becomes semantically meaningful overall, and how easy it is for the reader to follow.

Louwerse (2005) defines "cohesion as continuity in word and sentence structure, and coherence as continuity in meaning and context". While lexical cohesion can be detected and addressed to some extent, the semantics, meaning and contextual indicators necessary for coherence assessment are much more difficult to capture, even though judging coherence is an intuitive process for a human reader. Coherence is undeniably a complex cognitive process, which is however guided by elements of discourse that we believe can be modelled automatically to some extent.

Most previous computational models for assessing coherence have focused on entity transitions, syntactic patterns and discourse relations. The most popular models are detailed in Section 3. In what follows we describe these models, and our work to apply these models to MT. Lin et al. (2011) evaluate the coherence of texts from discourse role transitions in a grid-based model, on the basis that there is a preferential, canonical, ordering of discourse relations that leads to coherent texts. Burstein et al. (2010) use the entity-grid for student essay evaluation, which is a scenario closer to ours. They used a range of additional features specifically targeting grammar and style. These proved

useful for discriminating good from bad quality essays, but it is unclear how much of the problem with low quality essays was due to coherence issues. Their features are not publicly available for us to assess this.

Somasundaran et al. (2014) consider how lexical chains affect discourse coherence. They use lexical chaining features such as length, density, and link strength to detect textual continuity, elaboration, lexical variety and organisation, all vital aspects of coherent texts. They claim that the interaction between lexical chains and discourse cues can also show whether cohesive devices are organised in a coherent fashion.

Recently, Li and Hovy (2014) developed a coherence model based on distributed sentence representation. They used recurrent and recursive neural networks to perform ordering and readability tasks. They leverage semantic representations to establish coherent orderings, using original texts as positive examples and shuffled versions as negative ones for optimising the neural networks.

Li et al. (2015) train a hierarchical Long-Short Term Memory (LSTM) to explore neural Natural Language Generation, and assess whether local semantic and syntactic coherence can be represented at a higher level, namely paragraphs. In their model, different LSTM layers represents word embeddings, sentences, and paragraphs. They are then able to regenerate the text to a degree that indicates neural networks are able to capture certain elements of coherence.

Lin and Li (2015) use a hierarchical recurrent neural network language model (RNNLM) to combine a word-level model with a sentence-level model for document modeling. They claim that their model captures both intra- and inter-sentential sequences. They assess their model on an MT reranking task, progressively reranking consecutive sentences. In the MT domain, Xiong et al. (2013) attempt to improve lexical coherence with a topic-based model. They extract a coherence chain for the source sentence, and project it onto the target sentence to try and make lexical choices taken 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. 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.

While the task of automatically evaluating text coherence has been addressed previously within applications such as multi-document text summarisation or in terms of optimal ordering within shuffled texts, our aim is to further investigate these components in an MT context without the use of a reference translation. We ultimately expect to be able to bias the translation process to ensure coherence in MT.

3 Coherence Models

Here we describe some of the most popular coherence models, all of which we reimplement and test in our experiments, as well as our improvement over a syntax-based model (Section 3.4).

3.1 Entity-grid approach

The entity-based approach (Lapata, 2005; Barzilay and Lapata, 2008), in particular the Centering Theory (Grosz et al., 1995) it is based on, derives from the idea that entities in a coherent text are distributed in a certain manner. This theory states that coherent texts are characterised by salient entities in strong grammatical roles, such as subject or object. The focus of the entity-based approach uses this knowledge, via patterns in terms of prominent syntactic constructions, to distinguish coherent from non-coherent texts. Entity grids are constructed by identifying the discourse entities in the documents under consideration and constructing a 2D grid for each document, whereby each column corresponds to the entity, i.e. noun being tracked, and each row represents a particular sentence in the document.

An **entity transition** is defined as a consecutive occurrence of an entity with a given syntactic role, namely, subject (S), object (O), or other (X). Absences of entities in sentences, or nulls, are recorded with a dash. Transitions are observed by examining the grid vertically for each entity. The assumption is that incoherent texts have more breaks in the entity transitions, and thus lower scores.

Lapata (2005) introduces a generative model of document coherence based on entity transitions. Equation 1 shows this formulation, where m is the number of entities, n is the number of sentences in a document D, and $r_{s,e}$ is the role taken by entity e in sentence s. Observe that the model makes a Markov assumption, under which an entity's role is independent of all but its h preceding roles.

$$p(D) = \frac{1}{m \cdot n} \prod_{e=1}^{m} \prod_{s=1}^{n} p(r_{s,e} | r_{(s-h),e} \cdots r_{(s-1),e}) \tag{1}$$

Our objective with this model, as with all others in this paper, is to assess whether the coherence model allows us to discriminate between Human Translation (HT) and MT.

For our experiments, a POS tagger[1] is used to identify nouns and subsequently a parser[2] is used to establish the grammatical role of each of these nouns. The original model presumes that grids of coherent texts have a few dense columns and many sparse ones, and that entities occurring in the dense columns are more often be subjects or objects. It assumes that these characteristics are less common in texts exhibiting lower coherence (Lapata, 2005). In our experiments, the MT displays no more sparse columns than the reference counterpart. It would seem that given how preeminent the focused nouns are, these are captured in the MT output. There are, however, differences in transition patterns, in that some patterns are more common in the MT than the HT, such as 'OO', or other patterns with strong object positions. This seems to indicate a more simplistic style by MT systems. Quantitative results for the experiments with the entity-grid model are given in Section 4.

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

3.2 Entity graph approach

Guinaudeau and Strube (2013) adapted the entity-grid into a graph format using a bipartite graph which they claim avoids the data sparsity issues encountered by Barzilay and Lapata (2008) and achieves equal performance, without training. Additionally, their representation can track any cross-sentential references, as opposed to only those present in adjacent sentences.

The graph tracks the presence of all entities and connections to the sentences they occur in, taking all nouns in the document as discourse entities, as recommended by Elsner and Charniak (2011). The coherence of a text in this model is measured by calculating the average outdegree of a projection, summing the shared edges.

The general form of the coherence score assigned to a document in this approach is shown in Equation 2. This is a centrality measure based on the average outdegree across the N sentences represented in the document graph. The outdegree of a sentence s_i, denoted $o(s_i)$, is the total weight of edges leaving that sentence, a notion of how connected (or how central) it is. This weight is the sum of the contributions of all edges connecting s_i to any $s_j \in D$.

$$s(D) = \frac{1}{N}\sum_{i=1}^{N} o(s_i) \qquad = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} W_{i,j} \tag{2}$$

We reimplemented the algorithm in Guinaudeau and Strube (2013) (using syntactic projection) and ran experiments with the same objective and datasets as for the grid model. Quantitative results for the experiments with the entity-graph model are given in Section 4.

We have also experimented with other languages and noted that syntactic differences do indeed change the transition parameters. This varies depending on the language pair. In particular, it has been proven that the same patterns of syntactic constructions do not hold for German, for example, where topological fields are more relevant. We therefore limit ourselves to reporting results on English.

3.3 Syntax-based model

Motivated by the strong impact syntax has in text coherence, Louis and Nenkova (2012) propose both a local and a global coherence model based on syntactic patterns. Our implementation focuses on their local coherence model. It follows the hypothesis that in a coherent text consecutive sentences will exhibit syntactic regularities, and that these regularities can be captured in terms of co-occurrence of syntactic items.

The units of syntax can be context-free grammar productions (e.g. S $\rightarrow$ NP VP) or d-sequences (a sequence of sibling constituents at depth d starting from the root, possibly annotated with the left-most child node they dominate, e.g. NP_{NN} VP_{VB}). The model conditions each sentence on the immediately preceding sentence, both seen as sequences of syntactic patterns. Each sentence is assumed to be generated one pattern at a time and patterns are assumed independent of each other.

The parameters of the model are "unigram" and "bigram" patterns over a vocabulary of syntactic items (i.e. productions or d-sequences) which are directly observed from

training data by relative frequency counting.

$$p(D) = \prod_{(u_1^m, v_1^n) \in D} \prod_{j=1}^{n} \frac{1}{m} \sum_{i=1}^{m} \frac{c(u_i, v_j) + \alpha}{c(u_i) + \alpha|V|} \tag{3}$$

The coherence of a document under the model is given by Equation 3, where (u_1^m, v_1^n) represents adjacent sentences, and $c(\cdot)$ is a function that counts how often a pattern (or a pair of patterns) was observed in the training data. To account for unseen syntactic patterns at test time, their model is smoothed by a constant α (and $|V|$ is the size of the vocabulary of syntactic tokens).

In our experiments, we derived the syntactic items in the form of the d-sequence, defined as the leaves of the parse tree at a given depth (in our experiments of depth 2, 3, 4), and annotated with the left-most leaf. The choice of d-sequences results in what we believe to be an informative representation. Further experiments could use grammatical productions as an alternative.

3.4 Syntax-based model with IBM 1

The syntax model by Louis and Nenkova (2012) does not model latent alignments. This is possible under the assumption that all available alignment configurations have been directly observed in the training data. It is worth highlighting that in reality the training data is incomplete in the sense that it lacks alignment information. We introduce alignments between syntactic patterns in adjacent sentences as a latent variable. Our model does that based on the IBM model 1 (Brown et al., 1993), where the current sentence is generated by the preceding one, one pattern at a time, with a uniform prior over alignment configurations. The latent alignment variable allows us to model the fact that some patterns are more likely to trigger certain subsequent patterns.

In IBM model 1, a latent alignment function $a : j \mapsto i$ maps patterns in v_1^n (current sentence) to patterns in u_0^m (preceding sentence), where u_0 is a special NULL symbol which models insertion. The score of a document is given by Equation 4.

$$P(D) = \prod_{(u_1^m, v_1^n) \in D} p(v_1 \dots v_n, a_1 \dots a_n | u_0 \dots u_m) \tag{4}$$

Here n is the current sentence and m the preceding sentence. As the alignment is hidden, we marginalise over all possible configurations, which is tractable due to an independence assumption (that items align independently of each other). Equation 5 shows this tractable marginalisation.

$$p(D) = \prod_{(u_1^m, v_1^n) \in D} \prod_{j=1}^{n} \sum_{i=0}^{m} p(v_j | u_i) \tag{5}$$

We resort to Expectation Maximisation (EM) to estimate the parameters in Equation 5 (Brown et al., 1993): due to the convexity of IBM model 1, EM is guaranteed to converge to a global optimum. Moreover, as we observe more data this model converges to better parameters.

A similar solution was proposed in a different context by Soricut and Marcu, (2006) in their work on word co-occurrences.

Table 1: Number of documents and sentences in the training (Gigaword) and test (WMT14) sets.

Corpus	Portion	Documents	Sentences
Gigaword	12/2010	41,564	774,965
WMT14	de-en	164	3,003
WMT14	fr-en	176	3,003
WMT14	ru-en	175	3,003

To avoid assigning 0 probability to documents containing unseen patterns, we modify the training procedure to treat all the singletons as pertaining to an unknown category (UNK), thus reserving probability mass for future unseen items.[3] In addition to this special UNK item, we also include NULL alignments, which together with UNK will smooth the bigram counts.

4 Experiments and Results

4.1 Datasets

To estimate the parameters of the entity-grid and syntax-based models (e.g. distribution over entity role transitions and syntactic patterns), we use the most recent portion of English LDC Gigaword corpus, excluding 2 sections.[4] Table 1 displays information about the size of these datasets.

To test our models on the translation task, we use WMT14 test data as corpus (Bojar et al., 2014), considering submissions from all participating MT systems (including statistical, rule-based, hybrid) in the translation shared task for three language pairs, namely, 13 German-English (de-en) systems, 9 French-English (fr-en) systems, and 13 Russian-English (ru-en) systems.

We assume that the HT (reference) is a coherent text, and that the MT output may or may not be coherent. While the former is a fair assumption, we do acknowledge that many outputs from MT systems may be coherent. However, we are not aware of any datasets with translated data which have been annotated for coherence. This is a challenging task in itself, as judging coherence is a complex and subjective task which requires, at the very least, well trained annotators. Our hypothesis is that a good coherence model should be able to score human translations as having higher coherence than their counterpart machine translations in most cases.

For the shuffling task we also use the MT data, taking the HT documents as the coherent texts and shuffled versions of them to create incoherent ones.

4.2 Metrics

We evaluated the results according to a number of metrics, defined as follows: m is a model, $d \in D$ a document, r the reference or original (non-shuffled) version and $s \in S$

[3] The hypothesis, backed by the Zipf's law, is that unseen items are singletons that we have not yet observed, and that singletons we did observe would remain so if we had observed some more data.

[4] https://catalog.ldc.upenn.edu/LDC2003T05

shuffled or MT output. Then let $\mathrm{win}_m(d_r, d_s)$ return 1 if model m scores reference document d_r higher than a shuffled or MT document d_s, and 0 otherwise. We can define tie and lose analogously. Finally, $\mathrm{first}_m(d_r)$ returns 1 if the reference ranks first, and $\mathrm{solo}_m(d_r)$ returns 1 if the reference occupies a position alone in the ranking. Our various model evaluation methods are defined as follows:

ref$_>$ how often a model ranks reference documents higher than any of their shuffled or MT counterparts: $\frac{1}{|D||S|}\sum_d \sum_s \mathrm{win}_m(d_r, d_s)$

ref$_\geq$ how often a model ranks the reference no worse than any of their shuffled or MT counterparts: $\frac{1}{|D||S|}\sum_d \sum_s \mathrm{win}_m(d_r, d_s) + \mathrm{tie}_m(d_r, d_s)$

ref$_{1^*}$ how often the reference is ranked strictly higher than every other system: $\frac{1}{|D|}\sum_d \mathrm{first}_m(d_r) \times \mathrm{solo}_m(d_r)$

4.3 Results on shuffling task

To test our hypothesis that patterns of syntactic items between adjacent sentences can be better modelled through a latent alignment, we conducted the traditional shuffling experiment with our reference text and a randomly shuffled version of it. The aim was to check whether our formulation for the syntax model, based on IBM model 1, outperforms the original syntax model. Thus we are comparing grammatically correct and coherent sentences instead of MT output.

From our results (Table 3), it is clear that our adaptation (henceforth IBM1) improves over the original syntax model (LN) by a large margin. In fact, in most cases it also outperforms the entity grid. Noteworthy is the fact that the ref$_>$ metric discriminates how often a model ranks the unshuffled documents strictly higher than any other version, not just equal to them, as the ref$_\geq$ does. We experimented at varying depths, displayed as *d* in our results, but display only the best performing ones.

The difference between our experiment and those reported elsewhere (Barzilay and Lapata, 2008; Louis and Nenkova, 2012) is that the experiments elsewhere have been on a specific corpus widely used for coherence prediction, the Earthquakes and Accidents corpus[5]. The scores we report are therefore not as high. By way of comparison, we also include results on the aforementioned corpus (Table 2). Here the ref$_\geq$ metric results for our reimplementation of the syntax model are close those of the original local model with d-sequences (Louis and Nenkova, 2012).

Results for previous grid experiments were obtained using supervised training where the parameters are trained on this same Earthquakes and Accidents corpus, then tested on a heldout section of the same dataset. We adopted a more automated approach, training on more general data. This does, however, affect the results, particularly given the consistent nature of the Earthquakes and Accidents corpus.

4.4 Results on translation task

This evaluation is conducted under the assumption that the reference documents are coherent. An obvious benefit of such a strategy is that we can assess models automatically

[5] http://people.csail.mit.edu/regina/coherence/CLsubmission/

Table 2: Model comparisons for shuffling experiment on Earthquakes and Accidents corpus, ref_{1*} is "accuracy" used in previous work with this corpus.

Earthquakes	ref_{1*}	$\text{ref}_{\geq}$
IBM1-d2	80.88	80.88
IBM1-d3	77.10	77.10
GRID	66.21	66.21
GRAPH	60.53	60.58
LN-d2	57.62	71.73
LN-d3	57.00	67.69

Accidents	ref_{1*}	$\text{ref}_{\geq}$
GRAPH	86.51	86.51
IBM1-d3	72.61	72.61
IBM1-d2	67.32	67.37
GRID	50.25	50.25
LN-d4	46.58	55.89
LN-d2	38.82	57.15

Table 3: Model comparisons for shuffling experiment

fr-en	ref_{1*}	$\text{ref}_{\geq}$
IBM1-d3	82.95	85.23
GRID	75.00	77.84
IBM1-d4	71.59	73.86
GRAPH	50.00	53.98
LN-d3	46.59	59.66
LN-d4	41.48	54.55

de-en	ref_{1*}	$\text{ref}_{\geq}$
GRID	79.27	80.49
IBM1-d3	76.83	76.83
IBM1-d2	71.34	71.34
GRAPH	62.80	65.24
LN-d4	53.66	62.20
LN-d2	47.56	59.15

ru-en	ref_{1*}	$\text{ref}_{\geq}$
IBM1-d3	79.43	80.00
GRID	74.86	76.00
IBM1-d2	74.86	75.43
GRAPH	50.29	54.29
LN-d4	46.29	57.71
LN-d3	45.14	57.14

and objectively without the need for any particular type of annotation (e.g. reference translations). To provide a concise summary of our findings, we aggregate the results for all MT systems in this section.

Table 4 shows the performance of our models according to different evaluation methods (scores are percentages), ranked by the first method.

Table 4: Model comparisons for translation task.

fr-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1*}
IBM1-d4	58.24	58.66	20.45
GRID	55.54	56.68	22.16
IBM1-d3	54.19	54.62	17.61
LN-d4	45.17	55.82	14.77
GRAPH	41.62	45.60	11.93
LN-d3	41.26	59.23	15.34

de-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1*}
GRAPH	67.03	68.62	28.66
IBM1-d2	53.52	53.56	12.20
IBM1-d3	53.05	53.05	17.68
LN-d3	43.67	60.55	8.54
LN-d4	43.34	53.38	10.37
GRID	37.71	37.71	6.10

ru-en	$\text{ref}_{>}$	$\text{ref}_{\geq}$	ref_{1*}
GRAPH	60.84	63.21	20.57
IBM1-d3	58.02	58.02	10.86
IBM1-d2	57.41	57.54	13.14
LN-d3	48.62	63.47	9.14
LN-d4	47.21	58.42	8.57
GRID	31.38	31.38	5.14

Our results show that all the models tested are more limited in their ability to assess coherence in an MT context, as the task is more difficult than that of distinguishing shuffled from original texts. The models can score machine translated texts as well as reference translations, and in some cases, even better than reference translations.

Our extension of the syntax-based model – IBM1 – consistently outperforms LN (Louis Nenkova) according to all metrics. That is because IBM1 learns a distribution over hidden alignments between syntactic items. These alignments can be seen as more plausible explanations for certain syntactic patterns. Moreover, in experiments using

held out data (from WMT13), we noticed that increasing the amount of training data helps IBM1, which is guaranteed to move towards better parameters.

Overall, for a given language pair, we found that the best coherence model was able to score the human translations higher than any particular MT system for more than 58% of the documents. The best score was 67%, which is a good basis to make future improvements on.

Some models are clearly more heavily affected by the use of methods that disregard ties. The LN model typically clusters the reference together with MT systems. The other models, IBM1 and GRAPH, are less affected by differences in evaluation methods. While the figures change across methods, the trend in the ranking of models is maintained.

In general, IBM1 and GRAPH are the strongest in terms of scores, with GRID performing poorly (except for the `fr-en` language pair). Overall GRAPH performs better than GRID, perhaps because it offers a broader view of entity-based coherence, in that it captures links between all entities in all sentences in the text, including links over non-adjacent sentences, and as such is not as dependent on consecutive transitions. If ties are not considered, GRAPH features as the best model for two out of the three language pairs, with IBM1 performing similarly well.

Interestingly, there is a difference between language pairs. It is worth emphasising that among our three language pairs, `fr-en` is arguably the one with the highest MT quality. Low translation quality may have affected the performance of the models differently, as they rely on linguistic information to different extents. GRID, which performed the best for `fr-en`, relies heavily on the correct identification of nouns and their syntactic roles in sentences. Therefore, for the other languages, an excessive number of ungrammatical translations – and unreliable syntactic roles as a consequence – may have affected the model more significantly. Moreover, the `fr-en` language pair is closer than the other two, and therefore more likely to be similar syntactically in the output, which could improve performance of the GRID model. If the MT output remained similar syntactically to the source language, then GRID would not perform as well for other language pairs (it is known that the syntactic assumptions which hold for English do not do so for German). Although this potentially affects GRAPH too, it does not depend on entity transitions but models connections among all sentences in a document. Moreover, a closer inspection of the data showed that the quality of the `fr-en` reference translation was not as good as the `de-en` reference translation. Coupled with better MT output for the `fr-en` language pair, this would make it a harder task for the models to differentiate between HT and MT.

While GRID does well in the shuffling experiment, it does not do so well with the MT output. Clearly, shuffling and reordering is a different task entirely, as illustrated by the difference in the scores between Table 3 and Table 4. By comparison, the ability of GRAPH (as the other entity-based method) to distinguish between HT and MT output is presumably due to it not relying on the transitions between sentences, unlike GRID.

5 Conclusions and Future Work

Work on measuring text coherence has thus far been commonly limited to somewhat artificial scenarios such as sentence shuffling or insertion tasks. These operations natu-

rally tend to break the overall logic of the text. In this paper we have investigated local coherence models for a very different scenario, where texts are automatically translated from a given language by systems of various overall levels of quality. Coherence in this scenario is much more nuanced, as elements of coherence are often present in the translations to some degree, and their absence may be connected to various types of translation errors at different linguistic levels. There are undeniably grammatical issues, but arguably a proportion of these do indirectly affect coherence.

For a given language pair, we found that the best coherence model was able to score the human translations higher than any particular MT system for more than 67.03% of the documents. Our IBM1 model performs strongly, detecting MT output from HT 58.24% of the time, which is a strong result considering that it is based on syntax alone. This model did well in the standard shuffling experiment.

We believe that the source language of the training data is crucially important in this MT domain, as noted by others (Cartoni et al., 2011), as is whether the text is original or translated (Lembersky et al., 2012). We plan to investigate filtering input data and to further expand our coherence models to integrate discourse relations and distributed representations. In addition, by way of a supplementary test to determine that our models are indeed measuring coherence not simply the differences between the MT and HT, we intend to test them on an artificial corpus containing injected coherence errors (Sim Smith et al., 2015).

References

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

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 Ales Tamchyna. (2014). *Findings of the 2014 workshop on statistical machine translation*. In Proceedings of WMT, pages 12-58, Baltimore, Maryland.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. *The mathematics of statistical machine translation: parameter estimation*. Computational Linguistics, 19(2):263-311, June.

Jill Burstein, Joel R. Tetreault, and Slava Andreyev. 2010. *Using entity-based features to model coherence in student essays*. In HLT-NAACL, pages 681-684.

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

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. *How Comparable Are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives*, Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 78–86, Portland, Oregon, USA, Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011. *Extending the Entity Grid with Entity-Specific Features*. Proceedings of ACL, pages 125-129, Portland, Oregon, USA, Association for Computational Linguistics.

Zhengxian Gong and Min Zhang and Guodong Zhou. 2015. *Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion*. Proceedings of the Second Workshop on Discourse in Machine Translation, pages 52-58, September, Lisbon, Portugal, Association for Computational Linguistics.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. *Centering: A framework for modeling the local coherence of discourse*. Computational Linguistics, 21:203-225.

Camille Guinaudeau and Michael Strube. 2013. *Graph-based local coherence modeling*. In Proceedings of ACL, pages 93103.

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

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*. Prentice Hall, 2 edition.

Mirella Lapata. 2005. *Automatic evaluation of text coherence: models and representations*. In Proceedings of IJCAI, pages 1085-1090.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. *Language models for machine translation: Original vs. translated texts*. Comput. Linguist., 38(4):799-825, December.

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 (EMNLP), pages 2039-2048. Association for Computational Linguistics.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. *A hierarchical neural autoencoder for paragraphs and documents*. In Proceedings of ACL, pages 1106-1115, Beijing, China, July. Association for Computational Linguistics.

Liu Shujie Yang Muyun Li Mu Zhou Ming Lin, Rui and Sheng Li. 2015. *Hierarchical recurrent neural network for document modeling*. In Proceedings of EMNLP, page 899-907, Lisbon, Portugal, September. Association for Computational Linguistics.

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

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

Max M. Louwerse and Arthur C. Graesser, 2005. *Coherence in Discourse*, pages 216-218. Encyclopedia of linguistics.

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, pages 52-58, Lisbon, Portugal, September. Association for Computational Linguistics.

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

Radu Soricut and Daniel Marcu. 2006. *Discourse generation using utility-trained coherence models*. In Proceedings of the COLING/ACL, pages 803-810, Sydney, Australia.

Jorg 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.

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.

Received May 2, 2016 , accepted May 10. 2016