Modeling Lexical Cohesion for Document-Level Machine Translation

Deyi Xiong[1,2], Guosheng Ben[3], Min Zhang[1,2*], Yajuan Lü[3] and Qun Liu[3,4]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China 215006
{*dyxiong, minzhang*}*@suda.edu.cn*
[2]Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore138632
[3]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, China
{*benguosheng, lvyajuan, liuqun*}*@ict.ac.cn*
[4]Centre for Next Generation Localisation, School of Computing, Dublin City University, Ireland

Abstract

Lexical cohesion arises from a chain of lexical items that establish links between sentences in a text. In this paper we propose three different models to capture lexical cohesion for document-level machine translation: (a) a direct reward model where translation hypotheses are rewarded whenever lexical cohesion devices occur in them, (b) a conditional probability model where the appropriateness of using lexical cohesion devices is measured, and (c) a mutual information trigger model where a lexical cohesion relation is considered as a trigger pair and the strength of the association between the trigger and the triggered item is estimated by mutual information. We integrate the three models into hierarchical phrase-based machine translation and evaluate their effectiveness on the NIST Chinese-English translation tasks with large-scale training data. Experiment results show that all three models can achieve substantial improvements over the baseline and that the *mutual information trigger model* performs better than the others.

1 Introduction

Most statistical machine translation (SMT) systems translate a text in a sentence-by-sentence fashion. The major drawback of this kind of sentence-based document translation is the neglect of inter-sentence links and dependencies. From a linguistic perspective, **cohesion** is a well-known means to establish such inter-sentential links within a text. Widdowson [1979] defines cohesion as "the overt structural link between sentences as formal items". In other words, cohesion refers to various manifest linguistic links (e.g., references, word repetitions) between sentences within a text that hold the text together.

Appropriately establishing such cohesion links makes machine-generated texts cohesive. Unfortunately, Wong and Kit [2012] find that SMT systems tend to use less cohesion links than human translators. This is due to the fact that there is no independent cohesion model in SMT systems to capture these inter-sentence cohesion links.

*Corresponding author

In this paper, we study cohesion in the context of SMT and try to incorporate it into document-level translation. Halliday and Hasan [1976] identify 5 categories of cohesion devices that create cohesion in texts: reference, substitution, ellipsis, conjunction and lexical cohesion. The former 4 devices can be roughly grouped into grammatical cohesion in contrast to lexical cohesion. Here we focus on **lexical cohesion** that connects sentences in a text not through grammatical devices, but rather through lexical choices (i.e., lexical cohesion devices). We propose three different models to capture lexical cohesion for document-level SMT. In particular,

- *Direct Reward Model:* We introduce a direct reward model to encourage translation hypotheses where lexical cohesion devices are used. For instance, a word occurs in the best translation hypothesis of a sentence. If the same word reiterates or its synonym words appear in translation hypotheses of succeeding sentences, such hypotheses will be rewarded by the model.
- *Conditional Probability Model:* The direct reward model is apt to use lexical cohesion devices frequently. However, overuse of the same lexical cohesion devices may jeopardize the readability of a text [Wong and Kit, 2012]. Therefore we want to use lexical cohesion devices appropriately rather than frequently. We measure the appropriateness by calculating the likelihood that a lexical cohesion item is used again given its presence in preceding sentences in a text.
- *Mutual Information Trigger Model:* We extend the conditional probability model further to consider the occurrence of a lexical cohesion item and its reoccurrence in succeeding sentences as a trigger pair. We treat the first occurrence of a cohesion item as a trigger (presupposing) and its reoccurrence (e.g., the same word repeated or the synonym/near-synonym of the word) as the triggered item (presupposed). Then we build a mutual information trigger model to measure the dependencies between trigger pairs.

We integrate the three models into a hierarchical phrase-based SMT system. Experiment results show that they are all able to achieve substantial improvements over the baseline. The mutual information trigger model outperforms the baseline by up to 0.92 BLEU [Papineni *et al.*, 2002] points and also performs better than the other two models.

We begin in Section 2 with a brief overview of related work. Section 3 presents the definition and detection of lexical cohesion devices. Section 4 elaborates the proposed three models and Section 5 introduces how we integrate the three models into SMT. We conduct experiments in Section 6 to validate the effectiveness of the proposed models and the impact of lexical cohesion devices on translation quality. Finally, Section 7 presents conclusions and directions for future research.

2 Related Work

The exploration of cohesion in SMT is very limited. Most previous studies on document-level machine translation take different perspectives rather than cohesion. Lexical cohesion is used to facilitate machine translation evaluation at document level and integrated into a semantic language model, both of which are significantly different from our models. This section will briefly introduce these approaches and other work that is partly related to our cohesion models.

In document-level machine translation, Tiedemann [2010] integrates cache-based language and translation models that are built from recently translated sentences into SMT. Gong et al. [2011] further extend this cache-based approach by introducing two additional caches: a static cache that stores phrases extracted from documents in training data which are similar to the document in question and a topic cache with target language topic words. Xiao et al. [2011] try to solve the translation consistency issue in document-level translation by introducing a hard constraint where ambiguous source words are required to be consistently translated into the most frequent translation options. Ture et al. [2012] soften this consistency constraint by integrating three counting features into decoder. Although these studies partially explore word/phrase repetitions via cache or consistency constraint, lexical cohesion is not the focus, even not mentioned at all in these studies.

Wong and Kit [2012] incorporate various lexical cohesion features into automatic metrics to evaluate machine translation quality at the document level. The key difference between their work and ours is that we integrate lexical cohesion devices into decoder to improve translation quality rather than improve machine translation evaluation metrics.

Hardmeier et al. [2012] introduce a document-wide phrase-based decoder and integrate a semantic language model into the decoder. The semantic language model explores n-grams that cross sentence boundaries. Although they argue that the semantic language model is able to capture lexical cohesion, experiment results show a very small gain in BLEU score achieved by their language model.

Our conditional probability model is partially inspired by Church's study on adaptation [Church, 2000]. He models word repetition based on the probability that a word will occur in the "test" portion of a text given its presence in the "history" part of the text. Our third cohesion model is related to the mutual information trigger language model proposed by Xiong et al. [2011]. The key difference is that they capture intra-sentence trigger pairs while we explore inter-sentence triggers. Moreover we define our triggers on lexical cohesion devices instead of any words.

3 Lexical Cohesion Devices

Lexical cohesion arises from a chain of lexical items that establish links across sentences within a text. We call these lexical items **lexical cohesion devices**. In this study we focus on three classes[1] of lexical cohesion devices: *reiteration*, *synonym/near-synonym*, and *super-subordinate*. Such devices are extracted from content words (i.e., words left after stop words being filtered out). This section presents the definition and detection of the three types of lexical cohesion devices.

Reiteration: Reiteration refers to the repetition of the same words in a document. For example, in the following two sentences extracted from a document, the word *investigation* in the first sentence repeats itself in the second sentence.

> 1. During two days of *investigation*, they have inquired about more than 50 people.
> 2. Right now, the scope of *investigation* is being narrowed down.

According to Church [2000], supposing that the first occurrence of a word in a text has a probability p, the joint probability of two instances of the same word in the text is closer to $\frac{p}{2}$ rather than p^2. This indicates that reiteration is very common in texts.

Synonym/near-synonym: We use WordNet [Fellbaum, 1998] to define synonyms and near-synonyms. WordNet is a lexical database that clusters English nouns, verbs, adjectives and adverbs into sets of semantic groups that are called *synsets*. Let $synset(w)$ be a function that defines synonyms grouped in the same synset as word w in WordNet. We can use the function to find all synonyms and near-synonyms for word w. Let us denote the set of synonyms in $synset(w)$ as syn_0. Near-synonym set syn_1 is obtained as the union of all synsets that are defined by the function $synset(w)$ where $w \in syn_0$. It can be formulated as follows.

$$syn_1 = \bigcup_{w \in syn_0} synset(w) \tag{1}$$

Similarly syn_2 and syn_3 can be defined recursively as follows.

$$syn_2 = \bigcup_{w \in syn_1} synset(w) \tag{2}$$

$$syn_3 = \bigcup_{w \in syn_2} synset(w) \tag{3}$$

Obviously, the near-synonym sets syn_m become larger and larger as m increases. They have the following relationship

$$syn_0 \subseteq syn_1 \subseteq syn_2 \subseteq \cdots \subseteq \mathcal{V}$$

where $\mathcal{V}$ denotes the vocabulary. Although we can define any syn_m, we only keep near-synonym sets syn_m where $m \leq 3$

[1] Other lexical cohesion devices such as antonyms do not occur frequently. Therefore we do not involve them in our study.

for our experiments. The reason is twofold: 1) a larger near-synonym set results in a higher computational cost in finding near-synonym words; 2) larger near-synonym sets contain noisy words that are actually quite distant in WordNet.

Super-subordinate: Superordinate and subordinate are formed by words with an is-a semantic relationship, such as *apple* and *fruit* (hypernym), *furniture* and *cupboard* (hyponym), and so on. As the super-subordinate relation is also encoded in WordNet, we still use WordNet to detect hypernyms and hyponyms. Let $hypset(w)$ be a function that defines both hypernyms and hyponyms in WordNet for word w. Because the hypernym/hyponym relation is transitive, we can define hyp_0 - hyp_3 in the way that we formulate syn_m.

$$hyp_0 = hypset(w) \tag{4}$$

$$hyp_1 = \bigcup_{w \in hyp_0} hypset(w) \tag{5}$$

$$hyp_2 = \bigcup_{w \in hyp_1} hypset(w) \tag{6}$$

$$hyp_3 = \bigcup_{w \in hyp_2} hypset(w) \tag{7}$$

For notational convenience, we use rep, syn and hyp to represent the reiteration, synonym/near-synonym and super-subordinate lexical cohesion device respectively hereafter.

So how do the three lexical cohesion devices distribute in real-world texts? Table 1 presents the percentages of the three cohesion devices in our training data (see Section 6.1). We calculate the percentages according to the following formula.

$$\text{Percentage} = \frac{d}{c} \tag{8}$$

where d is the number of words that are used as a lexical cohesion device, c is the total number of content words in our corpus. If a word is a rep/syn/hyp of another word in a document, it is counted as a cohesion device as long as the two words are in the same document.

The reiteration cohesion device is the device that occurs most frequently, contributing nearly a third of all content words (30.85%). The percentage of the synonym/near-synonym device is varying from 15.84% to 18.01% while the super-subordinate device from 14.37% to 19.00% when we increase m from 0 to 3. Such a distribution of lexical cohesion devices in texts is similar to the finding in [Wong and Kit, 2012]. We empirically reconfirm that lexical cohesion devices are frequently used in real-world texts. Therefore modeling lexical cohesion and incorporating it into SMT would benefit document-level machine translation.

4 Models

This section elaborates the three models that we propose to capture lexical cohesion for document-level machine translation. They are the (A) direct reward model, (B) conditional probability model and (C) mutual information trigger model respectively.

Table 1: Distributions of lexical cohesion devices in the training data. The percentage of "not lexical cohesion device" depends on how we define syn_m/hyp_m.

Word Type		**Percentage (%)**
rep		30.85
syn	syn_0	15.84
	syn_1	17.10
	syn_2	17.58
	syn_3	18.01
hyp	hyp_0	14.37
	hyp_1	16.49
	hyp_2	18.04
	hyp_3	19.00
not lexical cohesion device		32.14 - 38.94
all content words		100

4.1 Model A: Direct Rewards

The most straightforward way that incorporates lexical cohesion into SMT is to directly reward a hypothesis whenever a lexical cohesion device occurs in the hypothesis. In order to reward translation hypotheses containing lexical cohesion devices, we maintain one counter for each class of lexical cohesion device rep, syn and hyp.

- rep: We accumulate the counter whenever a content word in the current hypothesis has already occurred in recently translated sentences.[2]
- syn: If a word of the current hypothesis is in the synonym/near-synonym set (i.e., $syn_0 - syn_3$) of any words in previously translated sentences, the counter is accumulated.
- hyp: If a word in the current hypothesis has a super-subordinate relationship with any words in recently translated sentences, the counter is accumulated.

The three counters are integrated into the log-linear model of SMT as three different features. Their weights are tuned via minimum error rate training (MERT) [Och, 2003]. Through the three counting features, the direct reward model can enable decoder to favor translations that establish lexical cohesion links with recently translated sentences.

4.2 Model B: Conditional Probabilities

The direct reward model is able to capture lexical cohesion links across sentences of a text. However, it tends to use lexical cohesion devices frequently, which may cause the overuse of some devices such as word repetitions. What we really concern is whether cohesion links between sentences are correctly established. Therefore we want to model how appropriately rather than frequently lexical cohesion devices are used.

Before we introduce the model that estimates the appropriateness of a lexical cohesion device, we formally define a lexical cohesion relation as follows.

$$x \mathcal{R} y \tag{9}$$

[2]Any sentences that have already been translated in the same document are considered as "recently translated sentences".

where x and y are two ordered lexical items in a text. x is located in a sentence that precedes the sentence where y occurs. $\mathcal{R}$ denotes that x and y have the relationship $\mathcal{R} \in \{rep, syn, hyp\}$ defined in the last section. Inspired by Church's study on repetition [Church, 2000], we measure the appropriateness of a lexical cohesion device by calculating the conditional probability of item y given x and their relationship $\mathcal{R}$. This probability estimates how possible that y (e.g., *fruit*) will be mentioned if x (e.g., *apple*) has already been mentioned in a text.

In particular, the conditional probability $p(y|x, \mathcal{R})$ for the $\mathcal{R}$ lexical cohesion device (i.e., $\{rep, syn, hyp\}$ device) can be calculated as follows.

$$p(y|x, \mathcal{R}) = \frac{b}{a} \tag{10}$$

where a denotes the number of documents with lexical item x and b denotes the number of documents with x and the corresponding item y that has the relationship $\mathcal{R}$ with x and occurs in a sentence after the sentence where x is present.

The three conditional probabilities $p(y|x, rep)$, $p(y|x, syn)$ and $p(y|x, hyp)$ for the reiteration, synonym/near-synonym and super-subordinate cohesion device are calculated as above.

Based on these probabilities, the conditional probability model for the three classes of lexical cohesion devices can be defined. Given a sentence y_1^m, the conditional probability model for the $\mathcal{R}$ lexical cohesion device is formulated as follows.

$$P_{\mathcal{R}}(y_1^m) = \prod_{y_i} p(y_i|\cdot, \mathcal{R}) \tag{11}$$

where y_i are content words in the sentence y_1^m. We may find multiple words x_1^q from recently translated sentences that have the $\mathcal{R}$ relationship with word y_i. The probability $p(y_i|\cdot, \mathcal{R})$ can be defined as the geometric mean[3] of all probabilities $p(y_i|x_j, \mathcal{R}),\ \ x_j \in x_1^q$.

$$p(y_i|\cdot, \mathcal{R}) = \sqrt[q]{\prod_{j=1}^{q} p(y_i|x_j, \mathcal{R})} \tag{12}$$

The conditional probability model for the reiteration device $P_{rep}(y_1^m)$, the synonym/near-synonym device $P_{syn}(y_1^m)$ and the super-subordinate device $P_{hyp}(y_1^m)$ can be defined as above. They are integrated as three features into the log-linear model of SMT and calculated separately.

4.3 Model C: Mutual Information Triggers

We further extend the conditional probability model to a trigger model by positing a lexical cohesion relation $x\mathcal{R}y$ ($\mathcal{R} \in \{rep, syn, hyp\}$) as a trigger pair: x being the trigger and y the triggered item. The possibility that y will occur given x is mentioned is equal to the chance that x triggers y. Therefore we use the pointwise mutual information (PMI) [Church and Hanks, 1990] between the trigger x and the triggered word y to measure the appropriateness of the lexical cohesion device y given the occurrence of x.

The PMI for the relation $x\mathcal{R}y$ is calculated as follows.

$$\text{PMI}(x\mathcal{R}y) = \log(\frac{p(x, y, \mathcal{R})}{p(x, \mathcal{R})p(y, \mathcal{R})}) \tag{13}$$

The joint probability $p(x, y, \mathcal{R})$ is defined as

$$p(x, y, \mathcal{R}) = \frac{C(x, y, \mathcal{R})}{\sum_{x,y} C(x, y, \mathcal{R})} \tag{14}$$

where $C(x, y, \mathcal{R})$ is the number of documents where both x and y occur with the relationship $\mathcal{R}$ in different sentences. Clearly, $\sum_{x,y} C(x, y, \mathcal{R}) = C(\mathcal{R})$ is the number of documents where the lexical cohesion relation $\mathcal{R} \in \{rep, syn, hyp\}$ occurs. The marginal probabilities of $(x, \mathcal{R})$ and $(y, \mathcal{R})$ can be calculated as follows.

$$p(x, \mathcal{R}) = \sum_{y} p(x, y, \mathcal{R}) \tag{15}$$

$$p(y, \mathcal{R}) = \sum_{x} p(x, y, \mathcal{R}) \tag{16}$$

The mutual information trigger model for the $\mathcal{R}$ lexical cohesion device on a given sentence y_1^m is defined as follows.

$$\text{MI}_{\mathcal{R}}(y_1^m) = \prod_{y_i} exp(\text{PMI}(\cdot\mathcal{R}y_i)) \tag{17}$$

where y_i are content words in the sentence y_1^m and $\text{PMI}(\cdot\mathcal{R}y_i)$ is defined as the maximum PMI value among all trigger words x_1^q from recently translated sentences that have the $\mathcal{R}$ relationship with word y_i

$$\text{PMI}(\cdot\mathcal{R}y_i) = max_{1 \le j \le q} \text{PMI}(x_j\mathcal{R}y_i) \tag{18}$$

Three models $\text{MI}_{rep}(y_1^m)$, $\text{MI}_{syn}(y_1^m)$ and $\text{MI}_{hyp}(y_1^m)$ for the reiteration device, the synonym/near-synonym device and the super-subordinate device can be formulated as above. They are trained separately and integrated into SMT as three different features.

5 Decoding

In this section, we discuss how the three models are integrated into SMT. We translate a document still in a sentence-by-sentence fashion. However, we maintain a vector for SMT decoder to capture lexical cohesion devices in documents. The vector is used to store target language content words from recently translated sentences.

When we translate a new sentence in a document, for each generated target language content word w, we search the vector to find all words that have the relationship of $\{rep, syn, hyp\}$ with content word w. We put these words in three sets $\mathcal{S}_{rep}$, $\mathcal{S}_{syn}$ and $\mathcal{S}_{hyp}$ respectively.

For the direct reward model, if $\mathcal{S}_{rep}$/$\mathcal{S}_{syn}$/$\mathcal{S}_{hyp}$ is not null, the corresponding counter $\{rep, syn, hyp\}$ will be accumulated. For the conditional probability model and the mutual

[3] We can also define $p(y_i|\cdot, \mathcal{R})$ as the maximum probability:

$$p(y_i|\cdot, \mathcal{R}) = max_{1 \le j \le q} p(y_i|x_j, \mathcal{R})$$

However our preliminary experiments show that the geometric mean is better than the maximum probability. Therefore we use the geometric mean to calculate $p(y_i|\cdot, \mathcal{R})$.

information trigger model, their model scores will be updated according to the equation (11) and (17) respectively.

Once a sentence is completely translated, target language content words in the best translation of the sentence go into the vector. When all sentences in a document are completely translated, the vector is cleared to store content words for the next document.

6 Experiment

We conducted a series of experiments to validate the effectiveness of the proposed three lexical cohesion models using a hierarchical phrase-based SMT [Chiang, 2007] system on large-scale training data. Additionally, we also want to find answers for the following questions through experiments.

1. What impact do these lexical cohesion devices have on translation quality in terms of BLEU?
2. Which model is the best model to incorporate lexical cohesion devices into SMT?

6.1 Setup

The bilingual training data are from LDC[4], which contains 3.8M sentence pairs with 96.9M Chinese words and 109.5M English words. We used a 4-gram language model trained on the Xinhua portion of the English Gigaword corpus (306 million words) via the SRILM toolkit [Stolcke, 2002] with Kneser-Ney smoothing.

In order to build the conditional probability model and the mutual information trigger model, we collected data with document boundaries explicitly provided. The corpora are selected from our bilingual training data and the whole Hong Kong parallel text corpus[5]. In total, the selected corpora contain 103,236 documents and 2.80M sentences. Averagely, each document consists of 28.4 sentences.

We used the NIST MT05 as the MERT [Och, 2003] tuning set, the NIST MT06 as the development test set and the MT08 as the final test set. The numbers of documents/sentences in the NIST MT05, MT06 and MT08 are 100/1082, 79/1664 and 109/1357 respectively. They contain 10.8, 21.1, and 12.4 sentences per document respectively.

We used the case-insensitive BLEU-4 as our evaluation metric. In order to alleviate the impact of the instability of MERT, we ran it three times for all our experiments and presented the average BLEU scores on the three runs following the suggestion by Clark et al. [2011].

6.2 Effect of the Direct Reward Model

Our first group of experiments were carried out to investigate the effectiveness of the direct reward model. We report the results in Table 2. First of all, we integrated only a single lexical cohesion device into decoder at a time. In other words, the rep, syn and hyp lexical cohesion devices are explored one by one.

[4]The corpora include LDC2002E18, LDC2003E07, LDC2003E14, LDC2004E12, LDC2004T07, LDC2004T08 (Only Hong Kong News), LDC2005T06 and LDC2005T10.

[5]They are LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hong Kong Hansards/Laws/News).

System		MT06	MT08	Avg
Base		30.43	23.32	26.88
rep		31.06	23.76	27.41
syn	syn_0	30.60	23.42	27.01
	syn_1	30.74	23.32	27.03
	syn_2	**30.88**	23.54	27.21
	syn_3	30.78	23.47	27.13
hyp	hyp_0	30.66	23.43	27.05
	hyp_1	30.92	23.75	27.34
	hyp_2	**30.95**	23.66	27.31
	hyp_3	30.68	23.68	27.18
$rep + syn_2 + hyp_2$		**31.29**	**23.91**	**27.60**

Table 2: BLEU-4 scores of the *direct reward model* with various lexical cohesion devices on the development test set MT06 and the final test set MT08.

As we define the synonym/near-synonym and super-subordinate cohesion devices at different levels from syn_0/hyp_0 to syn_3/hyp_3 (see Section 3), we ran experiments to find the level at which our model has the best performance using the NIST MT06 as the development test set. From Table 2, we observe that the direct reward model obtains steady improvements over the baseline for both the synonym/near-synonym and super-subordinate device when m is increased from 0 to 2. However, when we set m to 3, the performance drops for both devices. The highest BLEU scores 30.88 and 30.95 are obtained at the level 2 (i.e., syn_2 and hyp_2. See their definitions in the equation (2) and (6).) on the development test set MT06 for the two lexical cohesion devices. Hence we use syn_2 for the synonym/near-synonym cohesion device and hyp_2 for the super-subordinate device in all experiments thereafter.

Integrating a single lexical cohesion device into decoder, we gain 0.53, 0.33 and 0.43 BLEU points for rep, syn and hyp respectively over the baseline. The integration of the reiteration cohesion device achieves the largest improvement, which is consistent with the fact that this device is the most frequently used device in real-world texts according to Table 1.

We also want to investigate whether we can achieve further improvements if we integrate the three devices into decoder simultaneously with the direct reward model. Experiment results (displayed in the last row of Table 2) show that we do achieve further improvements. The final gain over the baseline is on average 0.72 BLEU points.

Overall, the substantial improvements over the baseline obtained by the direct reward model indicates that this model is a simple and yet effective model for incorporating lexical cohesion devices into SMT.

6.3 Effect of the Conditional Probability Model

We conducted the second group of experiments to study 1) whether the conditional probability model is able to improve the performance in terms of BLEU and 2) whether it can outperform the direct reward model. Results are shown in Table 3, from which we observe the following phenomena that are similar to what we have found in the direct reward model.

System	MT06	MT08	Avg
Base	30.43	23.32	26.88
rep	31.01	23.56	27.29
syn_2	31.00	23.59	27.30
hyp_2	30.59	23.35	26.97
$rep + syn_2 + hyp_2$	**31.21**	**24.06**	**27.64**

Table 3: BLEU-4 scores of the *conditional probability model* with various lexical cohesion devices on the development test set MT06 and the final test set MT08.

System	MT06	MT08	Avg
Base	30.43	23.32	26.88
rep	31.22	23.81	27.52
syn_2	31.00	23.63	27.32
hyp_2	30.93	23.58	27.26
$rep + syn_2 + hyp_2$	**31.35**	**24.11**	**27.73**

Table 4: BLEU-4 scores of the *mutual information trigger model* with various lexical cohesion devices on the development test set MT06 and the final test set MT08.

- With a single lexical cohesion device rep, syn or hyp, the conditional probability model also outperforms the baseline by up to 0.58 BLEU points on the MT06.
- The simultaneous incorporation of the three lexical cohesion devices into SMT can achieve a further improvement over the integration of a single device. We obtain an average improvement of 0.76 BLEU points over the baseline by integrating the combination of three lexical cohesion devices.

Comparing the improvement obtained by the combination $rep + syn_2 + hyp_2$ in the conditional probability model against the gain of the direct reward model, we can see that the conditional probability model is marginally better than the direct reward model.

6.4 Effect of the Mutual Information Trigger Model

The last group of experiments were conducted on the mutual information trigger model. Results are presented in Table 4. From the table, we find that the mutual information trigger model is also able to improve the performance over the baseline. Integrating a single lexical cohesion device into SMT, the model gains an improvement of up to 0.64 BLEU points on the MT06. Combining all three devices together, the model outperforms the baseline by an average improvement of 0.85 BLEU points.

The stable improvements obtained by the direct reward model, the conditional probability model and the mutual information trigger model strongly suggest that lexical cohesion devices are very useful for SMT and that the incorporation of them into document-level translation is indeed able to improve translation quality.

We also find that the mutual information trigger model is better than the other two models in terms of the performance achieved by the combination of three cohesion devices $rep + syn_2 + hyp_2$. The mutual information trigger model outperforms the direct reward model by up to 0.2 BLEU points on the MT08 test set. This suggests that we should integrate lexical cohesion devices into SMT appropriately rather than frequently.

7 Conclusions

In this paper we have presented three different models to incorporate three classes of lexical cohesion devices, namely the reiteration, synonym/near-synonym and super-subordinate device, into SMT. The three models are the first attempt, to our knowledge, to successfully integrate lexical cohesion into document-level machine translation and achieve substantial improvements over the baseline.

The direct reward model maintains a counter per device and accumulate the counters whenever lexical cohesion devices are used in translation hypotheses. The conditional probability model calculates the probability that a lexical cohesion device will be used again given its presence in recently translated sentences in a text. The mutual information trigger model treats the first instance of a lexical cohesion device and its re-occurrence in a document as the trigger and the triggered item, whose association strength is then measured by the pointwise mutual information.

We have integrated these three models into a hierarchical phrase-based SMT system[6] and conducted a series of experiments to verify their effectiveness. Results have shown that

- All three models are able to substantially improve translation quality in terms of BLEU over the baseline.
- The simultaneous incorporation of all three devices can provide further improvements.
- The mutual information trigger model outperforms the other two models.

As our experiment results suggest that we should use lexical cohesion devices appropriately rather than frequently, an important future direction is to use more features to measure the appropriateness of the occurrence of a lexical cohesion device in a sentence. Informative features such as global document context, lexical cohesion devices in source language documents can be incorporated as evidences that a lexical cohesion device will be used again given its presence in recently translated sentences.

Cohesion and *coherence* have often been studied together in discourse analysis. Cohesion is related to the surface structure link while coherence is concerned with the underlying connectedness in a text [Vasconcellos, 1989]. Compared with cohesion, coherence is not easy to be detected. In spite of this, it has been successfully explored and proven useful in document summarization [Barzilay and Lee, 2004; Barzilay and Lapata, 2008]. In the future, we plan to extend our models to capture both cohesion and coherence for document-level machine translation. This study could also uncover interesting connections between cohesion and coherence in bitexts.

[6]Our models are not limited to hierarchical phrase-based SMT. They can be easily applied to other SMT formalisms, such as phrase-based and syntax-based SMT.

Acknowledgements

The second, forth and fifth authors were supported by High-Technology R&D Program (863) Project No 2011AA01A207. The fifth author was also partially supported by Science Foundation Ireland (Grant No.07/CE/I1142) as part of the CNGL at Dublin City University. We would like to thank three anonymous reviewers for their insightful comments.

References

[Barzilay and Lapata, 2008] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

[Barzilay and Lee, 2004] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston, Massachusetts, USA, May 2 - May 7 2004.

[Chiang, 2007] David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

[Church and Hanks, 1990] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[Church, 2000] Kenneth W. Church. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 180–186, Stroudsburg, PA, USA, 2000.

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

[Fellbaum, 1998] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Gong *et al.*, 2011] Zhengxian Gong, Min Zhang, and Guodong Zhou. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, Scotland, UK., July 2011.

[Halliday and Hasan, 1976] M.A.K Halliday and Ruqayia Hasan. *Cohesion in English*. London: Longman, 1976.

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

[Och, 2003] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July 2003.

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

[Stolcke, 2002] Andreas Stolcke. Srilm–an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA, September 2002.

[Tiedemann, 2010] Jörg Tiedemann. 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, July 2010.

[Ture *et al.*, 2012] Ferhan Ture, Douglas W. Oard, and Philip Resnik. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426, Montréal, Canada, June 2012.

[Vasconcellos, 1989] Muriel Vasconcellos. Cohesion and coherence in the presentation of machine translation products. In James E.Alatis, editor, *Geogetown University Round Table on Languages and Linguistics 1989*, pages 89–105, Washington, D.C., 1989. Georgetown University Press.

[Widdowson, 1979] H.G. Widdowson. *Explorations in Applied Linguistics*. London and Edinburgh: Oxford University Press, 1979.

[Wong and Kit, 2012] Billy T. M. Wong and Chunyu Kit. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea, July 2012.

[Xiao *et al.*, 2011] Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. Document-level consistency verification in machine translation. In *Proceedings of the 2011 MT summit XIII*, pages 131–138, Xiamen, China, September 2011.

[Xiong *et al.*, 2011] Deyi Xiong, Min Zhang, and Haizhou Li. Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1288–1297, Portland, Oregon, USA, June 2011.