Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7

Hiroshi Echizen-ya
Hokkai-Gakuen University
S 26-Jo, W 11-Chome, Chuo-ku,
Sapporo, 064-0926 Japan
echi@eli.hokkai-s-u.ac.jp

Terumasa Ehara
Yamanashi Eiwa College
888 Yokone-machi, Kofu,
Yamanashi, 400-8555 Japan

Sayori Shimohata
Oki Electric Industry Co., Ltd.
1-16-8 Chuou Warabi-shi,
Saitama 335-8510, Japan

Atsushi Fujii
University of Tsukuba
1-2 Kasuga, Tsukuba,
Ibaraki, 305-8550, Japan

Masao Utiyama
National Institute of Information
and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto 619-0289, Japan

Mikio Yamamoto
University of Tsukuba
1-1-1, Tennodai, Tsukuba,
Ibaraki, 305-8573, Japan

Takehito Utsuro
University of Tsukuba
1-1-1, Tennodai, Tsukuba,
Ibaraki, 305-8573, Japan

Noriko Kando
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan

Abstract

Herein, we describe meta-evaluation based on various automatic evaluation methods for machine translation using patent translation data in NTCIR-7. We investigated the correlation between results obtained using automatic evaluation methods and human judges by particularly addressing sentence-level evaluation because the improvement of sentence-level evaluation is important to realize high-quality automatic evaluation that is equivalent to that of human judges. Through this meta-evaluation, we confirmed that some automatic evaluation methods can yield high correlation in sentence-level adequacy (about 0.6). However, the correlation of sentence-level fluency in the rule-based machine translation systems was insufficient (less than 0.5). These results indicate that automatic evaluation methods using grammatical knowledge must be developed to improve sentence-level correlation.

1 Introduction

In the field of machine translation, various tasks necessitate high-quality automatic evaluation systems that can evaluate translation quality quantitatively. A patent translation also requires such systems. One automatic evaluation method, BLEU (Papineni et al., 2002), is popular among various automatic evaluation methods because BLEU can yield high correlation with human judgments for document-level evaluation (Coughlin, 2007). Nevertheless, BLEU is inadequate for sentence-level evaluation. Therefore, a great difference exists between the language process used by BLEU and the language

process of humans. It is important to realize high-quality automatic evaluation that is equivalent to that of human judges by particularly addressing sentence-level evaluation. For sentence-level evaluation, several approaches have been proposed (Kluesza and Shieber, 2004; Gamon et al., 2005; Mutton et al., 2007) in recent years, but meta-evaluation based on various automatic evaluation methods has not been performed sufficiently. Therefore, we performed a meta-evaluation using various automatic evaluation methods with patent translation data in NTCIR-7 (Fujii et al., 2008). Experimental results indicate that some automatic evaluation methods can achieve high correlation of sentence-level adequacy (about 0.6). However, the correlation of sentence-level fluency in rule-based machine translation systems remains insufficient (under 0.5). Rule-based machine translation often generates correct or erroneous translation sentences that differ more from the reference in the phrase sequence than sentences obtained using statistical machine translation. Therefore, realizing automatic evaluation methods that can use grammatical knowledge effectively is important.

2 Automatic Evaluation Methods

For this meta-evaluation, we use the following eight automatic evaluation methods:

1) Recursive acquisition of Intuitive comMon PArts ConTinuum (IMPACT) (Echizen-ya and Araki, 2007)

2) Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) (Lin and Och, 2004)

3) Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002)

4) National Institute of Standards and Technology (NIST) (NIST, 2002)

5) Word Number (WN)

6) Normalized Mean Grams, Word Number (NMG_WN) (Ehara, 2007)

7) Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee and Lavie., 2005)

8) Word Error Rate (WER) (Leusch et al., 2003)

In method 3, BLEU used in this meta-evaluation is improved to perform sentence-level evaluation: the maximum N value, which exists between the translation sentence and the references, is used, not a fixed N value (*e.g.*, N=4). In method 5, WN is an automatic evaluation method using the word number of the translation sentence as the score. It is based on the presupposition that the quality of the translation is insufficient when the word number of the translation sentence is large. The evaluation of the translation sentence is high when the score is low in WN: the coefficient between the evaluation of this method and that of human judges is negative. In method 6, although NMG (Ehara, 2007) is based on n-gram, it counts words in the longest word sequence matches between the translation sentence and the references to evaluate fluency correctly. The following shows the NMG_WN definition.

$$\text{NMG\_WN} = \text{NMG\_REF} + \text{NMG\_COR} - 0.05 \times \text{WN} \quad (1)$$

In NMG_REF, the target language sentences corresponding to the source language sentences are referred. In NMG_COR, the large corpus in the target language sentences is referred. In method 7, the matching modules of METEOR used in this meta-evaluation are the exact matching module, stemmed matching module, and a WordNet based synonym-matching module.

3 Experiments

3.1 Experimental Data

For this meta-evaluation, 14 machine translation systems produced by 14 groups attending the NTCIR-7 workshop were used. Each machine translation system translated 100 Japanese sentences into 100 English sentences. Therefore,

Table 1: Groups and types of machine translation systems.

Group	tori	FDU -MCandWI	HIT2	JAPIO	KLE	MIT	NAIST-NTT
Type	SMT	SMT	SMT	RBMT	SMT	SMT	SMT
Group	NiCT-ATR	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt
Type	SMT	SMT	SMT	EBMT	SMT	SMT	RBMT

Table 2: Pearson's correlation coefficient in the sentence-level adequacy.

	tori	FDU -MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
IMPACT	**0.7639**	0.5276	**0.4487**	0.5980	**0.5371**	**0.6371**	**0.6255**	**0.7249**
ROUGE-L	0.7597	0.4840	0.4264	**0.6111**	0.5229	0.6183	0.5927	0.7079
BLEU	0.6473	0.4469	0.2463	0.4230	0.4336	0.3727	0.4124	0.5340
NIST	0.5135	0.3380	0.2756	0.4142	0.3086	0.2553	0.2300	0.3628
WN	0.5003	0.2995	0.3122	0.4317	0.3684	0.4586	0.4886	0.4227
NMG-WN	0.7010	0.4606	0.3432	0.6067	0.4719	0.5441	0.5885	0.5906
METEOR	0.4509	0.3121	0.0892	0.3907	0.2781	0.3120	0.2744	0.3937
WER	0.7464	**0.5405**	0.4114	0.5519	0.5185	0.5461	0.5970	0.6902

	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.
IMPACT	**0.7007**	0.5902	**0.7125**	0.5981	**0.7621**	0.5345	**0.6258**
ROUGE-L	0.6834	**0.5958**	0.7042	0.5691	0.7480	0.5293	0.6109
BLEU	0.5188	0.3469	0.5884	0.3697	0.5459	0.4357	0.4515
NIST	0.4218	0.2622	0.4092	0.1721	0.3521	0.4769	0.3423
WN	0.5211	0.3560	0.4268	0.5558	0.5154	0.4565	0.4367
NMG-WN	0.6658	0.4718	0.6068	**0.6116**	0.6770	**0.5740**	0.5653
METEOR	0.3881	0.3058	0.4947	0.3127	0.2987	0.4162	0.3370
WER	0.6656	0.5341	0.6570	0.5740	0.7491	0.5301	0.5937

all automatic methods and human judges evaluated English translation sentences. Table 1 presents groups and types of machine translation systems.

As presented in Table 1, SMT, RBMT, and EBMT respectively represent a statistical machine translation, a rule-based machine translation, and an example-based machine translation. Four English references by bilingual humans were used for each translated English sentence to calculate the scores.

Three human judges evaluated all translation sentences ($1,400 = 100 \times 14$ MT systems). In that case, the human judges scored all translation sentences from the perspective of fluency and adequacy on a scale of 1–5. Moreover, we used the median value in the results of three human judges as the final scores of 1–5.

3.2 Experimental Results

For this meta-evaluation, we calculated the respective Pearson's correlation and the Spearman's rank correlation coefficients between the scores of the automatic evaluation methods and the scores of human judgments in sentence-level adequacy and fluency. Table 2 shows the Pearson's correlation coefficient for sentence-level adequacy. Table 3 presents the Pearson's correlation coefficient for sentence-level fluency. Table 4 shows the Spearman's rank correlation coefficient for sentence-level adequacy. Table 5 shows the Spearman's rank correlation coefficient for sentence-level fluency. In Tables 2–5, the ab-

Table 3: Pearson's correlation coefficient in sentence-level fluency.

	tori	FDU -MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
IMPACT	0.5581	0.3621	**0.3407**	0.5821	0.4586	**0.5768**	0.4852	**0.6896**
ROUGE-L	0.5551	0.3156	0.3056	**0.5925**	0.4391	0.5666	0.4475	0.6756
BLEU	0.4793	0.3142	0.0963	0.4488	0.3033	0.4690	0.3602	0.5272
NIST	0.4139	0.2091	0.0257	0.4987	0.1682	0.3923	0.2236	0.3749
WN	0.4730	0.3305	0.2974	0.3371	0.3923	0.3571	0.4450	0.3313
NMG-WN	**0.5782**	**0.3824**	0.3090	0.5434	**0.4680**	0.5070	**0.5234**	0.5363
METEOR	0.4050	0.1608	0.1405	0.4420	0.1825	0.4259	0.2336	0.4873
WER	0.5143	0.3485	0.3031	0.5220	0.4262	0.4936	0.4405	0.6351

	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.
IMPACT	**0.5612**	**0.5458**	0.6320	0.3492	0.6034	0.4166	**0.5115**
ROUGE-L	0.5414	0.5360	0.6347	0.3231	0.5889	0.4127	0.4953
BLEU	0.5040	0.4225	0.5521	0.2134	0.4783	0.4078	0.3983
NIST	0.3682	0.2584	0.3811	0.1682	0.3116	**0.4484**	0.3030
WN	0.3839	0.3514	0.4248	0.4413	0.4772	0.2770	0.3799
NMG-WN	0.5526	0.5138	0.5799	**0.4509**	**0.6308**	0.4124	0.4992
METEOR	0.2511	0.3245	0.4153	0.1376	0.3351	0.2902	0.3022
WER	0.5492	0.5312	**0.6421**	0.3962	0.6228	0.4063	0.4879

solute values are used as the correlation coefficients. Therefore, the correlation coefficients between the scores of the automatic evaluation method and the scores of human judgments are high when the values in the tables are high. Bold typeface signifies the maximum correlation coefficients among eight automatic evaluation methods. In Tables 2–5, "Avg." shows the average of the correlation coefficients obtained by 14 machine translation systems in each automatic evaluation method.

3.3 Discussion

Results of the meta-evaluation confirmed that some automatic evaluation methods can yield high correlation in sentence-level adequacy. In "Avg." presented in Tables 2 and 4, the correlation coefficients of IMPACT, ROUGE-L and NMG_WN are about 0.6. In Tables 2 and 4, the values of average for "Avg." by eight automatic evaluation methods are, respectively, 0.4954 and 0.4804. However, in Tables 3 and 5, the average values for "Avg." by eight automatic evaluation methods are, respectively, 0.4222 and 0.4047. In Pearson's correlation coefficient, eight "Avg." of sentence-level fluency in Table 3 are lower than the eight "Avg." of sentence-level adequacy in Table 2. In Spearman's rank correlation coefficient, eight "Avg." of sentence-level fluency in Table 5 are lower than eight "Avg." of sentence-level adequacy in Table 4, as they were for Pearson's correlation coefficient. These results show that the correlation in sentence-level fluency is quite lower than the correlation in sentence-level adequacy.

Moreover, we respectively investigated the correlation coefficients of SMT (*i.e.*, tori, FDU-MC and WI, HIT2, KLE, MIT, NAIST-NTT, NiCT-ATR, NTT, TH, MIBEL and Moses), and RBMT (*i.e.*, JAPIO and tsbmt). In use of the automatic evaluation methods, the evaluation for sentences translated by RBMT is known to be inadequate compared with that for the sentences translated using SMT. Table 6 portrays the Pearson's correlation coefficient of SMT and RBMT. Table 7 presents the Spearman's rank correlation coefficient of SMT and RBMT. In that case, we calculated the correlation coefficient between the scores of the automatic eval-

Table 4: Spearman's rank correlation coefficient in sentence-level adequacy.

	tori	FDU -MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
IMPACT	0.7336	0.4604	**0.4881**	0.5992	**0.4741**	**0.6382**	**0.5841**	**0.6409**
ROUGE-L	0.7304	0.4327	0.4822	**0.6092**	0.4572	0.6135	0.5365	0.6368
BLEU	0.5525	0.4557	0.2206	0.4327	0.3449	0.3230	0.2805	0.4375
NIST	0.5032	0.3363	0.2438	0.4218	0.2489	0.2342	0.1534	0.3529
WN	0.5586	0.3224	0.3821	0.3178	0.3843	0.4847	0.5106	0.5009
NMG-WN	**0.7541**	0.4758	0.3829	0.5579	0.4472	0.5560	0.5828	0.6263
METEOR	0.4409	0.2689	0.1509	0.4018	0.2580	0.3085	0.1991	0.4115
WER	0.6566	**0.4966**	0.4147	0.5478	0.4272	0.5524	0.4884	0.5539

	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.
IMPACT	0.6703	0.5627	**0.7067**	0.5617	**0.7411**	0.5583	**0.6014**
ROUGE-L	0.6603	**0.5732**	0.6983	0.5340	0.7280	0.5281	0.5872
BLEU	0.4571	0.2813	0.5827	0.3220	0.4987	0.4302	0.4014
NIST	0.4255	0.3095	0.4424	0.1313	0.2950	0.4785	0.3269
WN	0.5584	0.3698	0.4162	0.6290	0.5504	0.3966	0.4558
NMG-WN	**0.6863**	0.5013	0.6524	**0.6412**	0.7015	**0.5728**	0.5813
METEOR	0.4242	0.3714	0.4776	0.3335	0.2861	0.4455	0.3413
WER	0.6234	0.4322	0.6480	0.5463	0.7131	0.5684	0.5478

uation methods and the scores of human judgments using the sentences translated by SMT ($1,100 = 100 \times 11$ SMT systems) and the sentences translated by RBMT ($200 = 100 \times 2$ RBMT systems), respectively. In "Avg." of Tables 6 and 7, we confirmed that the correlation coefficients in RBMT are lower than those in SMT. Especially, the correlation of sentence-level fluency in RBMT is insufficient (about 0.44).

Table 8 presents examples of sentence-level fluency in RBMT. For Table 8, we selected the scores of IMPACT and NMG_WN because they indicated the highest correlation coefficients among eight automatic evaluation methods. In IMPACT and NMG_WN, the evaluation for translation sentences is high when their scores are high. Example No. 1 in Table 8 is an example for which the scores of IMPACT and NMG_WN are low, but for which the score of human judgment is high. In such a case, the difference between the scores of the automatic evaluation methods and the scores of human judgment depend on the conjugated forms and the ambiguous words (*e.g.* "percentage" and "ratio" in example No. 1). Example No. 2 of Table 8 shows the example for which the scores of IMPACT and NMG_WN are high, although the score of human judgment is low because it is difficult for automatic evaluation methods to determine whether the translation sentence corresponds grammatically to the references. In particular, RBMT generates correct or erroneous translation sentences that differ from references in the phrase sequence. Therefore, automatic evaluation methods that can use grammatical knowledge effectively are necessary for sentence-level evaluation.

Patent translation involves the processing of many long sentences. Therefore, it is difficult to obtain higher correlation coefficient of sentence-level fluency in the patent translation. In evaluation experiments using news articles, 0.4552 and 0.5246 were obtained as Pearson's correlation coefficients for the sentence-level adequacy and fluency respectively when IMPACT was used as the automatic evaluation method (Echizen-ya and Araki, 2007): in

Table 5: Spearman's rank correlation coefficient in sentence-level fluency.

	tori	FDU -MCandWI	HIT2	JAPIO	KLE	MIT	NAIST -NTT	NiCT -ATR
IMPACT	0.5481	0.3419	0.3285	0.5572	0.3976	**0.5960**	0.4317	**0.6334**
ROUGE-L	0.5470	0.3087	0.3041	**0.5646**	0.3661	0.5638	0.3879	0.6255
BLEU	0.4157	0.3472	0.0559	0.4286	0.2018	0.4475	0.2569	0.4909
NIST	0.4209	0.2456	0.0185	0.4559	0.1093	0.3186	0.1898	0.3634
WN	0.4645	0.3051	0.3224	0.3058	0.3993	0.3421	0.4653	0.3503
NMG-WN	**0.5569**	**0.3726**	**0.3461**	0.5381	**0.4300**	0.5052	**0.5264**	0.5328
METEOR	0.4608	0.1533	0.1429	0.4438	0.1783	0.4073	0.1596	0.4821
WER	0.4469	0.2889	0.2395	0.5087	0.3292	0.4995	0.3482	0.5637

	NTT	TH	Kyoto-U	MIBEL	Moses	tsbmt	Avg.
IMPACT	0.5471	0.4635	0.6454	0.3222	0.6319	0.4358	0.4915
ROUGE-L	0.5246	0.4579	0.6428	0.2949	0.6159	0.3928	0.4712
BLEU	0.4882	0.3497	0.5419	0.1407	0.4740	0.4176	0.3612
NIST	0.4150	0.2521	0.4193	0.0889	0.3006	**0.4752**	0.2909
WN	0.3519	0.3284	0.3819	**0.4783**	0.4610	0.2902	0.3747
NMG-WN	**0.5684**	**0.4935**	0.5850	0.4451	**0.6502**	0.4387	**0.4992**
METEOR	0.2911	0.3176	0.4267	0.1735	0.3264	0.3512	0.3034
WER	0.5320	0.3958	**0.6505**	0.3828	0.6501	0.4003	0.4454

sentence-level adequacy, the correlation coefficient of IMPACT using the patent translation (0.6258 in Table 2) is higher than that using the news articles (0.4552). However, for sentence-level fluency, the correlation coefficient of IMPACT using the patent translation (0.5115 in Table 3) is lower than that using the news articles (0.5246). These results indicate that the correlation coefficient of sentence-level fluency in the patent translation is insufficient. Consequently, it is important to realize automatic evaluation methods that can deal efficiently with grammatical knowledge.

4 Conclusion

As described herein, we explained meta-evaluation based on various automatic evaluation methods for machine translation using patent translation data in NTCIR-7. The experimental results indicate that some automatic evaluation methods can achieve a high correlation of sentence-level adequacy. However, the correlation of sentence-level fluency in RBMT is insufficient in particular.

Future studies will develop an automatic evaluation method that can use grammatical knowledge efficiently to improve the correlation of sentence-level fluency in RBMT.

Acknowledgments

This work was done under research in the AAMT/JAPIO Special Interest Group on Patent Translation. The authors express their sincere acknowledgements to group members for their useful discussions. The author sincerely acknowledges the support the Japan Patent Information Organization (JAPIO), who provided the corpora used for this work. This work was performed as a joint research project with the National Institute of Informatics.

References

Banerjee, Satanjeev, and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. pp.65–72. *In Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.*

Table 6: Pearson's correlation coefficient of SMT and RBMT.

	Adequacy		Fluency	
	SMT	RBMT	SMT	RBMT
IMPACT	**0.6858**	0.5650	**0.5676**	**0.4960**
ROUGE-L	0.6719	0.5691	0.5523	0.4988
BLEU	0.5291	0.4263	0.4749	0.4238
NIST	0.4272	0.4432	0.3966	0.4706
WN	0.3875	0.4370	0.3263	0.3042
NMG-WN	0.5845	**0.5850**	0.5184	0.4732
METEOR	0.4098	0.4019	0.3976	0.3625
WER	0.6435	0.5375	0.5255	0.4595
Avg.	0.5425	0.4956	0.4699	0.4361

Table 7: Spearman's rank correlation coefficient of SMT and RBMT.

	Adequacy		Fluency	
	SMT	RBMT	SMT	RBMT
IMPACT	**0.6506**	**0.5811**	**0.5482**	0.4951
ROUGE-L	0.6363	0.5701	0.5292	0.4783
BLEU	0.4625	0.4364	0.4347	0.4213
NIST	0.3995	0.4505	0.3706	0.4597
WN	0.4150	0.3596	0.3309	0.3016
NMG-WN	0.6020	0.5755	0.5283	**0.4959**
METEOR	0.4005	0.4245	0.3813	0.3943
WER	0.5753	0.5573	0.4803	0.4512
Avg.	0.5177	0.4944	0.4504	0.4372

Coughlin, Deborah. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. pp.63–70. *In Proc. of MT Summit IX.*

Echizen-ya, Hiroshi, and Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. pp.151–158. *In Proc. of MT Summit XII.*

Ehara, Terumasa. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. pp.13–18. *In Proc. of MT Summit XII Workshop on Patent Translation.*

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. pp.389–400. *In Proc. of 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access.*

Gamon, Michael, Anthony Aue and Martine Smets. 2005. Sentence-level evaluation without reference translations: Beyond language modeling. pp.103–111. *In Proc. of EAMT'05.*

Kluesza, Alex, and Stuart M. Shieber. 2004. A Learning Approach Improving Sentence-Level MT Evaluation. pp.75–84. *In Proc. of TMI'04.*

Leusch, Gregor, Nicola Ueffing and Hermann Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. pp.240–247. *In Proc. of MT Summit IX.*

Lin, Chin-Yew and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip–Bigram Statistics. pp.606–613. *In Proc. of ACL'04.*

Mutton, Andrew., Mark Dras, Stephen Wan and Robert Dale. 2007. GLEU: Automatic Evaluation of Sentence-Level Fluency. pp.344–351. *In Proc. of ACL'07.*

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.

Table 8: Examples of automatic evaluation based on sentence-level fluency in RBMT.

Example No. 1		human	IMPACT	NMG_WN
source sentence	これらのガスは、所定の割合で混合して用いてもよい。	4	0.3917	-0.6988
translation sentence	you may use these gases mixing it by the given percentage.			
reference No. 1	these gases might be used in mixture in a prescribed proportion.			
reference No. 2	these gases might be used by mixing at a predetermined percentage.			
reference No. 3	these gases might be mixed at a prescribed ratio and be used.			
reference No. 4	those gases can be mixed and used at a predetermined ratio.			
Example No. 2		human	IMPACT	NMG_WN
source sentence	なお、図中同一または相当部分には同一符号を付してその説明は繰返さない。	2	0.5221	-0.4284
translation sentence	the same code is fixed to identical or the equivalent part in figure and the clarification is not repeated.			
reference No.1	the same or corresponding portion in the drawings will be represented by the same reference character and the description thereof will not be repeated.			
reference No.2	it is noteworthy that, in the figure, the same code is given to the same or equivalent area, and the explanation is not repeated.			
reference No.3	incidentally, the same or corresponding parts in the figure will be referred to by the same symbols, and descriptions thereof will not be repeated.			
reference No. 4	moreover, the same code is assigned to the same portion or an equivalent portion in the drawing and the description is not repeated.			

http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.

Papineni, Kishore., Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. pp.311–318. *In Proc. of ACL'02.*