Colouring Summaries BLEU

Katerina Pastra
Department of Computer Science
University of Sheffield
katerina@dcs.shef.ac.uk

Horacio Saggion
Department of Computing Science
University of Sheffield
saggion@dcs.shef.ac.uk

Abstract

In this paper we attempt to apply the IBM algorithm, BLEU, to the output of four different summarizers in order to perform an intrinsic evaluation of their output. The objective of this experiment is to explore whether a metric, originally developed for the evaluation of machine translation output, could be used for assessing another type of output reliably. Changing the type of text to be evaluated by BLEU into automatically generated extracts and setting the conditions and parameters of the evaluation experiment according to the idiosyncrasies of the task, we put the feasibility of porting BLEU in different Natural Language Processing research areas under test. Furthermore, some important conclusions relevant to the resources needed for evaluating summaries have come up as a side-effect of running the whole experiment.

1 Introduction

Machine Translation and Automatic Summarization are two very different Natural Language Processing (NLP) tasks with -among others- different implementation needs and goals. They both aim at generating text; however, the properties and characteristics of these target texts vary considerably. Simply put, in Machine Translation, the generated document should be an accurate and fluent translation of the original document, in the target language. In Summarization, the generated text should be an informative, reduced version of the original document (single-document summary), or sets of documents (multi-document summary) in the form of an abstract, or an extract. Abstracts present an overview of the main points expressed in the original document, while extracts consist of a number of informative sentences taken directly from the source document. The fact that, by their very nature, automatically generated extracts carry the single sentence qualities of the source documents[1], may lead one to the conclusion that evaluating this type of text is trivial, as compared to the evaluation of abstracts or even machine translation, since in the latter, one needs to be able to evaluate the content of the generated translation in terms of grammaticality, semantic equivalence to the source document and other quality characteristics (Hovy et al., 2002).

Though the evaluation of generated extracts is not as demanding as the evaluation of Machine Translation, it does have two critical idiosyncratic aspects that render the evaluation task difficult:

- the compression level (word or sentence level) and the compression rate of the source document must be determined for the selection of the contents of the extract ; the values of these variables may greatly affect the whole evaluation setup and the results obtained

[1] Even if coherence issues may arise beyond the sentence boundaries i.e. at the text level

- the very low agreement among human evaluators on what is considered to be "important information" for inclusion in the extract, reaching sometimes the point of total disagreement on the focus of the extract (Mani, 2001; Mani et al., 2001). The nature of this disagreement on the adequacy of the extracts is such that - by definition - cannot manifest itself in Machine Translation; this is because it refers to the adequacy of the contents chosen to form the extract, rather than what constitutes an adequate way of expressing all the contents of the source document in a target language.

The difference on the parameters to be taken into consideration when performing evaluation within these two NLP tasks presents a challenge for porting evaluation metrics from the one research area to the other. Given the relatively recent success in achieving high correlations with human judgement for Machine Translation evaluation, using the IBM content-based evaluation metric, BLEU (Papineni et al., 2001), we attempt to run this same metric on system generated extracts; this way we explore whether BLEU can be used reliably in this research area and if so, which testing parameters need to be taken into consideration. First, we refer briefly to BLEU and its use across different NLP areas, then we locate our experiments relatively to this related work and we describe the resources we used, the tools we developed and the parameters we set for running the experiments. The description of these experiments and the interpretation of the results follows. The paper concludes with some preliminary observations we make as a result of this restricted, first experimentation.

2 Using BLEU in NLP

Being an intrinsic evaluation measure (Sparck Jones and Galliers, 1995), BLEU compares the content of a machine translation against an "ideal" translation. It is based on a "weighted average of similar length phrase matches" (n-grams), it is sensitive to longer n-grams (the baseline being the use of up to 4-grams) and it also includes a brevity penalty factor for penalising shorter than the "gold standard" translations (Papineni et al., 2001; Doddington, 2002). The metric has been found to highly correlate with human judgement, being at the same time reliable even when run on different documents and against different number of model references. Experiments run by NIST (Doddington, 2002), checking the metric for consistency and sensitivity, verified these findings and showed that the metric distinguishes, indeed, between quite similar systems. A slightly different version of BLEU has been suggested by the same people, which still needs to be put into comparative testing with BLEU before any claims for its performance are made.

BLEU has been used for evaluating different types of NLP output to a small extent. In (Zajic et al., 2002), the algorithm has been used in a specific Natural Language Generation application: headline generation. The purpose of this work was to use an automated metric for evaluating a system generated headline against a human generated one, in order to draw conclusions on the parameters that affect the performance of a system and improve scoring similarity. In (Lin and Hovy, 2002) BLEU has been applied on summarization. The authors argue on the unstable and unreliable nature of manual evaluation and the low agreement among humans on the contents of a reference summary. Lin and Hovy make the case that automated metrics are necessary and test their own modified recall metric, along with BLEU itself, on single and multi-document summaries and compare the results with human judgement. Modified recall seems to reach very high correlation scores, though direct comparative experimentation is needed for drawing conclusions on its performance in relation to BLEU. The latter, has been shown to achieve 0.66 correlation in single-document summaries at 100 words compression rate and against a single reference summary. The correlation achieved by BLEU climbs up to 0.82 when BLEU is run over and compared against multiply judged document units, that could be thought of as a sort of multiple reference summaries. The correlation scores for multi-document summaries are similar. Therefore, BLEU has been found to correlate quite highly with human judge-

ment for the summarization task when multiple judgement is involved, while -as Lin and Hovy indicate- using a single reference is not adequate for getting reliable results with high correlation with the human evaluators.

It is this conclusion that Lin and Hovy have drawn, that contradicts findings by the IBM and NIST people for the importance of using multiple references when using BLEU in Machine Translation. The use of either multiple references or just a single reference has been proved not to affect the reliability of the results provided by BLEU (Papineni et al., 2001; Doddington, 2002), which seems not to be the case in summarization. This is not a surprise; comparisons of content-based metrics for summarization in (Donaway et al., 2000) have led the authors to the conclusion that such metrics correlate highly with human judgement when the humans do not disagree substantially. The fact that more than one reference summaries are needed because of the low agreement between human evaluators has been repeatedly indicated in automatic summarization evaluation (Mani, 2001).

We attempt to test BLEU's reliability when changing various evaluation parameters such as the source documents, the reference summaries used and even parameters unique to the evaluation of summaries, such as the compression rate of the extract. In doing so, we explore whether the metric is indeed reliable only when using more than a single reference and whether any other testing parameter could compensate for lack of multiple references, if used appropriately.

3 Evaluation Experiment

In this section, we will present a description of the experiments themselves, along with the results obtained and their analysis, preceded by information on the corpus we used for our experiments and the tools we developed for setting their parameters and running them automatically.

3.1 Testing corpus

We make use of part of the language resources (HKNews Corpus) developed during the 2001 Workshop on Automatic Summarization of Multiple (Multilingual) Documents (Saggion et al., 2002).

The documents of each cluster are all relevant to a specific topic-query, so that they form, in fact, thematic clusters. The texts are marked up on the paragraph, sentence and word level. Annotations with linguistic information (Part of speech tags and morphological information), though marked up on the documents have not been used in our experiments at all. Three judges have assessed the sentences in each cluster and have provided a score on a scale from 0 to 10 (i.e. utility judgement), expressing how important the sentence is for the topic of the cluster (Radev et al., 2000). In our experiments, we have used three document clusters, each consisting of ten documents in English.

3.2 Summarizers

It is important to note, that our objective is not to demonstrate how a particular summarization methodology performs, but to analyse an evaluation metric. The summaries used for the evaluation were produced as extracts at different 'sentence' (and not word) compression rates[2]. In order to produce summarizers for our evaluation, we use a robust summarisation system (Saggion, 2002) that makes use of components for semantic tagging and coreference resolution developed within the GATE architecture (Cunningham et al., 2002). The system combines GATE components with well established statistical techniques developed for the purpose of text summarisation research. The system supports "generic" and query-based summarisation addressing the need for user adaptation[3]. For each sentence, the system computes values for a number of 'shallow' summarization features: position of the sentence, term distribution analysis, similarity of the sentence with the document, similarity with the sentence at the leading part of the document, similarity of the sentence with the query, named entity distribution analysis, statistic cohesion, etc. The values of these features are linearly combined to produce the sentence fi-

[2]We have to note that the level of compression i.e sentence or word level, affects probably the evaluation of the summarizers' output. Comparative testing could indicate whether this is a crucial parameter for system evaluation.

[3]The software can be obtained from `http://www.dcs.shef.ac.uk/~saggion`

nal score. Top-ranked sentences are annotated until the target n% compression is achieved (an annotation set is produced for each summary that is generated). Different summarization systems can be deployed by setting-up the weights that participate in the scoring formula. Note that as the summarization components are not aware of the compression parameter, one would expect specific configurations to produce good extracts at different compression rates and across documents.

We have configured four different summarizers, namely, the "query-based system" that computes the similarity of each sentence of the source document with the documents topic-query, in order to decide whether to include a sentence in the generated extract or not. We also have the "Simple 1 system", whose main feature is that it computes the similarity of a sentence with the whole document, the "Simple 2 system" which is a lead based summarizer and the "Simple 3 system" that blindly extracts the last part of the source document.

3.3 Judge-based Summaries

Following the same methodology used in (Saggion et al., 2002), we implemented a judge-based summarization system that given a judge number (1, 2, 3, or all), it scores sentences based on a combination of the utility that the sentence has according to the judge (or the sum of the utilities if 'all') and the position of the sentence (leading sentences are preferred). These 'extracts' represent our gold-standards for evaluation in our experiments. In order to use the documents in a stand-alone way, we have enriched the initial corpus mark-up and added to each document information about cluster number, cluster topic (or query) and all the information about utility judgement (that information was kept in separate files in the original HKNews corpus).

3.4 Evaluation Software

We have developed a number of software components to facilitate the evaluation and we make use of the GATE development environment for testing and processing. The evaluation package allows the user to specify different reference extracts (judge-based summarizers) and summarization systems to be compared.

Co-selection comparison (i.e., precision and recall) is being done with modules obtained from the GATE library (AnnotationDiff components). Content-based comparison by the Bleu algorithm was implemented as a Java class. The exact formula provided by the developers of BLEU has been implemented following the baseline configurations i.e use of 4-grams and uniform weights summing to 1:

$$Bleu(S,R) = K(S,R) * e^{Bleu_1(S,R)}$$

$$Bleu_1(S,R) = \sum_{i=1,2,...n} w_i * \lg(\frac{|(S_i \bigcap R_i)|}{|S_i|})$$

$$K(S,R) = \begin{cases} 1 & \text{if } |S| > |R| \\ e^{(1-\frac{|R|}{|S|})} & \text{otherwise} \end{cases}$$

$$w_i = \frac{i}{\sum_{j=1,2,...n} j} \qquad \text{for } i = 1,2,...,n$$

where S and R are the system and reference sets. S_i and R_i are the "bags" of i-grams for system and reference. n is a parameter of our implementation, but for the purpose of our experiments we have set n to 4.

3.5 Experiments

In our experiments we have treated compression rates and clusters as variables each one being a condition for the other and both dependent to a third variable, the gold standard summary. We ran BLEU in all different combinations in order to see the main effects of each combination and the interactions among them. In particular, we have used three different text clusters, consisting of texts that refer to the same topic: cluster 1197 on "Museum exhibits and hours", cluster 125 which deals with "Narcotics and rehabilitation" and cluster 241 which refers to "Fire safety and building management". For the texts of each cluster we have three different reference summaries (created according to the utility judgement score assigned by human evaluators cf. 3.1 and 3.2). We will refer to these as Reference1, Reference2 and Reference3. The judges behind these references are

all the same for the three text clusters with one exception: Reference1 in cluster 241 has not been created by the same human evaluator as the Reference 1 summaries for the other two clusters. Last, we ran the experiments at five different compression rates [4]: 10%, 20%, 30%, 40% and 50%.

We first ran BLEU on the reference summaries in order to check whether BLEU is consistent in the data it produces concerning the agreement among human evaluators. We tried all possible combinations for comparing the reference summaries; using at first Reference 1 as the gold standard, we ran BLEU over References 2 and 3 and we did this for two clusters (since the third's -241- Reference 1 set of summaries had been created by another judge - a fourth one). We did this for all five compression rates separately. We repeated the experiment changing the gold standard and the references to be scored accordingly (i.e Reference 1 and 3 against 2, Reference 1 and 2 against 3). The results we got were consistent neither across clusters, nor within clusters across compression rates; however the latter, did show a general tendency for consistency which allows for some observations to be made. In cluster 1197, References 1 and 2 are generally in higher agreement than with 3, a fact verified regardless the reference chosen as a gold standard. The fact that References 1 and 2 are very close was also evident when both compared against Reference 3; though the latter is generally closer to Reference 2, the scores assigned to Reference 1 and 2 are extremely close. In cluster 125, Reference 1 is consistently closer to 3, while 2 is closer to 1 at some compression rates and closer to 3 at others. These very close scores indicate that all three references are similarly "distant" one from another, and no groupings of agreement can actually be made. Agreement between reference summaries augments as the compression rate also increases, with the higher similarity scores always found at the 50% compression rate and the lower ones consistently found at 10%. Table 1 shows a consistent ranking across compression rates in cluster 1197 and an inconsistent one in cluster 125, using in both cases Reference 2 as the gold standard. From this first experiment, the rankings of the reference summaries seem to depend on the different values of the variables used. If that is the case, then one should use BLEU in summarization only when determining specific values for the evaluation experiment, that will guarantee reliable results; but how could one determine which value(s) should be chosen? To explore things further we decided to proceed with a second experiment set up in a similar way.

In our second experiment we try to compare the system generated extracts (and therefore the performance of the four summarizers) against the different human references. Again, the different rounds of the experiment involve multiple parameters; the generated extracts of all three text clusters are compared against each reference summary, against all reference summaries (integrated summary) and at all five compression rates. Going through the different stages of this experiment we observe that:

- For Reference X within Cluster Y across Compressions, the ranking of the systems is not consistent

One does not get the same system ranking at different compression rates. The similarity of a generated extract to a specific reference summary is the same at some compression rates, similar at others (e.g the order of two of the systems swaps) and totally different at other rates. No patterns arise in the way that rankings are similar at specific compression rates; for example, in table 2, there seems to be a prevailing ranking common in four compression rates; however, the ranking provided at 10% is totally different, and no apparent reason seems to justify this deviation (e.g. very close scores). Furthermore, this agreement among the four highest compression rates does not form a pattern i.e it does not appear as such across clusters or references.

- For Reference X at Compression Y across Clusters, the ranking of the systems is not consistent

In our experiments we were able to observe 15 different realisations of these testing configurations and hardly did a case of consistency at a compression rate across clusters appeared.

[4] In our experiments compression is always performed at the sentence level

Ref 2 - 1197	**10%**	**20%**	**30%**	**40%**	**50%**
Reference 1	0.50 - 1	0.67 - 1	0.73 - 1	0.73 - 1	0.79 - 1
Reference 3	0.34 - 2	0.51 - 2	0.52 - 2	0.63 - 2	0.69 - 2
Ref 2 - 125	**10%**	**20%**	**30%**	**40%**	**50%**
Reference 1	0.36 - 1	0.41 - 1	0.59 - 2	0.67 - 2	0.78 - 1
Reference 3	0.20 - 2	0.46 - 2	0.66 - 1	0.73 - 1	0.73 - 2

Table 1: Reference summary similarity scores and rankings across clusters and compression rates

Reference 3	**10%**	**20%**	**30%**	**40%**	**50%**
Query-based	0.44 - 2	0.50 - 1	0.58 - 1	0.66 - 1	0.71 - 1
Simple 1	0.10 - 3	0.23 - 3	0.48 - 3	0.57 - 3	0.64 - 3
Simple 2	0.52 - 1	0.45 - 2	0.53 - 2	0.62 - 2	0.68 - 2
Simple 3	0.03 - 4	0.07 - 4	0.08 - 4	0.11 - 4	0.11 - 4

Table 2: System scores and rankings for cluster 241, against Reference 3, at different compression rates

- For Reference All across Clusters at multiple Compressions, the ranking of the systems is consistent

Estimating similarity scores against Reference All (use of multiple references cf. 3.2), proves to provide reliable, consistent results across clusters and compression rates. Table 3 presents the scores and corresponding system rankings for two different clusters and at the five different compression rates. The prevailing system ranking is [1324], which is what we would intuitively expect according to the features of the summarizers we compare. Some deviations from this ranking are due to very small differences in the similarity scores assigned to the systems[5], which indicates the need for using a larger testing corpus for the experiments.

So, the need for multiple references is evident; BLEU is a consistent, reliable metric, but when used in summarization, one has to apply it to multiple references in order to get reliable results. This is not just a way to improve correlation with human judgement (Lin and Hovy, 2002); it is a crucial evaluation parameter that affects the quality of the automatic evaluation results. In our case we had a balanced set of reference summaries to work with, i.e none of them was too similar to another. The more reference summaries one has and the larger one's testing corpus, the safer the conclusions drawn will be. However, what happens when there is lack of such resources and especially of multiple reference summaries? Is there a way to use BLEU with a single reference summary and still get reliable results back?

Looking at the results of our experiments, when using each reference summary separately as a gold standard, we realised that estimating the average ranking of each system across multiple compression rates might lead to consistent rankings. Following the average rank aggregation techique (Rajman and Hartley, 2001), we transfered the average scores each system got per text cluster at each compression rate into ranks and computed the average rank of each system across all five compression rates per text cluster and against each reference summary. Table 4, shows the average system rankings we got for each system at clusters 1197 and 125, using Reference 1, 2, and 3 separately. [1324] is the average system ranking that is clearly indicated in the vast majority of cases. The two exceptions to this are due to extremely small differences in average scores at specific compression rates and indicate the need for scaling up our experiment, a fact that has already been indicated by the results of our experiment using multiple references (Reference All).

4 Conclusions and Future Work

BLEU has been developed for measuring content similarity in terms of length and wording between texts. For the evaluation of automatically generated extracts, the metric is expected to capture similarities between sentences not shared by both the generated text and the model sum-

[5]For example, at the 10% compression rate, cluster 1197, systems Simple 1 and Simple 2 swap places in the final ranking with a 0.005 difference in their similarity scores

Ref All - 1197	**10%**	**20%**	**30%**	**40%**	**50%**
Query based	0.55 - 1	0.47 - 1	0.49 - 1	0.62 - 1	0.63 - 2
Simple 1	0.3184 - 2	0.32 - 3	0.40 - 3	0.49 - 3	0.62 - 3
Simple 2	0.3134 - 3	0.39 - 2	0.44 - 2	0.56 - 2	0.67 - 1
Simple 3	0.02 - 4	0.03 - 4	0.07 - 4	0.11 - 4	0.13 - 4
Ref All - 125	**10%**	**20%**	**30%**	**40%**	**50%**
Query based	0.44 - 1	0.43 - 1	0.57 - 1	0.72 - 1	0.7641 - 2
Simple 1	0.18 - 3	0.3684 - 2	0.54 - 2	0.60 - 3	0.68 - 3
Simple 2	0.32 - 2	0.3673 - 3	0.44 - 3	0.66 - 2	0.7691 - 1
Simple 3	0.03 - 4	0.06 - 4	0.07 - 4	0.10 - 4	0.14 - 4

Table 3: Systems' similarity scores and rankings using Reference All as gold standard

	10%	**20%**	**30%**	**40%**	**50%**	**Average Rank**
Ref 1 - 125	1324	1234	2134	1324	1234	**1234**
Ref 2 - 125	1324	1324	1324	1324	2314	**1324**
Ref 3 - 125	2314	2314	1324	1324	2314	**2314**
Ref 1 - 1197	1324	2314	1324	1324	2314	**1324**
Ref 2 - 1197	1324	1324	1324	1324	2314	**1324**
Ref 3 - 1197	1324	1324	1324	1324	2314	**1324**

Table 4: Systems' average rankings resulting from ranks at multiple compression rates in clusters 125 and 1197. (Systems assumed to be listed in alphabetical order: Query-based, Simple1, Simple2, Simple3)

mary. Going through the texts scored in the above experiments, we found cases in which BLEU does not actually capture content similarity to such a granularity that a human would. Sometimes, this is because the order of the words forming n-grams differs slightly but still conveys the same meaning (e.g. "...abusers reported..." vs. "...reported abusers...") and most of the times because there is no way to capture cases of synonymy, paraphrasing (e.g. "downward tendency"/"falling trend"/"decrease") and other deeper semantic equivalence (e.g. "number of X" vs. "9,000 of X"). Such phenomena are -of course- expected from a statistical metric which involves no linguistic knowledge at all. Our aim in this paper was to shed some light on the conditions under which the metric performs reliably within summarization, given the different parameters that affect evaluation in this NLP research area. From the results obtained by our preliminary experiments, we have generally concluded that:

- Running BLEU over system generated summaries using a single reference affects the reliability of the results provided by the metric. The use of multiple references is a *sine qua non* for reliable results
- Running BLEU over system generated summaries at multiple compression rates and estimating the average rank of each system might yield consistent and reliable results even with a single reference summary and therefore compensate for lack of multiple reference summaries

In order to draw more safe conclusions, we need to scale our experiments considerably, and this is already in progress. Many research questions need still to be answered, such as how BLEU scores correlate with results produced by other content-based metrics used in summarization and elsewhere. We hope that this preliminary, experimental work on porting evaluation metrics across different NLP research areas will function as a stimulus for extensive and thorough research in this direction.

References

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *ACL 2002*.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence

statistics. In *Proceedings of HLT 2002, Human Language Technology Conference, San Diego, CA*.

R. Donaway, K. Drummey, and L. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the ANLP-NAACL 2000 Workshop on Automatic Summarization, Advanced Natural Language Processing - North American of the Association for Computational Linguistics Conference, Seattle, DC*.

E. Hovy, M. King, and A. Popescu-Belis. 2002. An introduction to machine translation evaluation. In *Proceedings of the LREC 2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics, Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Ch. Lin and E. Hovy. 2002. Manual and automatic evaluation of summariess. In *Proceedings of the ACL 2002 Workshop on Automatic Summarization, Association for Computation Linguistics, Philadelphia, PA*.

I. Mani, T. Firmin, and B. Sundheim. 2001. Summac: A text summarization evaluation. *Natural Language Engineering*.

I. Mani. 2001. Summarization evaluation: an overview. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization, North Chapter of the Association for Computational Linguistics, Pittsburgh, PA*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0 109-022), IBM Research Division.

Dr. Radev, J. Hongyan, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.

M. Rajman and A. Hartley. 2001. Automatically predicting mt system rankings compatible with fluency, adequacy or informativeness scores. In *Proceedings of the MT Summit 2001 Workshop on Machine Translation Evaluation: Who did what to whom, European Association for Machine Translation, Santiago de Compostella, Spain*.

H. Saggion, D. Radev., S. Teufel, L. Wai, and S. Strassel. 2002. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 747–754, Las Palmas, Gran Canaria, Spain.

H. Saggion. 2002. Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA, December, 14.

Karen Sparck Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

D. Zajic, B. Dorr, and R. Schwartz. 2002. Automatic headline generation for newspaper stories. In *Proceedings of the ACL 2002 Workshop on Automatic Summarization, Association for Computation Linguistics, Philadelphia, PA*.