Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation

Yves Lepage **Etienne Denoual**
ATR – Spoken language communication research labs
Keihanna gakken tosi, 619-0288 Kyoto, Japan
{yves.lepage, etienne.denoual}@atr.jp

Abstract

We propose a method that automatically generates paraphrase sets from seed sentences to be used as reference sets in objective machine translation evaluation measures like BLEU and NIST. We measured the quality of the paraphrases produced in an experiment, *i.e.*, (i) their grammaticality: at least 99% correct sentences; (ii) their equivalence in meaning: at least 96% correct paraphrases either by meaning equivalence or entailment; and, (iii) the amount of internal lexical and syntactical variation in a set of paraphrases: slightly superior to that of hand-produced sets. The paraphrase sets produced by this method thus seem adequate as reference sets to be used for MT evaluation.

1 Introduction

We present and evaluate a method to automatically produce paraphrases from seed sentences, from a given linguistic resource. Lexical and syntactical variation among paraphrases is handled through commutations exhibited in proportional analogies, while well-formedness is enforced by filtering with sequences of characters of a certain length. In an experiment, the quality of the paraphrases produced, *i.e.*, (i) their grammaticality, (ii) their equivalence in meaning with the seed sentence, and, (iii) the internal lexical and syntactical variation in a set of paraphrases, was assessed by sampling and objective measures.

2 Motivation

Paraphrases are an important element in the evaluation of many natural language processing tasks. Specifically, in the automatic evaluation of machine translation systems, the quality of translation candidates is judged against reference translations that are paraphrases in the target language. Automatic measures like BLEU (PAPINENI et al., 2001) or NIST (DODDINGTON, 2002) do so by counting sequences of words in such paraphrases.

It is expected that such reference sets contain synonymous sentences (*i.e.*, paraphrases) that explicit possible lexical and syntactical variations in order to cope with translation variations in terms and structures (BABYCH and HARTLEY, 2004).

In order to produce such reference sets, we propose a method to generate paraphrases from a seed sentence where lexical and syntactical variations are handled by the use of commutations as captured by proportional analogies whereas N-sequences are used to enforce fluency of expression and adequacy of meaning.

3 The linguistic resource used

The linguistic resource used in the experiment presented in this paper relies on the C-STAR collection of utterances called Basic Traveler's Expressions[1]. This is a multilingual resource of expressions from the travel and tourism domain that contains 162,318 aligned translations in several languages, among which English. The items are quite short as the following examples show (one line is one item in the corpus), and as the figures in Table 1 show.

[1] http://www.c-star.org/.

Number of ≠ sentences	Avg. size ± std. dev. in characters	in words
97,769	35.14 ± 18.81	6.86 ± 3.57

Table 1: Some statistics about the linguistic resource

Number of ≠ sentences	Avg. size ± std. dev. in characters	in words
42,249	33.15 ± 9.31	6.44 ± 1.90

Table 2: Some statistics about the paraphrases produced

Thank you so much. Keep the change.
Bring plenty of lemon, please.
Please tell me about some interesting places [near here.
Thank you. Please sign here.
How do you spell your name?

The quality of this resource is of at least 99% correct sentences (p-value = 1.92%). The few incorrect sentences contain spelling errors or slight syntactical mistakes.

4 Our paraphrasing methodology

4.1 Our algorithm

The proposed method consists in two phases: firstly, paraphrase detection through equality of translation and secondly, paraphrase generation through linguistic commutations based on the data produced in the first phase:

- Detection: find sentences which share a same translation in the multilingual resource (4.2);
- Generation: produce new sentences by exploiting commutations (4.3); limit combinatorics by contiguity constraints (4.4).

Each of the steps of the previous algorithm is explained in details in the following sections.

4.2 Initialisation by paraphrase detection

In a first phase we initialise our data by paraphrase detection. By definition, paraphrase is an equivalence in meaning, thus, different sentences having the same translation ought to be considered equivalent in meaning, *i.e.*, they are paraphrases[2]. As the linguistic resource used in the present experiment is a multilingual corpus, we have at our disposal the corresponding translations in different languages for each of its sentences. For instance, the following English sentences share a common Japanese translation shown in bold face below. Therefore, they are paraphrases.

A beer, please.	**ビールをください。**
	ビールを一本。
	ビールを一本ください。
Beer, please.	ビール。
	ビールをお願いします。
	ビールをください。
	ビールを下さい。
	ビール一杯ください。
Can I have a beer?	**ビールをください。**
Give me a beer, please.	**ビールをください。**
I would like beer.	**ビールをください。**
I'd like a beer, please.	**ビールをください。**

4.3 Commutation in proportional analogies for paraphrase generation

In a second phase, we implement paraphrase generation. Any given sentence may share commutations with other sentences of the corpus. Such commutations are best seen in analogical relations that explicit syntagmatic and paradigmatic variations (de SAUSSURE, 1995, part 3, chap 4). For instance, the seed sentence

A slice of pizza, please.

[2]This is basically the same approach as (OHTAKE and YAMAMOTO, 2003, p. 3 and 4).

I'd like a beer, please.	:	*A beer, please.*	::	*I'd like a slice of pizza, please.*	:	*A slice of pizza, please.*
I'd like a twin, please.	:	*A twin, please.*	::	*I'd like a slice of pizza, please.*	:	*A slice of pizza, please.*
I'd like a bottle of red wine, please.	:	*A bottle of red wine, please.*	::	*I'd like a slice of pizza, please.*	:	*A slice of pizza, please.*

Table 3: Some analogies formed with sentences of the linguistic resource that show commutations with the sentence *A slice of pizza, please.*

(i) *I'd like a beer, please.*	:	*A beer, please.*	::	*I'd like a slice of pizza, please.*	:	*A slice of pizza, please.*
(ii) *I'd like a beer, please.*	:	*Can I have a beer?*	::	*I'd like a slice of pizza, please.*	:	*x*
(iii) *I'd like a beer, please.*	:	*Can I have a beer?*	::	*I'd like a slice of pizza, please.*	:	*Can I have a slice of pizza?*

Table 4: Generating a paraphrase for the seed sentence *A slice of pizza, please.* using proportional analogies. (i) The original proportional analogy taken from Table 3. (ii) Replacing the sentence *A beer, please.* with one of its paraphrases acquired during the detection phase: *Can I have a beer?* The last sentence of the proportional analogy becomes unknown. (iii) Solving the analogical equation, *i.e.*, generating a paraphrase of *A slice of pizza, please.*

enters in the analogies of Table 3. The replacement of some sentences with known paraphrases in such analogies allows us to produce new sentences. This explains why we needed some paraphrases to start with. For instance, by replacing the sentence:

A beer, please.

with the sentence:

Can I have a beer?

in the first analogy of Table 3, one gets the following analogical equation, that is solved as indicated.

I'd like a beer, please. : *Can I have a beer?* ::
I'd like a slice of pizza, please. : x
$\Rightarrow$ $x =$ *Can I have a slice of pizza?*

It is then legitimate to say that the produced sentence:

Can I have a slice of pizza?

is a paraphrase of the seed sentence (see Table 4).

Such a method alleviates the problem of creating templates from examples which would be used in an ulterior phase of generation (BARZILAY and LEE, 2003). Here, all examples in the corpus *are* potential templates in their actual raw form, with the advantage that the choice of the places where commutations may occur is left to proportional analogy.

4.4 Limitation of combinatorics by contiguity constraints

During paraphrase generation, spurious sentences may be produced. For instance, the replacement in the previous analogy, of the sentence:

A beer, please.

by the following paraphrase detected during the first phase:

A bottle of beer, please.

produces the unfortunate sentence:

**A bottle of slice of pizza, please.*

Moreover, as no complete and valid formalisation of linguistic analogies has yet been proposed, the algorithm used (LEPAGE, 1998) may deliver such unacceptable strings as:

43	*Could we have a table in the corner?*
43	*I'd like a table in the corner.*
43	*We would like a table in the corner.*
28	*Can we have a table in the corner?*
5	*Can I get a table in the corner?*
5	~~*In the corner, please.*~~
4	~~*We'd like to sit in the corner.*~~
2	~~*I'd like to sit in the corner.*~~
2	~~*I would like a table in the corner.*~~
2	*We'd like a table in the corner.*
1	~~*I'd prefer a table in the corner.*~~
1	*I prefer a table in the corner.*

Table 5: Paraphrase candidates for *Can we have a table in the corner?* Candidates filtered out by unseen N-sequences ($N = 20$) are struck out. Notice that the seed sentence itself has been generated again by the method (4th sentence from the top). The figures on the left are the frequencies with which the sentence has been generated.

**A slice of pizzthe, pleaset for tha, please.*

In order to ensure a very high rate of well-formedness among the sentences produced, we require a method that extracts well-formed sentences from the set of generated sentences with a very high precision (to the possible prejudice of the recall).

To this end, we eliminate all sentences containing sequences of characters of a given length unseen in the original data[3]. It is clear that, by adequately tuning the given length, such a method will be able to retain a satisfactory number of sentences that will be undoubtedly correct, at least in the sense of the linguistic resource.

5 Experiments

During the first phase of paraphrase detection, $26,079$ sentences (out of $97,769$) got at least one possibly incorrect paraphrase candidate with an average of 5.35 paraphrases by sentence. However, the distribution is not uniform: 60 sentences get more than 100 paraphrases.

The maximum is reached with 529 paraphrases for the sentence *Sure.* Such a sentence has a variety of meanings depending on the context, which explains the high number of its possible paraphrases as illustrated below.

[3]This is conform to the trend of using N-sequences to assess the quality of outputs of various NLP systems like (LIN and HOVY, 2003) for summary generation, (DODDINGTON, 2002) for machine translation, *etc.*.

Sure. Here you are.
Sure. This way, please.
Certainly, go ahead, please.
I'm sure I will.
No, I don't mind a bit.
Okay. I understand quite well, thank you.
Sounds fine to me.
Yes, I do.
...

However, such an example shows also that the more the paraphrases obtained by this method, the less reliable their quality.

During the second phase of paraphrase generation, the method generated $4,495,266$ English sentences on our linguistic resource. An inspection of a sample of 400 sentences shows that the quality lies around 23.6% of correct sentences (p-value = 1.19%) in syntax and meaning. The set of paraphrase candidates obtained on an example sentence are shown in Table 5.

To ensure fluency of expression and adequacy of meaning, the method then filtered out any sentence containing an N-sequence unseen in the corpus (see Section 4.4). The best value for N that allowed us to obtain a quality rate at the same level to that of the original linguistic resource was 20.

As a final result, the number of seed sentences for which we obtained at least one paraphrase is $16,153$. With a total number of $147,708$ para-

phrases generated[4], the average number of paraphrases per sentence is 8.65 with a standard deviation of 16.98 which means that the distribution is unbalanced. The graph on the left of Figure 1 shows the number of seed sentences with the same number of paraphrases. while the graph on the right shows the number of paraphrases against the length of the seed sentence in words.

6 Quality of the generated paraphrases

6.1 Well-formedness of the generated paraphrases

The grammatical quality of the paraphrase candidates obtained was evaluated on a sample of 400 sentences: at least 99% of the paraphrases may be considered grammatically correct (p-value $= 2.22\%$). This quality is approximately the same as that of the original resource: at least 99% (p-value $= 1.92\%$).

An overview of the errors in the generated paraphrases suggests that they do not differ from the ones in the original data. For instance, one notes that an article is lacking before the noun phrase *tourist area* in the following sentence:

Where is tourist area?

Although we are not able to trace the error back to its origin, such a mistake is certainly due to a commutation with a sentence like:

Where is information office?

that contains a similar mistake and that is found in the original linguistic resource.

6.2 Equivalence in content between generated paraphases and seed sentence

The semantic quality of the paraphrases produced was also checked by hand on a sample of 470 paraphrases that were compared with their corresponding seed sentence. We not only checked for strict equivalence, but also for meaning entailment[5].

[4]The same sentence may have been generated several times for different seed sentences. Overall there were $42,249$ different sentences generated. Their lengths in characters and words are given in Table 2.

[5]Bill Dolan, Chris Brockett, and Chris Quirk, Microsoft Research Paraphrase Corpus, `http://-research.microsoft.com/research/nlp-/msr_paraphrase.htm`.

The following three paraphrases on the left with their corresponding seed sentences on the right are examples that were judged to be strict equivalences.

Can I see some ID?	*Could you show me some ID?*
Please exchange this.	*Could you exchange this, please.*
Please send it to Japan.	*Send it to Japan, please.*

The following are examples in which there is a lack of information either in the paraphrase produced or in the seed sentence. This is precisely what entailment is.

Coke, please.	*Miss, could I have a coke?*
I want to change money.	*Please exchange this.*
Sunny-side up, please.	*Fried eggs, sunny-side up, please.*

The result of the sampling is that the paraphrase candidates can be considered valid paraphrases in at least 94% of the cases either by equivalence or entailment (p-value $= 3.05\%$). The following sentences exemplify the remaining cases where two sentences were not judged valid paraphrases of one another.

Do you charge extra if I drop it off?	*There will be a drop off charge.*
Here's one for you, sir.	*You can get one here.*
There it is.	*Yes, please sit down.*

Table 6 summarises the distribution of paraphrase candidates according to the abovementionned classification.

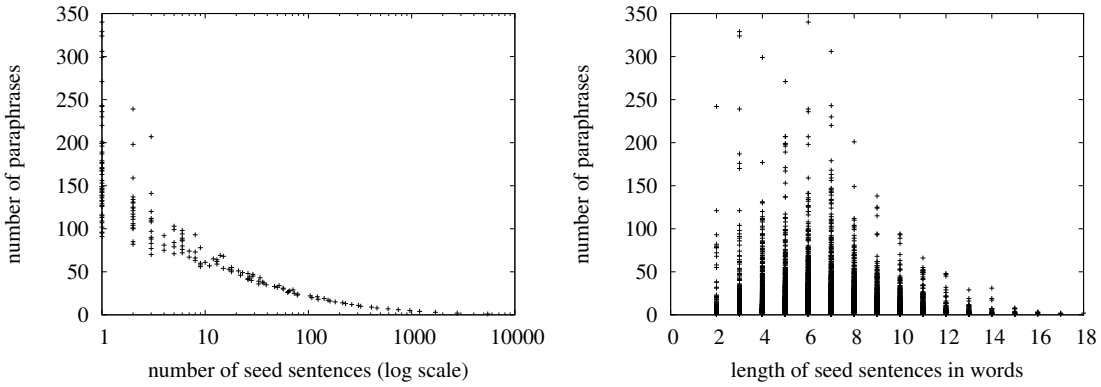

Figure 1: Number of seed sentences with the same number of paraphrases (on the left). Number of paraphrases by length of seed sentence in words (on the right).

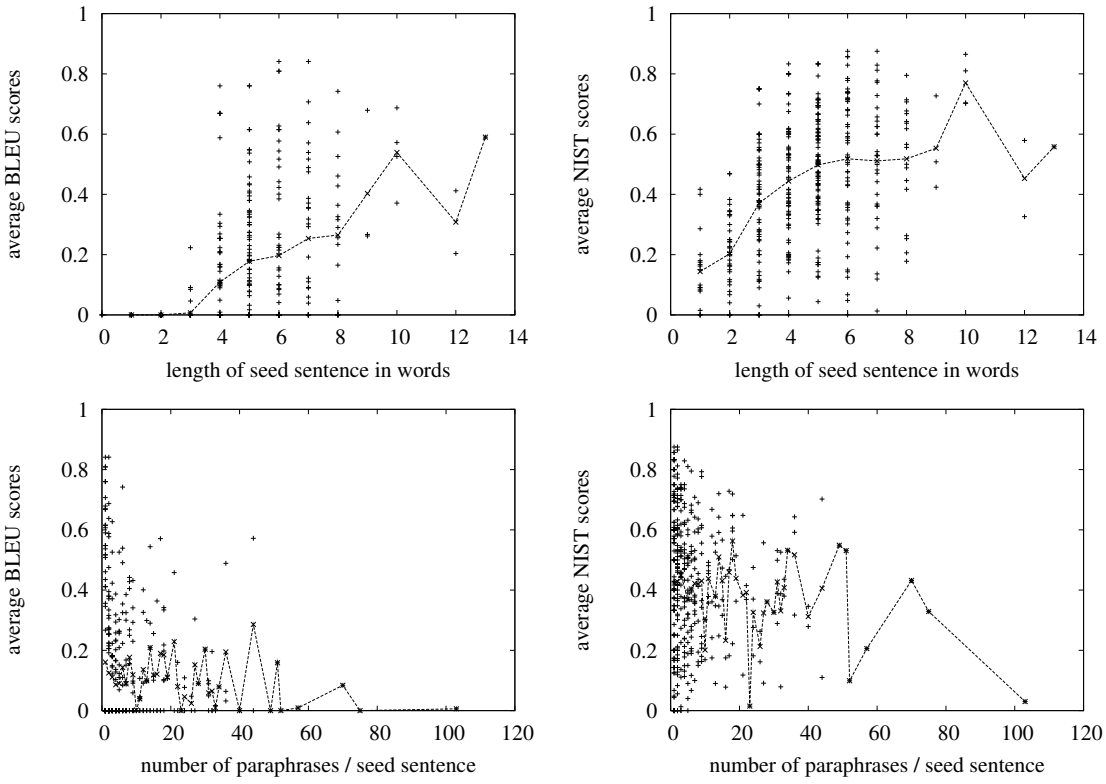

Figure 2: BLEU and NIST scores by length of seed sentence (upper graphs) and by number of paraphrases per seed sentence (lower graphs). In these graphs, each point is the score of a set of paraphrases against the seed sentence they were produced for. Lower scores indicate a greater lexical and syntactical variation in paraphrases. The connected points show mean values along the axis of abscissae.

Paraphrase		Not a paraphrase
Equivalence	Entailment	
346	104	20

Table 6: Equivalence or entailment in meaning of the paraphrases produced, on a sample of 470 paraphrases from various seed sentences.

7 Measure of lexical and syntactical variation in paraphrases

7.1 Objective measures

We assessed the lexical and syntactical variation of our paraphrases on a sample of 400 seed sentences using BLEU and NIST. On the contrary to evaluation of machine translation where the goal is to obtain high scores in BLEU and NIST, our goal here, when comparing a paraphrase to the seed sentence it has been produced for, is to get low scores. Indeed, high scores reflect some high correlation with translation references that is a lesser variation. As our goal is precisely to prepare data for evaluation with BLEU and NIST, it is thus to generate sets of paraphrases that would contain as much variation as possible to express the same meaning as the seed sentences, *i.e.* we look for low scores in BLEU and NIST.

Again, all this can be done safely as long as one is sure that the sentences compared are valid sentences and valid paraphrases. This is the case of our data, as we have already shown that the paraphrases produced are 99% grammatically and semantically correct sentences and that they are paraphrases of their corresponding seed sentences in 94% of the cases.

As for the meaning of BLEU and NIST, they are supposed to measure complementary characteristics of translations: namely fluency and adequacy (AKIBA et al., 2004, p. 7). BLEU tends to measure the quality in form of expression (fluency), while NIST[6] tends to measure quality in meaning (adequacy).

7.2 Results

The scores in BLEU and NIST (both on a scale from 0 to 1) shown in Figure 2 are interpreted as a measure of the lexical and syntactical variation among paraphrases. The lower they are, the greater the variation. The upper graphs show that this variation depends clearly on the lengths of the seed sentences. The shorter the seed sentence, the greater the variation among the paraphrases produced by this method. This is no surprise as the detection phase introduces a bias as was mentionned in Section 5 with the example sentence *Sure.*

The lower graphs show that the variation does not depend on the number of paraphrases per seed sentence. Hence, on the contrary to a method that would produce more variations as more paraphrases are generated, in our method, the variation is not expected to change when one produces more and more paraphrases (however, the grammatical quality or the paraphrasing quality could change). In this sense, the method is scalable, *i.e.*, one could tune the number of paraphrases wished without considerably altering the lexical and syntactical variation.

7.3 Comparison with reference sets produced by hand

We compared the lexical and syntactical variation of our paraphrases with paraphrases created by hand for a past MT evaluation campaign (AKIBA et al., 2004) in two language pairs: Japanese to English and Chinese to English.

For every reference set, we evaluated each sentence against one chosen at random and left out. The mean of all these evaluation scores gives an indication on the overall internal lexical and syntactical variation inside the reference sets. The lower the scores, the better the lexical and syntactical variation. This scheme was applied to both reference sets created by hand, and to the one automatically produced by our method. The scores obtained are shown on Figure 7. Whereas BLEU scores are comparable for all reference sets, which indicates no notable difference in flu-

[6]Formally, NIST is an open scale. Hence, scores cannot be directly compared for different seed sentences. We thus normalised them by the score of the seed sentence against itself. In this way, NIST scores become comparable for different seed sentences.

	Average BLEU	Average NIST
Automatically produced set	0.11	**0.39**
Hand-produced set 1	**0.10**	0.49
Hand-produced set 2	0.11	0.49

Table 7: Measure of the lexical and syntactical variation of various reference sets produced by hand and automatically produced by our method. The lower the scores, the better the lexical and syntactical variation.

ency, NIST scores are definitely better for the automatically produced reference set: this hints at a possibly richer lexical variation.

8 Conclusion

We reported a technique to generate paraphrases in the view of constituting reference sets for machine translation evaluation measures like BLEU and NIST. In an experiment with a linguistic resource of $97,769$ sentences we generated $8,65$ paraphrases in average for $16,153$ seed sentences.

The grammaticality was evaluated by sampling and was shown to be of at least 99% grammatically and semantically correct sentences (p-value = 2.22%), a quality comparable to that of the original linguistic resource. In addition, at least 96% of the candidates (p-value = 1.92%) were correct paraphrases either by meaning equivalence or entailment.

Finally, the lexical and syntactical variation within each paraphrase set was assessed using BLEU and NIST against the seed sentence. It was found that the lexical and syntactical variation did not depend upon the number of paraphrases generated, but on the length of the seed sentence.

Going back to the view of constituting reference sets for machine translation evaluation, not only are the paraphrase sets produced by this method correct sentences and valid paraphrases, but they also exhibit an internal lexical and syntactical variation which was shown to be slightly superior to that of two evaluation campaign sets of paraphrases produced by hand.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

References

Yasuhiro AKIBA, Marcello FEDERICO, Noriko KANDO, Hiromi NAKAIWA, Michael PAUL, and Jun'ichi TSUJII. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Bogdan BABYCH and Anthony HARTLEY. 2004. Extending the BLEU MT evaluation method with frequency weighting. In *Proceedings of ACL 2004*, pages 621–628, Barcelone, July.

Regina BARZILAY and Lillian LEE. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL 2003: Main Proceedings*, pages 16–23.

Ferdinand de SAUSSURE. 1995. *Cours de linguistique générale*. Payot, Lausanne et Paris. [1^e éd. 1916].

George DODDINGTON. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of Human Language Technology*, pages 128–132, San Diego, March.

Yves LEPAGE. 1998. Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume I, pages 728–735, Montréal, August.

Chin-Yew LIN and Eduard HOVY. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Edmonton, May.

Kiyonori OHTAKE and Kazuhide YAMAMOTO. 2003. Applicability analysis of corpus-derived paraphrases toward example-based paraphrasing. In *Language, Information and Computation, Proceedings of 17th Pacific Asia Conference*, pages 380–391.

Kishore PAPINENI, Salim ROUKOS, Todd WARD, and Wei-Jing ZHU. 2001. Bleu: a method for automatic evaluation of machine translation. Research report RC22176, IBM, September.