Adding paraphrases of the same quality to the C-STAR BTEC

Yves Lepage & Etienne Denoual
ATR 音声言語コミュニケーション研究所
619-0288「けいはんな学研都市」光台 2-2-2
{yves.lepage,etienne.denoual}@atr.jp

1 Introduction

We present a method to expand a linguistic resource with paraphrases, which combines two techniques whose drawbacks neutralise reciprocally. The first step over-generates sentences by using analogy, while the second step over-eliminates erroneous sentences which do not meet a criterion on N-gram occurrences. In a practical experiment, we added 17,862 paraphrases to a linguistic resource of 97,769 English sentences. The new paraphrases are 99% correct English sentences (p-value = 2.22%), a quality approximately equal to that of the original linguistic resource (p-value = 1.92%).

2 Justification

Similarly to natural resources, there exists a risk of "depletion" of linguistic resources used in successive evaluation campaigns of machine-translation systems: in such campaigns, parts of linguistic resources which are released cannot be used in subsequent campaigns, as they are not "new" anymore. This shows the need to generate "on-demand" new sentences to be added to linguistic resources in specific domains.

In addition, the use of automatic measures like NIST (Doddington, 2002) or BLEU (Papineni et al., 2001) requires paraphrases. This induces the need to annotate linguistic resources with multiple paraphrases. Such paraphrases are not easily gathered from, say, the Web. They should be synonymous sentences that explicit possible lexical or syntactical variations in order to cope with translation variations in terms and structures (Babych and Hartley, 2004a).

Relatively to the aforementioned topics, we propose a technique to expand a linguistic resource. It addresses the following two concerns. Firstly, lexical or syntactical variations are dealt with using a linguistic operation capturing commutations: analogy (de Saussure, 1995, part 3, chap 4). Secondly, it follows the trend of using N-grams to reflect naturalness and partly judge adequacy (Babych and Hartley, 2004b), as it uses a filtering technique based on the absence of unseen N-grams (Brill and Soricut, 2004) (Lin and Hovy, 2003).

3 The linguistic resource used

For this study, we used 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. It contains 162,318 aligned translations in several languages, and especially in English and Japanese. The sentences are quite short as the figures in Table 1 show. As for English, there are $97,769$ different sentences (some sentences may appear several times) with an average length of 35.14 characters and a standard deviation of 18.81.

The quality of this resource is of 99% (p-value = 1.92%) correct sentences. Errors include spelling and syntactical mistakes as is exemplified in Table 2.

4 Over-generation by analogy

4.1 Method

The method relies firstly on the equality of translation for different sentences and secondly on linguistic commutations at work in analogies. This is detailed in the sequel.

Firstly, it is reasonable to say that the equality in translation of different sentences implies a paraphrase equivalence. For instance, the three following English sentences *A beer, please.*, *Can I have a beer?* and *Give me a beer, please.*, share a common Japanese translation in our linguistic

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

Table 1: Some statistics about the linguistic resource

	$\sharp$ of $\neq$ sentences	size in characters		
		avg.	±	std. dev.
English	97,769	35.14	±	18.81
Japanese	103,274	16.21	±	7.84

Table 2: Some incorrect sentences in the linguistic resource with their description.

* *What famous store do you recommend?.*	(superfluous fullstop)
* *Yes, This one.*	(uppercase after comma)
* *Please fill out this registration from.*	(*from* instead of *form*)
* *Good-by.*	(instead of *Good-bye*)
* *Any massages for me?*	(*massages* instead of *messages*)
* *I couldn't here the announcement.*	(*here* instead of *hear*)
* *I'm locked myself out.*	(incorrect syntax)

resource (ビールをください。). Therefore, they are paraphrases. In our linguistic resource, one Japanese sentence corresponds to 1.57 English sentence in average.

Secondly, a given sentence may share commutations with other sentences of the corpus. Such commutations are best seen in analogical relations. For instance, the sentence *A slice of pizza, please.* enter in the analogies of Table 3. By replacing in such analogies some sentences with known paraphrases, it is possible to produce new sentences that do not already exist in the linguistic resource. For instance, by replacing *A beer, please.* with *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 new paraphrase of *A slice of pizza, please.*

4.2 Results

With our linguistic resource, the application of the method generated $4,495,266$ English sentences. An inspection of a sample of 400 sentences shows that the quality lies around 23.6% of correct sentences (p-value = 1.19%).

As a matter of fact, analogy has a well-known drawback: it overgenerates. For instance, with the same analogy as previously, the replacement of *A beer, please.* by *A bottle of beer, please.* will produce the unfortunate following 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 strings which go against the linguistic feeling. Two examples of such unacceptable strings are as follows: **A slice to get of pizza, please.* and: **A slice of pizzthe, pleaset for tha, please.*

In order to retain paraphrases that are valid at the same quality level of that of the original linguistic resource, the second step of the method has to spot and eliminate such sentences.

5 Over-elimination by unseen N-grams

5.1 Method

The task is thus to retain only those sentences which are undoubtedly correct. As the number of generated sentences is high, it is well permissible to eliminate sentences in case of doubt. Therefore, while the method need have a precision as close as possible to 100%, a poor recall is well acceptable. To this end, we eliminate any sentence containing any N-gram unseen in

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

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 4: Some of the 678 strings output as paraphrase candidates for *A slice of pizza, please.*

ieasiCe of pizza, please.
A cup of a slice of pizza, please.
A non-smoking table, I’d like a slice.
A room slice.
Bring me a slice of pizza, please.
Bring me a slice.
I’d like a a slice bed of pizza, please.
I’d like a a slice bed.
I’d like to have a a slice seat.
I’d like to have a a slice, please.
Some a slicee, plase.
Some slice of pizza, please.
...

the original data. Results obtained on two sentences with different values of N are shown in Tables 5 and 6. The next section shows that acceptable results are obtained which leave space for improvement by more subtle N-gram scoring methods.

5.2 Results

In order to obtain a quality rate for the paraphrases at approximately the same level to that of the original linguistic resource, the best value for N was 20, a value to be compared with 35.14, the average length of sentences in the original data. The number of sentences retained after filtering was $51,911$, of which $34,049$ were sentences already contained in the original resource. Adding the newly $17,862$ generated sentences increases the linguistic resource by one fifth. The quality of the filtered paraphrases was evaluated by sampling: with a p-value of 2.22%, 99% of the paraphrases may be considered correct. This quality is approximately the same as that of the original resource (99% with a p-value of 1.92%). An overview of the errors in the generated paraphrases (e.g. *Where is tourist area?*) suggests that they do not differ from the ones in the original data (*Where is information office?* is found in the original data, see 3).

6 Conclusion

In this paper, we reported a technique to increase the size of a linguistic resource with paraphrases. The technique works in two steps. The first step over-generates whilst the second one over-eliminates candidate strings. In an experiment with $97,769$ English sentences we added $17,862$ paraphrases, which increases the original resource by 18.32%. The quality is left untouched at 99% (p-value = 2.22%).

The originality of the technique is that all work is done at the level on characters. Hence, the technique is applicable to languages without word segmentation, like Japanese.

Our results still leave room for improvement. Firstly, the analogical technique used here consists in a one-shot application: recursive application should deliver a much greater number of sentences. Secondly, N-gram filtering may be improved by using values of N depending on the length of the sentences at hand, and by using the probabilities of the occurrences of N-grams to perform standard N-gram scoring of the generated sentences.

Table 5: Paraphrases for *A slice of pizza, please.* after N-gram filtering ($N = 15$). These sentences did not exist in the linguistic resource. The number on the left is the number of times the paraphrase was produced.

51	*Can I have a slice of pizza, please?*
45	*Could I have a slice of pizza, please?*
12	*A slice of pizza, please?*
12	*I'll have a slice of pizza, please.*
9	*Can I have a slice of pizza, please.*
9	*I'd like a slice of pizza, please?*
3	*slice of pizza, please.*

Table 6: Paraphrases for *Can we have a table in the corner?* after N-gram filtering ($N = 20$). These sentences did not exist in the linguistic resource. The number on the left is the number of times the paraphrase was produced.

1678	*Could we have a table in the corner?*
1658	*We would like a table in the corner.*
1652	*I'd like a table in the corner.*
878	*Can we have a table in the corner?*
50	*Can I get a table in the corner?*
8	*We'd like a table in the corner.*
2	*I prefer a table in the corner.*

Acknowledgements

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

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

Bogdan BABYCH and Anthony HARTLEY. 2004b. Modelling legitimate translation variation for automatic evaluation of MT quality. In *Proceedings of LREC 2004*, volume III, pages 833–836, Lisbonne, May.

Eric BRILL and Radu SORICUT. 2004. A unified framework for automatic evaluation using n-gram co-occurence statistics. In *Proceedings of ACL 2004*, pages 613–620, Barcelone, July.

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.

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.