Using Paradigm Tables to Generate New Utterances Similar to those Existing in Linguistic Resources

Yves Lepage and Guilhem Peralta

ATR- Spoken Language Translation Research Laboratories
619-0288, Keihanna Science City, Japan
yves.lepage@atr.jp

Abstract

We inspect the possibility of creating new linguistic utterances (small sentences) similar to those already present in an existing linguistic resource. Using paradigm tables ensures that the new generated sentences resemble previous data, while being of course different. We report an experiment in which 1,201 new correct sentences were generated starting from only 22 seed sentences.

1. Introduction

In this article, we are concerned with the addition of new sentences that resemble those contained in an already existing linguistic resource.

For a definite task, just collecting texts, for instance from the Web, does not suffice as the data required are always very dependent on the task at hand. Collecting a large amount of representative data is time consuming and monetarily expensive.

The previous reasons lead to the idea of automatically expanding an existing corpus. Starting from a corpus already tuned for the task, one wants to automatically produce more and more utterances or sentences in the task domain that resemble in a certain way the sentences of the initial corpus.

2. Linguistically Justified Production of New Sentences from Observed Ones

2.1. Commutation Series, Paradigms

Our method relies on the notion of *paradigms*. A paradigm is built on a number of series of commutations among sentences. For instance, the sentences on the left below show a series of commutations of *Japanese* with *Spanish*, *French*, etc. These commutations do not just exchange nationality-related adjectives; they cross the boundaries between (derivational) morphology and syntax as the last examples clearly show: *seafood*, *almost all kinds of food*. Another more complex example of commutations, is shown on the right below. Commutations happen both at the front and at the end of sentences with a certain degree of freedom.

I like Japanese food.	*Japanese food would be fine.*
I like Spanish food.	*I'd prefer Japanese food.*
I like French food.	*Japanese food is fine with me.*
I like seafood.	*I'd like to have Japanese food.*
I like almost all kinds of food.	*Does Japanese food suit your*
...	... [*taste?*

Paradigms are revealed by the existence of several series of commutations around a given sentence of the corpus, the *seed sentence*. Figure 1 shows this in the form of a table with several columns starting from the seed sentence *I like Japanese food.*, placed at the top left corner of the table.

In fact, paradigms may involve many different commutation series. Thus, the exact representation of a complete paradigm should be a multi-dimensional space. For reasons of visibility, and because this is always possible, we use a projection on two dimensions: both the border line and the border column contain the same sentences in the same order, so that the table is symmetrical. We call such tables "complete tables".

As is examplified in Fig. 1, paradigm tables are usually rather hollow. For instance, in Fig. 1, the total number of cells in the visible part of the table is $14 \times 8 = 112$, with only 10 inner cells filled with sentences from the corpus. Our goal will be to fill in the other cells of the table.

2.2. Analogy

Sentences in the inner cells of paradigm tables meet a linguistic relationship with the top left sentence and the corresponding cells on the border. This relationship is an *analogy*, usually noted $A : B :: C : D$, which states that A *is to* B *as* C *is to* D[1]. It may be characterised on different levels of abstraction and between different types of objects (HOLYOAK and THAGARD, 1995). We shall only be concerned with the formal type of analogy described in linguistics (PAUL, 1920). Sentences generated in this way, although not necessarily correct (de SAUSSURE, 1995), are much more linguistically constrained (ITKONEN and HAUKIOJA, 1997) than simple strings of characters produced by, say, n-gram models used in generation. We use a purely formal characterisation of analogy between strings of symbols which is based on the verification of a similarity criterion:

$$A : B :: C : D \quad \Rightarrow \quad \begin{cases} \mathrm{d}(A,B) & = & \mathrm{d}(C,D) \\ \mathrm{d}(A,C) & = & \mathrm{d}(B,D) \end{cases}$$

and of a contiguity criterion:

$$A : B :: C : D \quad \Rightarrow \quad \forall a, |A|_a + |D|_a = |B|_a + |C|_a$$

$\mathrm{d}(A,B)$ stands for the edit distance between A and B with deletions and insertions as the sole edit operations. $|A|_a$ is the number of occurrences of the symbol a in the string A.

By contrapositive implication, these criteria may be used to test whether an analogical equation has a solution. For instance, the contiguity criterion implies that there is no solution to the equation below[2].

I like Japanese food. : *I prefer Japanese food.* :: *I enjoyed the food.* : x

With this, one can already mark cells in paradigm tables where no sentence can be created. In Fig. 1, such cells have been marked with a •.

[1] For example, the sentence *I prefer Italian food.* is to *I prefer Japanese food.* as the sentence *I like Italian food.* is to *I like Japanese food.*

[2] If n is the number of *l*s in the solution of the analogical equation, then according to the contiguity criterion, n should meet the following constraint: $1 + n = 0 + 0$. This is impossible.

I like Japanese food.	*I prefer Japanese food.*	*I'd prefer Japanese food.*	*I feel like Japanese food.*	*I enjoyed the food.*	*I prefer French food.*	*I like Italian food.*	*I like seafood.*	*I'd like local food.*	...
I prefer Japanese food.	•	•		•	•	*I prefer Italian food.*	*I prefer seafood.*		...
I'd prefer Japanese food.	•	•		•	•				...
I feel like Japanese food.							*I feel like seafood.*		...
I enjoyed the food.	•	•		•	•	•	•	•	...
I prefer French food.	•	•		•	•	•	•	•	...
I like Italian food.	*I prefer Italian food.*			•	•	•	•	•	...
I like seafood.	*I prefer seafood.*		*I feel like seafood.*	•	•	•	•	•	...
I'd like local food.				•	•	•	•	•	...
I like Mexican food.			*I feel like Mexican food.*	•	•	•	•	•	...
I'd like the local food.				•	•	•	•	•	...
I like Western food.		*I'd prefer Western food.*		•	•	•	•	•	...
I'd like some Italian food.				•	•	•	•	•	...
I'd like Western food.	*I'd prefer Western food.*			•	•	•	•	•	...
I like Chinese food.	*I prefer Chinese food.*			•	•	•	•	•	...
...	...	...	...	...	...	...	...	...	...

Figure 1: A chunk of the paradigm table for the seed sentence *I like Japanese food.* The table is symmetrical relative to the diagonal. Cells marked with • are explained in 2.2..

Our goal is to fill in the blank cells of paradigm tables with new sentences. Consequently, we also need a procedure to enumerate sentences that meet the criteria previously mentioned. *i.e.* to solve analogical equations. Ours is based on the computation of edit distances between strings of symbols (LEPAGE, 1998), and outputs the first encountered sentence that meets both criteria. For instance, the following solution is output for the following analogical equation.

$$\textit{I like Japanese food.} : \textit{I feel like Japanese food.} :: \textit{I like seafood.} : x \quad \Rightarrow \quad x = \textit{I feel like seafood.}$$

3. Generating New Sentences in Paradigm Tables

Any inner cell in a complete paradigm table stands for an analogical equation. When the cell is filled, this means that the analogical equation has a solution which is observed in the corpus. When the cell does not contain anything, we may want to solve the analogical equation. There are two possible issues: either no solution exists to the analogical equation, and the cell remains empty; or there exists a solution, a sentence, which can be inserted into the table. By definition, since such a sentence is not observed in the corpus, it is new.

To evaluate the efficiency of our method, we picked up 22 seed sentences at random from the Basic Traveller's Expressions Corpus (`http://www.c-star.org`). This is a collection of sentences representative of various travel situations, like hotels, restaurants, post offices, trains, etc. There are 97,769 unique sentences, with an average of 5.85 words per sentence. Because this corpus is quite large, for each of our 22 seed sentences, we kept only several thousand sentences as a sub-corpus by filtering with a typical keyword for the seed sentence.

We computed all complete paradigm tables and solved all possible analogical equations. All sentences generated by analogy were checked by hand for grammaticality. Table 1 summarizes the results in the columns entitled "complete table": the number of attested sentences in the paradigm table is compared with the total number of cells in the table, this latter number represents the total number of theoretically possible sentences in the paradigm. Also, the number of new correct sentences is compared with the total number of new generated sentences.

4. Correctness of the New Generated Sentences

As column 4 of Table 1 shows, the number of correct sentences generated with 22 complete paradigm tables is quite high although we limited ourselves to sub-corpora: 1,427 new correct sentences were obtained. Column 2 shows that the percentage of correct sentences is also high: almost 2/3 (62 %).

We encountered problems to classify the generated sentences into incorrect and correct ones. Hereafter a star indicates a sentence that we finally rejected:

Let me think about your passport for a while, please.
**Okay. Let me weigh your passport.*
**Where's the passport control office and immigration card?*

Other examples are simpler. The following sentences reflect the fact that our analogical equation solver works on a character unit level.

**I'dt's Western food.*
**What's e differencyour in price between them?*
**Whate rgular price do you have in mind?*
**Is this the be differencest in price between them?*

A spellchecker of English could help in rejecting such incorrect sentences, but simply filtering by the vocabulary of the corpus is not satisfying, as an advantage of analogy is precisely its ability of generating new words by morphological formation. Also, a trigram model of English words may well be able to reject sentences like **May you show me your passport.*, but it may well reject *I like classical food.*[3] because of *classical food*, although some people may find it correct.

Our purpose here is to inspect ways of improving the production of new sentences in paradigm tables directly so as to increase the reliability (in percentage) of the generated sentences.

5. Densifying Paradigm Tables to Make New Generated Sentences More Reliable

A first observation of the complete paradigm tables is that a line with a lot of cells that cannot be filled (marked with an • in Fig. 1) is to be interpreted as a line which does not satisfyingly commute with the other sentences on the borders. In other words, the membership of such a sentence to the overall paradigm is weak, and it may be questioned. As a consequence, we should look for solutions to decrease the number of holes which cannot be filled in paradigm tables.

A second observation is that a cell which could be filled with a new sentence is filled with more reliability when it is surrounded by a greater number of sentences observed in the corpus. As a consequence, we should try to increase the number of attested analogical sentences in the paradigm table.

These two tasks of minimising the number of theoretically "unfillable" cells and maximising the number of cells filled with sentences from the corpus can be described in similar terms: the first goal consists of increasing the *paradigmatic density*, the second one consists of increasing the *observed paradigmatic density*.

[3]Produced by the analogy: *I like French music.* : *I like classical music.* :: *I prefer French food.* : x.

6. Comparison of Complete and Densified Paradigm Tables

We performed similar countings as with the complete paradigm tables, for new densified tables obtained automatically by program on the same seed sentences. The results are shown in Table 1. On the left part of this table, the increase in the ratio of attested sentences over all theoretically possible sentences is to be interpreted as a measure of the increase of the paradigmatic density. This density was more than doubled, from $7\ \%$ to $16\ \%$. On the right part of Table 1, the ratio between the number of new correct sentences and the total number of new generated sentences stands for the quality of the analogical generation. It increased from 62 % to 70 %. Clearly, the densification of the paradigm tables using our method increased the reliability of the new generated sentences.

Going back to our initial goal, which is the automatic production of new sentences to be added to a linguistic resource, during the above-reported experiments, we were able to generate 1,427 new correct sentences from only 22 seed sentences (and restricting ourselves to subcorpora) using complete paradigm tables. By densifying the paradigm tables, only $226/1,427 = 16\ \%$ of the new correct sentences were left aside. In other words, the proportion of new correct sentences retained in the densified tables is $84\ \%$, which makes 1,201 new correct sentences. Combined with the increase in reliability, this shows that densification privileges quality over quantity.

7. Conclusion

We have inspected the possibility of increasing the size of a corpus by using paradigm tables. The sentences generated with this method still have to be checked by hand for morphological, syntactical, semantic and pragmatical correctness. Standard techniques of n-grams, spellcheckers, or syntax checkers could and should be used to filter out incorrect sentences. We inspected the possibility of making the proposed method intrinsically more reliable. For that, we proposed to "densify" paradigm tables obtained from a seed sentence by first reducing the number of unsolvable analogies and then augmenting the relative number of observed analogical sentences. This densification increased the paradigmatic density simultaneously with the reliability of the sentences generated by analogy.

8. Acknowledgements

This research was supported in part by the Telecommunications Advancement Organization of Japan.

9. References

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

HOLYOAK, K.J. and P. THAGARD, 1995. *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

ITKONEN, Esa and Jussi HAUKIOJA, 1997. *A rehabilitation of analogy in syntax (and elsewhere)*. pages 131–177.

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

PAUL, Hermann, 1920. *Prinzipien der Sprachgeschichte*. Tübingen: Niemayer. 5^e éd., [1^e éd. 1880].

I like Japanese food.	*I prefer Japanese food.*	*I'd prefer Japanese food.*	*I feel like Japanese food.*
I enjoyed the food.	•	•	
I prefer French food.	•	•	
I like Italian food.	*I prefer Italian food.*		
I like seafood.	*I prefer seafood.*		*I feel like seafood.*
I'd like local food.			
I like Mexican food.			*I feel like Mexican food.*
I'd like the local food.			
I like Western food.		*I'd prefer Western food.*	
I'd like some Italian food.			
I'd like Western food.	*I'd prefer Western food.*		
I like Chinese food.	*I prefer Chinese food.*		
I like Spanish food.			
I'd like some famous local food.			
Do you like Italian food?			*Do you feel like Italian food?*

Figure 2: The densified paradigm table for the seed sentence *I like Japanese food.*

I like Japanese food.	*I prefer Japanese food.*	*I'd prefer Japanese food.*	*I feel like Japanese food.*
I enjoyed the food.	•	•	△*I feel enjoyed the food.*
I prefer French food.	•	•	△*I feel prefer French food.*
I like Italian food.	*I prefer Italian food.*	△*I'd prefer Italian food.*	△*I feel like Italian food.*
I like seafood.	*I prefer seafood.*	△*I'd prefer seafood.*	*I feel like seafood.*
I'd like local food.	△*I'd prefer local food.*	△*I'd'd prefer local food.*	△*I feel'd like local food.*
I like Mexican food.	△*I prefer Mexican food.*	△*I'd prefer Mexican food.*	*I feel like Mexican food.*
I'd like the local food.	△*I'd prefer the local food.*	△*I'd'd prefer the local food.*	△*I feel'd like the local food.*
I like Western food.	△*I prefer Western food.*	*I'd prefer Western food.*	△*I feel like Western food.*
I'd like some Italian food.	△*I'd prefer some Italian food.*	△*I'd'd prefer some Italian food.*	△*I'd feel like some Italian food.*
I'd like Western food.	*I'd prefer Western food.*	△*I'd'd prefer Western food.*	△*I feel'd like Western food.*
I like Chinese food.	*I prefer Chinese food.*	△*I'd prefer Chinese food.*	△*I feel like Chinese food.*
I like Spanish food.	△*I prefer Spanish food.*	△*I'd prefer Spanish food.*	△*I feel like Spanish food.*
I'd like some famous local food.	△*I'd prefer some famous local food.*	△*I'd'd prefer some famous local food.*	△*I'd feel like some famous local food.*
Do you like Italian food?	△*Do you prefer Italian food?*	△*Do you'd prefer Italian food?*	*Do you feel like Italian food?*

Figure 3: Filling a paradigm table. Cells with an • have no analogical solution. Cells with a △ were produced by analogy.

	complete table	densified table	complete table	densified table
seed sentence label	observed / possible sentences	observed / possible sentences	correct / all new sentences	correct / all new sentences
20dollars	62 / 1770 = 3 %	62 / 731 = 8 %	371 / 421 = 88 %	365 / 403 = 91 %
CatchTaxi	16 / 171 = 9 %	16 / 70 = 22 %	23 / 28 = 82 %	18 / 21 = 86 %
FeelBlue	40 / 630 = 6 %	33 / 224 = 14 %	51 / 86 = 59 %	40 / 63 = 64 %
GetPostOffice	19 / 136 = 13 %	19 / 42 = 45 %	14 / 21 = 67 %	14 / 21 = 67 %
HardTime	8 / 66 = 12 %	6 / 35 = 17 %	22 / 29 = 76 %	13 / 15 = 87 %
JapFood	50 / 820 = 6 %	43 / 418 = 10 %	178 / 468 = 38 %	137 / 290 = 47 %
LeftTrain	15 / 105 = 14 %	14 / 36 = 38 %	13 / 16 = 81 %	11 / 14 = 79 %
LetmeSee	32 / 595 = 5 %	22 / 304 = 7 %	26 / 50 = 52 %	19 / 22 = 86 %
OnBusiness	12 / 66 = 18 %	10 / 36 = 27 %	14 / 39 = 36 %	9 / 26 = 35 %
OutOffice	30 / 435 = 6 %	27 / 81 = 33 %	26 / 83 = 31 %	19 / 46 = 41 %
PleaseTaxi	57 / 946 = 6 %	53 / 475 = 11 %	111 / 157 = 71 %	83 / 120 = 69 %
PreferSeafood	27 / 153 = 17 %	27 / 77 = 35 %	26 / 47 = 55 %	26 / 42 = 62 %
SeePass	20 / 171 = 11 %	19 / 78 = 24 %	40 / 49 = 82 %	36 / 41 = 88 %
TakeTrain	20 / 231 = 8 %	20 / 57 = 35 %	39 / 39 = 100 %	37 / 37 = 100 %
TrainTime	21 / 276 = 7 %	19 / 44 = 43 %	11 / 25 = 44 %	10 / 24 = 42 %
WantBookRoom	8 / 45 = 17 %	8 / 25 = 32 %	6 / 6 = 100 %	6 / 6 = 100 %
WantCoffee	39 / 351 = 11 %	37 / 126 = 29 %	30 / 53 = 57 %	25 / 46 = 54 %
WhatPrice	26 / 378 = 6 %	24 / 195 = 12 %	157 / 272 = 58 %	107 / 170 = 63 %
WhereOffice	18 / 231 = 7 %	14 / 121 = 11 %	77 / 136 = 57 %	52 / 82 = 64 %
WorkTrading	27 / 435 = 6 %	25 / 104 = 24 %	85 / 93 = 91 %	69 / 73 = 95 %
YenDollars	30 / 351 = 8 %	29 / 182 = 15 %	64 / 149 = 43 %	62 / 104 = 60 %
YouFeelTired	30 / 253 = 11 %	30 / 76 = 39 %	43 / 48 = 90 %	43 / 46 = 94 %
average	607 / 8,615 = 7 %	557 / 3,537 = 16 %	1,427 / 2,315 = 62 %	1,201 / 1,712 = 70 %

Table 1: Results for 22 seed sentences.