THE SNOW-BALL EFFECT OF ANALOGY

Yves LEPAGE, ANDO Shin-ichi, IIDA Hitoshi
ATR Interpreting Telecommunications Research Labs,
Hikaridai 2-2, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
e-mail: {lepage, ando, iida}@itl.atr.co.jp

Abstract

In previous work, a computational explanation of analogy, a phenomenon by which a new sentence is generated from three other sentences, has been proposed. For a given set of sentences, the number of analogy relations in which a new sentence may be involved is thus *a priori* proportional to the cube of the size of the set. In this paper, we study this number while reading a corpus, sentence by sentence. We also study the smallest number of previously read sentences which would ensure just a certain number of analogies. We also show how to construct a significantly reduced set of sentences from which any sentence of the corpus can be generated by analogy.

Introduction

Ananlogy, a phenomenon of major importance in language has not received much attention from the NLP community because of its lack of a computational explanation. As such an explanation has been recently proposed, the study of this phenomenon on a large corpus is made possible.

1 Analogy

1.1 Linguistic phenomenon

Analogy is a synchronic phenomenon, described by the Neugrammatiker and Saussure [Saussure 16], which explains the creation of regular, understandable but unregistered words. Examples are frequent in child language or in risky word formations[1]:

$$\textit{therm} : \textit{thermodynamics} = \textit{think} : \mathbf{x}$$
$$\mathbf{x} = \textit{thinkodynamics}$$

Analogy may be considered *the* general operation at work behind morphology because it captures the regularity of conjugation or derivation (and does not, of course, explain exceptions nor differences in paradigms!) In all these morphological operations, analogy involves string operations like suffixing, prefixing and even infixing, as shown in the following example.

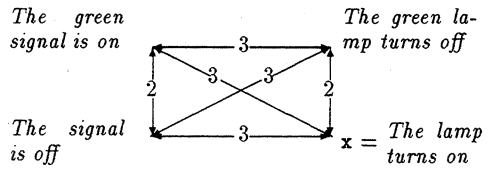
$$\textit{fable} : \textit{fabulous} = \textit{miracle} : \mathbf{x}$$
$$\mathbf{x} = \textit{miraculous}$$

Hermann Paul himself, one of the Neugrammatiker, and Bloomfield too, had the idea that, maybe, analogy was also at work in syntax, *i.e.*, it was also a means of creating new sentences with the help of (at least three) already existing sentences, as illustrated below.

$$\begin{array}{l}\textit{The green signal is on} : \textit{The signal is off} = \textit{The green lamp turns off} : \mathbf{x}\\ \mathbf{x} = \textit{The lamp turns on}\end{array}$$

1.2 Computational explanation

The four terms intervening in the analogy relation may be placed at the corners of a rectangle. By definition of a rectangle, the sizes of opposite sides and diagonals, representing the distances between the sentences, are equal respectively.



According to this interpretation, analogy is equivalent to the following system of three equations [Lepage 96].

Definition 1 (Analogy)

$$u : v = w : \mathbf{x} \overset{\Delta}{\Longleftrightarrow} \begin{cases} \mathrm{dist}(u,v) = \mathrm{dist}(w,\mathbf{x}) \\ \mathrm{dist}(u,w) = \mathrm{dist}(v,\mathbf{x}) \\ \mathrm{dist}(v,w) = \mathrm{dist}(u,\mathbf{x}) \end{cases}$$

Distances may be the classical edit distances between strings as defined in [Levenshtein 65] or [Wagner & Fischer 74] from simple edit operations: insertions, deletions and substitutions.

- deletion (*a red lamp* → *a lamp*)
- insertion (*you see* → *you should see*)

[1] Found in [Hofstadter et al. 94]!

- substitution (*a big cat* → *a small cat*)

Edit distances give the minimum number of edit operations which have to be performed to "transform" one sentence into another one.

1.3 The snow–ball effect

By definition, analogy involves four sentences. Given a set of sentences of size n, if we wanted to compute the number of analogies in which a new sentence intervenes, the number of times the analogy relation has to be checked is $A_n^3 = n \times (n-1) \times (n-2)$, because the order of the sentences does not matter in analogy[2]. This is almost a cubic explosion.

Among the total number of combinations, only a certain number will verify the analogy relation. We may expect this number to grow extremely fast too. It is the case and it is what we call the *snow–ball effect* of analogy.

We shall study this snow–ball effect:

- firstly, *for itself*: What is the shape of the curve representing the average number of analogies while reading a corpus? Does it grow for ever, or does it stop after a certain rank?
- secondly, so as to *limit* it: Is it possible to remember only a fixed number of past sentences in order to ensure a certain number of analogies verified by the next sentence to be read?
- thirdly, for *representativity* purposes: Is it possible to build a smaller corpus representative of the entire corpus?

1.4 The data

We conducted experiments on a collection of texts from the ATR-Lancaster tree–bank [Black et al. 96]. These texts come from various sources on the Internet, ranging from homepages to economical or medical reports, geographical descriptions, posts to Internet newsgroups, science-fiction novels, *etc.* Altogether, our excerpt contains 5 000 sentences (as a comparison, this article contains less that 100 sentences). The order of the sentences is left untouched in all our experiments.

These texts come in two forms:

- the *real texts*. The number of different words is about 10 600;
- the *tagged* form: each word of the text is replaced by a tag, which represents grammatical and semantic information. The set of tags is that of the ATR-Lancaster team. 960 different tags occurr in our texts.

For each sentence (under its two forms), we first computed the number of analogies involving only sentences of lesser rank. We call this the *number of past analogies*.

2 Number of past analogies

2.1 Experimental results

In Figure 1, for each sentence, the number of past analogies is plotted against the sentence number until rank 2 300. These points have the function $y = (x/9)^3$ as an envelop.

[2] $A_n^m = n!/(n-m)!$ is the number of sorted lists of m objects taken from a set of n.

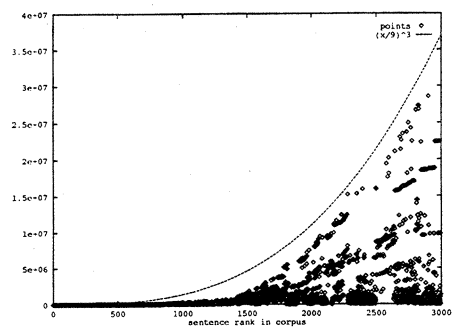


Figure 1: Number of past analogies.

2.1.1 Approximation

The previous points can be smoothed into a curve by taking the average on the past 100 values. This is shown in Figure 2 for tagged texts only, as the curve for raw texts is almost superimposed to this one.

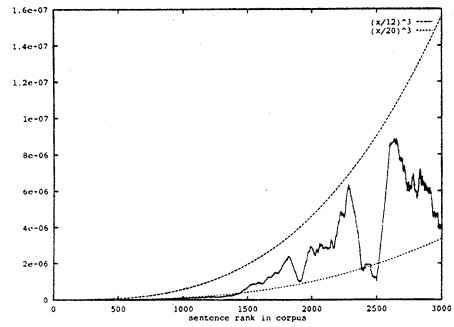


Figure 2: Number of past analogies (smoothed).

The curve obtained may be roughly framed (after rank 1 500) between two cubic curves:

$$y = (x/12)^3 \text{ and } y = (x/20)^3$$

which seems natural, as four objects intervene in analogy and we fixed one of them.

We still do not know if the number of past analogies stops after a certain rank, which would reflect some "saturation" of analogy, or not. As this kind of experiments is very much time–consuming, we are not in the position of giving an answer to this question at the moment.

3 Short-term memory

A first reasonable goal is to determine, locally, a fixed number of past sentences which would ensure a certain number of analogies for any new sentence, while reading the corpus. For a given sentence, we determine the smallest size of memory, in number of sentences which have to be remembered, in order to get a fixed number of analogies on average. In other words, we compute the smallest $m(r,n)$ such that sentence $r+1$ can be obtained

n times by analogy on triples of sentences whose ranks are in $[r - m(r,n); r]$.

This result is of practical importance, as, in any application, the size of the memory should be minimal so that computations should be as fast as possible.

3.1 Global shape

Figure 3 shows the (smoothed) results obtained for real texts and tagged sentences so as to obtain 1, 10 or 100 analogies.

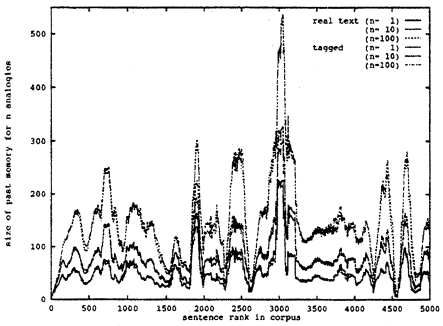


Figure 3: Size of necessary memory to get a fixed number of analogies.

A striking feature is the very similar shape of all these curves. The curves for real texts and tagged sentences are almost superimposed in both cases (number of analogies: 1 or 10). This means that real words or tags make small difference. We conclude that the set of tags (about 960 different tags) is a very good representation of words (about 10 600 different words) as far as analogy is concerned.

3.2 Absolute height and mean memory

The size of the memory reflects the coherence of the text: the smaller this size, the more analogical the sentences in a near context; the higher this size, the more diverse the sentences.

On our data, a context of about 150 past sentences (5 to 6 pages) on average is sufficient to obtain 100 analogies for a new sentence. Fewer, *i.e.* 84 (or 51) sentences on average, are sufficient to get 10 analogies (or only 1 analogy) if desired.

3.3 Local minima

Minima are observed at some ranks. If we think that analogy represents some similarity between sentences, then, it is reasonable to think that sentences sharing the same style in a text would have the same memory. Hence, we should observe a sudden growth for a sudden change in styles of texts, and a fall as soon as the style remains the same. In other words, minima should correspond to some of the text boundaries (two consecutive texts may have the same style).

On the entire corpus, we verified that minima correspond well to text boundaries, This is illustrated in Figure 4, on a portion of the corpus, where the boundaries visualised by vertical lines are the nearest ones to

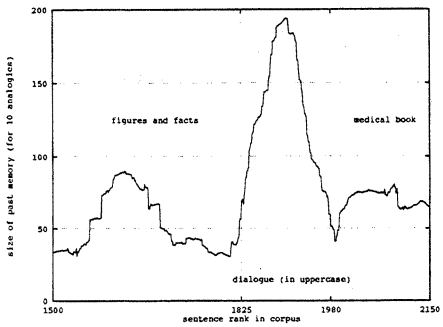


Figure 4: Style discrimination based on analogical memory.

minima. Rank 1 500 to 1 739 is a list of facts like superficy, population, *etc.* about African countries, a list of the meaning of the international maritime flags, *etc.*; from rank 1 826 to 1 980 is a dialogue written in uppercase (our program is case-sensitive); the third part is an excerpt of a medical book on cancer.

4 Reduced representative corpus

A second goal is to reduce, globally, a text to a certain number of sentences so that any sentence discarded can be reconstructed by analogy with three sentences kept in the reduced text. This is very much related with text representativity.

4.1 Definition of a reduced corpus

For that, we build a reduced set of sentences, which we call a *reduced corpus*, and which verifies the following two properties:

- for any quadruples of sentences in the reduced corpus, the analogy relation does not hold;
- any sentence of the entire corpus can be obtained by analogy from three sentences from the reduced corpus.

4.2 Construction of a reduced corpus

Because of the first property of the reduced corpus, it can be simply constructed by application of a greedy algorithm:

- initialise the reduced corpus with the first three sentences of the entire corpus.
- for each new sentence, if analogy is not verified with any triple of sentences from the reduced corpus, add this sentence to the reduced corpus, else do nothing.

As a remark, we have to say that the reduced set obtained is not necessarily the smallest in size of all possible reduced sets. At each step, we do not check if a different set of sentences, also built on past sentences, would capture more coming sentences or not. In other words, the quality of the contraction is not optimised. These problems are addressed in on-going research.

4.3 Logarithmic growth

The results obtained for the size of the reduced corpus on real texts as well as on tagged sentences are shown in Figure 5.

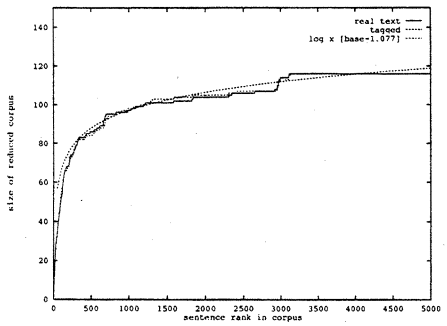


Figure 5: Size of reduced set.

The two curves, for real texts and tagged sentences, are again almost superimposed. As for these actual data, a logarithmic function seems a good approximation. Its base has been obtained experimentally: it is 1.077.

Because the method does not ensure that the bases are the same if the order of sentences in the corpus is different, we have at least to check that the cardinals of different bases are in the same range. As a gross test, we constructed a couple of reduced sets on the same corpus of 5 000 sentences, but with randomly permuted indices. Of course, a couple of experiments can't represent all the possible permutations over 5 000 elements, but in all the cases the shape of the curve is the same as in the first experiment, and fortunately, the number of elements in the bases obtained differs by only some elements.

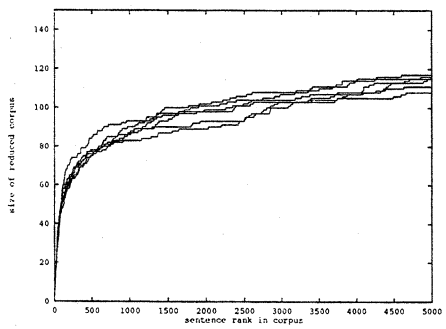


Figure 6: Size of different reduced sets.

From the practical point of view, these results are extremely excouraging, because they mean that a smaller corpus, from which any sentence of the real corpus can be obtained by analogy, has a size which is logarithmic in the size of a real corpus. Practically, 5 000 sentences are reduced to less than 120 sentences, and adding thousands of sentences in the real corpus would necessitate to add only few sentences in the reduced corpus.

Conclusion

We studied the behaviour of Saussurian analogy, as defined by a computational explanation, on real data: some thousands of sentences from texts of the ATR–Lancaster tree–bank.

Firstly, the absolute number of analogies in a text, while reading it, has been measured. The cubic growth has been experimentally confirmed.

Secondly, we inspected the size of the necessary memory of previously read sentences in order to obtain a a given number of analogies for a next sentence. Sudden rises in the curve obtained roughly correspond to changes in the style of the texts.

Thirdly, we constructed a reduced corpus, from which any sentence in the corpus can be generated and in which no analogy relation holds. This corpus is thus representative of the entire corpus relative to analogy.

References

[Black et al. 96] Ezra Black, Stephen Eubank, Kashioka Hideki, David Magerman, Roger Garside and Geoffrey Leech
Beyond Skeleton Parsing: Producing a Comprehensive Large–Scale General–English Treebank with Full Grammatical Analysis
Proceedings of COLING-96, Copenhagen, August 1996, pp. 107-112.

[Hofstadter et al. 94] Douglas Hofstadter and the Fluid Analogies Research Group
Fluid Concepts and Creative Analogies
Basic Books, New-York, 1994.

[Lepage 96] Yves Lepage
Ambiguities in analysis by analogy
Proceedings of MIDDIM-96, post-COLING seminar on interactive desambiguation, Christian Boitet ed., August 1996, pp. 93–100.

[Levenshtein 65] V.I. Levenshtein
Binary codes capable of correcting deletions, insertions and reversals
Dokl. Akad. Nauk SSSR, vol. 163, No. 4, August 1965, pp. 845-848.
English translation in *Soviet Physics-doklady* vol. 10, No. 8, February 1966, pp. 707-710.

[Saussure 16] Ferdinand de Saussure
Cours de linguistique générale
publié par Charles Bally et Albert Sechehaye, Payot, Lausanne et Paris, 1916.
フェルディナン ド ソシュール
一般言語学講義
小林英夫訳 岩波書店

[Wagner & Fischer 74] Robert A. Wagner and Michael J. Fischer
The String-to-String Correction Problem
Journal for the Association of Computing Machinery, Vol. 21, No. 1, January 1974, pp. 168-173.