Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus

Yves Lepage
ATR Spoken Language Translation Research Labs
Hikari-dai 2-2-2, Keihanna Science City
619-0288 Kyōto
Japan
yves.lepage@atr.jp

Abstract

The reality of analogies between words is refuted by noone (*e.g.*, *I walked* is to *to walk* as *I laughed* is to *to laugh*, noted *I walked* : *to walk* :: *I laughed* : *to laugh*). But computational linguists seem to be quite dubious about analogies between sentences: they would not be enough numerous to be of any use. We report experiments conducted on a multilingual corpus to estimate the number of analogies among the sentences that it contains. We give two estimates, a lower one and a higher one. As an analogy must be valid on the level of form as well as on the level of meaning, we relied on the idea that translation should preserve meaning to test for similar meanings.

1 Introduction

A long tradition in linguistics views analogy as a means for the speaker to analyze or produce new sentences[1]. To be linguistically relevant, an analogy should hold on the level of form as well as on the level of meaning.

In contrast to that, in Greek and Latin antiquity, *anomaly* designated those cases[2] where an analogy of meaning is not reflected by an analogy of form. (*e.g.*, 'I drink.' : 'I'd like to drink.' :: 'I can swim.' : 'I'd like to be able to swim.'[3]).

Conversely, the existence of analogies of form that are not verified on the level of meaning has been taken by the Generativists to indicate the independence of syntax (*e.g.*, *Abby is baking vegan pies.* : *Abby is baking.* :: *Abby is too tasteful to pour gravy on vegan pies.* : *Abby is too tasteful to pour gravy on.*[4]).

The purpose of this study is to estimate the number of "true analogies" present in a large corpus, *i.e.*, analogies which hold both on the level of form, as well as on the level of meaning.

Formally, let us denote 'A' as some meaning, and $\mathcal{L}(\text{'A'})$ as the set of all possible ways of realising 'A' in a particular language $\mathcal{L}$. Let us denote A as some realisation of 'A', *i.e.*, $A \in \mathcal{L}(\text{'A'})$. With these notations, we want to count, in a given corpus, all cases where the following holds[5].

$$\begin{array}{l} A \in \mathcal{L}(\text{'A'}) \ \wedge \\ B \in \mathcal{L}(\text{'B'}) \ \wedge \\ C \in \mathcal{L}(\text{'C'}) \ \wedge \\ D \in \mathcal{L}(\text{'D'}) \ \wedge \end{array}$$

$$A : B :: C : D \qquad \wedge \qquad \text{'A'} : \text{'B'} :: \text{'C'} : \text{'D'}$$

The reason for estimating the number of "true analogies" in a large corpus comes from the fact that it has been felt that "true analogies" between sentences are rare. There is a general feeling that analogy is well attested between words, *i.e.*, on the level of morphology, but not so much between sentences[6]. We will show that, at least in the corpus we used, this feeling has to be reconsidered.

[1] See, *inter alia*, (PAUL, 1920, chap.5), (de SAUSSURE, 1995, 3rd part, chap. iv), (BLOOMFIELD, 1933, p.276), (MOUNIN, 1968, p.119–120), (ITKONEN, 1994, p.48–50), (PULLUM, 1999, p.340–343).

[2] In those times, the cases considered were in fact in morphology (See, *.e.g.*, VARRO, *De lingua latine*).

[3] The meaning 'I'd like to be able to swim.' cannot be construed as **I'd like to can swim.*

[4] In the third sentence, *gravy* is poured on *vegan pies*, while it is poured on *Abby* (!) in the fourth sentence. This is not parallel to the first and second sentences where *Abby* plays the same role. This would imply that there is something about *to pour* which does not come directly from its form nor from its meaning.

[5] Needless to say, we disregard trivial cases of the form $A : A :: A : A$ and $A : A :: C : C$.

[6] It is not our purpose to address this issue, but the claim that some necessary analogies cannot be built from linguistic data available to children constitutes in fact the basis of the "arguments from the poverty of the stimulus." See (PULLUM and SCHOLTZ, 2002) and (LEGATE and YANG, 2002).

2 The corpus used

For this study, we used the Basic Traveler's Expression Corpus, or BTEC, for short[7]. This is a multilingual corpus of expressions from the travel and tourism domain. It contains 162,318 aligned translations in several languages. Here, we shall use Chinese, English and Japanese. There are 96,234 different sentences in Chinese, 97,769 in English and 103,274 in Japanese[8]. The sentences in BTEC are quite short as the figures in Table 1 show.

3 Analogies on the level of form

3.1 Method

On the level of form, a possible formalisation of analogy between strings of symbols has been proposed (LEPAGE, 2001) which renders an account of some analogies[9].

$$A : B :: C : D \;\Leftrightarrow\; \begin{cases} \forall a,\ |A|_a + |D|_a = |B|_a + |C|_a \\ \mathrm{dist}(A,B) = \mathrm{dist}(C,D) \\ \mathrm{dist}(A,C) = \mathrm{dist}(B,D) \end{cases}$$

Here, a is a character, whatever the writing system, and A, B, C and D are strings of characters. $|A|_a$ stands for the number of occurrences of a's in A. $\mathrm{dist}(A,B)$ is the edit distance between strings A and B, *i.e.*, the minimal number of insertions and deletions[10] of characters necessary to transform A into B.

Obviously, applied to sentences considered as strings of characters (not strings of words), this formalisation can only render an account of analogies on the level of form. Figure 1 shows examples of analogies meeting the above definition.

3.2 Results

It takes some ten days to gather all possible analogies of form using the above definition on a Pentium 4 computer at 2.8 Hz with 2 Gb memory for a corpus of around 100,000 sentences. Of course, we do not inspect all possible quadruples of sentences. Rather, a hierarchical coding of sentences based on counts of characters allows us to infer the absence of any analogy within large sets of sentences. This cuts the computational load. To compute edit distances, a fast bit string similarity computation algorithm (ALLISON and DIX, 1986) is used.

We counted the number of analogies of form in each of the monolingual Chinese, English and Japanese parts of the corpus using the previous formula. The examples of Figure 1 are actual examples of analogies retrieved. Table 2 shows the counts for each language. The numbers obtained are quite large. For English, we report around 2.5 million analogies of form involving more than 50,000 sentences. That is to say, half of the sentences of the corpus are already in immediate analogy with other sentences of the same corpus.

3.3 Discussion

The average number of analogies of form per sentence in each different language over all unique sentences may be estimated in the following way: $1,639,068 \;/\; 96,234 = 17.03$ for Chinese, $2,384,202 \;/\; 97,769 = 24.39$ for English and $1,910,065 \;/\; 103,274 = 18.50$ for Japanese. Averaging the sentences involved, this becomes: $5,059,979 \;/\; 49,675 = 33,00$ for Chinese, $2,384,202 \;/\; 53,250 = 44.77$ for English and $1,910,065 \;/\; 53,572 = 35.65$ for Japanese, which indicates that, on average, there are dozens of different ways to obtain these sentences by analogy with other sentences.

These counts are necessarily higher bounds of the numbers of "true analogies", as they rely on form only. For instance, the first analogy in Figure 1 is not a "true analogy". However, it is quite difficult to spot such analogies, so that the overall impression is that analogies of form which are not analogies of meaning are exceptions. So, our next problem will be to try to retain only those analogies which are also analogies of meaning.

[7] `http://www.c-star.org/`.

[8] The difference in size between Japanese and the other languages may be explained by the indifferent use of kanji or hiragana: *e.g.*, ください or 下さい /kudasai/ (*please*).

[9] Some cases of analogies are not considered by this definition, like reduplication: *e.g., I play tennis.* : *I play tennis. Do you play tennis too?* :: *I play guitar.* : *I play guitar. Do you play guitar too?*, or mirroring: *stressed* : *desserts* :: *reward* : *drawer*. Also, in reality, this formalisation is only an implication. But we shall use it as if it were an equivalence.

[10] Substitutions and transpositions are not considered as basic edit operations.

4 A lower estimate: meaning preservation through translation

4.1 Method

Computing analogies between structural representations is possible[11]. Unfortunately, the corpus we have at our disposal does not offer any structural representation. And it does not seem that tools are yet available which would deliver semantic (not syntactic) representations for all sentences of our corpus in all three languages we deal with.

Fortunately, common sense has it that translation preserves meaning[12], and, by definition, a multilingual corpus, like the one we use, contains corresponding utterances in different languages. Consequently, we shall assume that if two sentences A_1 and A_2 in two different languages are translations of one another (noted $A_1 \leftrightarrow A_2$), then, they should be the linguistic realisations of the same meaning, and reciprocally[13].

$$\exists \text{`A'} \;/\; \left\{ \begin{array}{l} A_1 \in \mathcal{L}_1(\text{`A'}) \\ A_2 \in \mathcal{L}_2(\text{`A'}) \end{array} \right. \quad \Leftrightarrow \quad A_1 \leftrightarrow A_2 \qquad (i)$$

Suppose that at least one analogy of form can be found to hold in every possible language of the world for some possible realisations of four given meanings. Then, for sure, the analogy of meaning can be said to hold.

$$\begin{array}{ll} \forall \mathcal{L}, & \exists A \in \mathcal{L}(\text{`A'}), \\ & \exists B \in \mathcal{L}(\text{`B'}), \\ & \exists C \in \mathcal{L}(\text{`C'}), \\ & \exists D \in \mathcal{L}(\text{`D'}), \qquad A : B :: C : D \end{array}$$

$$\Rightarrow \quad \text{`A'} : \text{`B'} :: \text{`C'} : \text{`D'}$$

If we suppose that the number of languages is finite, let us denote it n, counting the number of "true analogies" in a set of sentences in a given language, say $\mathcal{L}_1$, is tantamount to counting the cases described by the following formula (*ii*).

$$\begin{array}{l} A_1 \in \mathcal{L}_1(\text{`A'}) \;\wedge\; \ldots \;\wedge\; A_n \in \mathcal{L}_n(\text{`A'}) \;\wedge \\ B_1 \in \mathcal{L}_1(\text{`B'}) \;\wedge\; \ldots \;\wedge\; B_n \in \mathcal{L}_n(\text{`B'}) \;\wedge \\ C_1 \in \mathcal{L}_1(\text{`C'}) \;\wedge\; \ldots \;\wedge\; C_n \in \mathcal{L}_n(\text{`C'}) \;\wedge \\ D_1 \in \mathcal{L}_1(\text{`D'}) \;\wedge\; \ldots \;\wedge\; D_n \in \mathcal{L}_n(\text{`D'}) \;\wedge \\ \forall i \in \{1, \ldots, n\}, \quad A_i : B_i :: C_i : D_i \end{array}$$

$$\Rightarrow \quad \text{`A'} : \text{`B'} :: \text{`C'} : \text{`D'}$$

Of course, the problem is: how to test again all possible languages? Obviously, relying on more languages should give a higher accuracy to the method. Here, we have only three languages at our disposal. By relying on languages which are typologically different like Chinese, English and Japanese, it is reasonable to think that we somewhat counterbalance the small number of languages used.

To summarize, by using Equivalence (*i*), and by considering only sentences attested in our corpus, Formula (*ii*) can be restated as follows, when restricted to three languages.

$$\begin{array}{ccccccccc} A_1 & : & B_1 & :: & C_1 & : & D_1 & & \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \\ A_2 & : & B_2 & :: & C_2 & : & D_2 & \Rightarrow & \text{`A'} : \text{`B'} :: \text{`C'} : \text{`D'} \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow & & \\ A_3 & : & B_3 & :: & C_3 & : & D_3 & & \end{array}$$

Practically, thus, the number of "true analogies" is just the cardinal of the intersection set of the sets of analogies for each possible language.

4.2 Results

4.2.1 Pairwise intersection

Out of a total of 2,384,202 English analogies on the level of form, 238,135 are common with Chinese. They involve 25,554 sentences. Consequently, 10% of the English analogies of form may be thought to be analogies of form and meaning, *i.e.*, "true analogies", when relying only on Chinese.

Between English and Japanese the number of analogies in common is 336,287 (involving 24,674 sentences) which represents 14% of the English analogies. An example is given in Figure 2.

Between Chinese and Japanese very similar figures are obtained, as the number of analogies in common between these two languages is 329,429 (involving 25,127 sentences).

4.2.2 Chinese ∩ English ∩ Japanese

Taking the intersection of Chinese, English and Japanese leads to a figure of 68,164 "true analogies", involving 13,602 different sentences.

[11] (ITKONEN and HAUKIOJA, 1997) show how "true analogies" can be computed by relying at the same time on the surface and the structural representation of sentences.

[12] See (CARL, 1998) for an attempt at classifying machine translation systems relying on this idea.

[13] Note that, in this formula, $\mathcal{L}_1$ and $\mathcal{L}_2$ need not be different. If the language is the same, then, A_1 and A_2 are paraphrases.

4.3 Discussion

Although the number of analogies dropped from 2.5 million analogies of form in English, down to less than 70,000 when intersecting with Chinese and Japanese, one cannot say that the obtained figure is small.

The average number of "true analogies" per sentence over all the corpus is: $162,318 \ / \ 68,184 = 0.42$. In other words, in this corpus, one sentence is involved in about half a "true analogy" in average, taking it for granted that the linguistic differences between Chinese, English and Japanese filter real oppositions in meaning out of the oppositions captured by analogies of form.

The number of sentences involved in at least one analogy is 13,602, so that, more than one tenth of the sentences of the corpus are in an immediate analogical relation with other sentences of the corpus. Such a figure is not negligeable.

Averaging those sentences involved in at least one analogy gives the figure of $162,318 \ / \ 13,602 = 11.93$ "true analogies", which indicates that, on average, there are ten different ways to obtain these sentences by analogy with other sentences.

It is questionable whether those analogies that were lost in the successive intersections were really not analogies on the meaning level. In fact, the impression is that our experiment yielded a figure which is excessively low. An inspection by hand convinced us that almost all analogies which were discarded would have been considered by a human evaluator as "true analogies". Figure 1 shows two such examples. The problem is that the corresponding translations in other languages did not make an analogy of form. Other ways of saying could have made valid analogies of form. Consequently, the low number of translation equivalents available in our corpus is responsible of the low number of "true analogies" found by this method.

5 A higher estimate: translation by enforcement of "true analogies"

5.1 Method

The corpus we used is rather poor in translation equivalents, or paraphrases: an English sentence gets only 1.20 equivalent sentences on average when translated into Chinese, and only 1.52 into Japanese. If we would like to get a more accurate estimate of the number of "true analogies" in English, then our problem becomes that of increasing the number of possible translations of English sentences in Chinese and in Japanese, *i.e.*, to increase the number of paraphrases in Chinese and Japanese.

To address this problem, we adopted a view which is the opposite of our previous view. We decided to enforce "true analogies": given an analogy of form in a first language we forced it, when possible, to be reflected by an analogy of form in the second language. This should yield an estimate of the number of analogies in common between two languages which, if not necessarily more accurate, will at least be a higher estimate.

$$\begin{array}{ccccccccc} A_1 & : & B_1 & :: & C_1 & : & D_1 & & D_1 \\ \updownarrow & & \updownarrow & & \updownarrow & & & \Rightarrow & \updownarrow \\ A_2 & : & B_2 & :: & C_2 & : & D_2 & & D_2 \end{array}$$

To do so, the formula mentioned in section 3.1 is used in production, *i.e.*, D_2 is generated from the three sentences A_2, B_2 and C_2 when it is possible.

5.2 Results

Using the method described above, we automatically produced Chinese translations for those English sentences of the corpus which intervene in at least one analogy of form. This delivered an average of 51 different candidate sentences. As a whole, 48,351 sentences among 53,250 could be translated. By doing the same for Japanese, the average number of different sentences is higher: 174 for 47,702 translated sentences[14]. (For the reader to judge, Figure 3 shows examples of Japanese-to-English translations, rather than English-to-Japanese.)

The obtained translations were added to the corpus so as to increase the number of paraphrases in Chinese and Japanese. Then all counts were redone, and the new figures are listed under the title "Higher estimate" in Table 2.

5.3 Discussion

The new figure of 1,507,380 analogies for 49,052 sentences involved should be compared with the previous figures for the lower estimate. It is much higher, but it seems closer to the impression one gets when screening the analogies: analogies of form which are not analogies of meaning are very rare. However, the sentences that were obtained by enforcing analogies and then included in the corpus, are not always valid sentences. Figure 3 shows some such examples.

[14] Here again, we suspect the cause of the difference to be the indifferent use of kanji and hiragana.

Future works should thus consider the problem of filtering in some ways the translations obtained automatically using, for example, N-gram statistical models. After such a filtering, new counts should be performed again. However, the problem with such a filtering is that it may lose the morphological productivity of analogy.

6 Conclusion

In this paper, we reported experiments of counting the number of "true analogies," *i.e.*, analogies of form *and* meaning, between sentences contained in a large multilingual corpus, making the assumption that translation preserves meaning. We computed a lower and a higher estimates.

Using an English corpus of almost 100,000 different sentences, we obtained a lower estimate of almost 70,000 "true analogies" involving almost 14,000 sentences by intersecting analogies of form between Chinese, English and Japanese.

A higher estimate was obtained by enforcing analogies of form, *i.e.*, generating new sentences to fulfil analogies of form, so as to increase the number of paraphrases. More than a million and a half "true analogies" were found. They involve almost 50,000 sentences, *i.e.*, half of the sentences of the corpus. This meets our impression that almost all analogies of form between the English sentences of our corpus are also analogies of meaning.

Although we do not claim that analogy can explain everything about language, this work shows that, even when considering the lower estimate obtained, the number of "true analogies" that can be found in a corpus is far from being negligeable. Further research should focus on the way analogies are distributed over sentences, *i.e.*, on the characterisation of sentences involved in analogies.

Finally, as a speculative remark, similar countings as the ones reported above could contribute to the debate about "the argument from the poverty of the stimulus" if it were possible to reproduce them on such corpora as the CHILDES corpus[15].

7 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".

[15] `http://childes.psy.cmu.edu/`

References

Lloyd ALLISON and Trevor I. DIX. 1986. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310.

Leonard BLOOMFIELD. 1933. *Language*. Holt, New York.

Michael CARL. 1998. Meaning preservation in machine translation. In *ESSLI'98*, pages ?–?, Saarbrücken, March.

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

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

Esa ITKONEN. 1994. Iconicity, analogy, and universal grammar. *Journal of Pragmatics*, 22:37–53.

Julie Anne LEGATE and Charles D. YANG. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19:151–162.

Yves LEPAGE. 2001. Analogy and formal languages. In *Proceedings of FG/MOL 2001*, pages 373–378, Helsinki, August.

Georges MOUNIN. 1968. *Clefs pour la linguistique*. Bibliothèques 10/18, Seghers, Paris.

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

Geoffrey K. PULLUM and Barbara C. SCHOLTZ. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9–50.

Geoffrey K. PULLUM, 1999. *Generative grammar*, pages 340–343. The MIT Press, Cambridge.

	number of different sentences	size of sentences in characters mean	±	std.dev.
Chinese	96,234	11.00	±	5.77
English	97,769	35.14	±	18.81
Japanese	103,274	16.21	±	7.84

Table 1: Some statistics on the BTEC multilingual corpus.

	number of analogies	number of sentences involved	average number of analogies per sentence all corpus	involved only
	(i)	(ii)	(i) / 162,318	(i) / (ii)
Chinese	**1,639,068**	49,675	10.10	33.00
English	**2,384,202**	53,250	14.69	44.77
Japanese	**1,910,065**	53,572	11.77	35.65
Lower estimate:				
Chinese ∩ English	**238,135**	25,554	1.47	9.32
Chinese ∩ Japanese	**329,429**	25,127	2.03	13.11
English ∩ Japanese	**336,287**	24,674	2.07	13.63
"true analogies"	**68,164**	13,602	0.42	5.01
Higher estimate:				
Chinese ∩ English	**1,536,298**	49,297	9.46	31.16
Chinese ∩ Japanese	**1,569,037**	51,442	9.67	30.50
English ∩ Japanese	**1,901,689**	50,536	11.72	37.63
"true analogies"	**1,507,380**	49,052	9.29	30.73

Table 2: Number of analogies in the BTEC multilingual corpus.

Yea.	:	*Yep.*	::	*At five a.m.*	:	*At five p.m.*
Do you like music?	:	*Do you go to concerts often?*	::	*I like classical music.*	:	*I go to classical concerts often.*
I've lost my credit card.	:	*Do you accept credit card?*	::	*I've lost my travelers checks.*	:	*Do you accept travelers checks?*

Figure 1: Examples of analogies of form in English. The first one is not an analogy of meaning. The second and the third ones are analogies of meaning. However, their corresponding translations in the corpus (into Japanese for the second one, and into both Chinese and Japanese for the third one) do not make analogies of form.

I prefer Mexican food.	:	*I prefer Chinese food.*	::	*Is there a Mexican restaurant around here?*	:	*Is there a Chinese restaurant around here?*
↕		↕		↕		↕
メキシコ料理のほうが好きです。	:	中華料理のほうが好きです。	::	この辺りにメキシコ料理店はありますか。	:	この辺りに中華料理店はありますか。

Figure 2: An example of an analogy of form in two different languages that is an analogy of meaning.

ここで観光バスの切符を買えますか。
/koko de kankou basu no kippu wo kaemasu ka./
Can I buy a ticket for a sightseeing bus here?

9×	Can I buy a ticket for the sightseeing bus here?
6×	Can I get a any ticket for the sightseeing bus here?
3×	Could I buy sightseeing bus tickets here

児童書をください。
/zidousyo wo kudasai./
I'd like a children's book, please.

13×	I'd like a children's book, please
2×	I'd like a children's book, please.
2×	I'd like ae, pleas children's book
2×	Please give me a children's book
1×	Can I have a children's book
1×	Can I have a children's book, please
1×	Give me some children's book
1×	I would like a children's book, please
1×	I'd like a children's books.
1×	May I have a children's book

Figure 3: Actual translations in the corpus (above the dotted lines) and paraphrases produced by automatically enforcing analogies (under the dotted lines, with their output frequencies) for two sentences.