Errors of Omission in Translation

Graham Russell
RALI, Université de Montréal
C.P. 6128, succ. Centre-ville, Montréal (QC), Canada
russell@iro.umontreal.ca

Abstract

Automatic detection of translation errors represents one of the more promising applications of NLP techniques to this domain. This paper concentrates on one class of error, the inadvertent omission. To a greater extent than 'false friends', terminological inconsistency, etc., the detection of omissions raises problems both theoretical and practical in nature. These problems are discussed, and a technique is presented for identifying possible omissions in a completed translation by employing a model of translational equivalence between words. Examples are taken from a varied corpus of French-English bitext, and illustrate how different settings of the parameters of the system affect its performance. The approach is implemented as part of a translation-checking program.

1 Introduction

It has long been recognized that the provision of aids for translators is a promising area for the application of NLP techniques in the domain of translation. A recurring theme (Bashkansky et al. 1998, Macklovitch 1993, Picchi et al. 1992) is the "translator's workstation", integrating dictionaries, text editors, spelling and grammar checkers, terminological databases, translation memories, etc. Broadly speaking, assistance with the translation process takes two forms: on one hand access to information (via dictionaries and term banks), and on the other the detection of errors. While spelling and grammar checkers are of use in the translation context as elsewhere, more interesting from our perspective is the class of bilingual errors.

Most such errors can be described in terms of violation of constraints on the target-language translation E_T of a given source-language expression E_S. For example, deceptive cognates or 'false friends' such as English *library*, French *librairie* "bookshop" correspond to cases where E_T is one of an established list of 'forbidden' equivalents. Conversely, terminological inconsistency arises when E_T is not one of a list of 'obligatory' equivalents. A related problem arises in connection with proper names, some of which have standardized, term-like, translations (English *New Zealand*, French *Nouvelle-Zélande*) while others must be rendered unchanged (English/French *New Delhi*) in the target text.

In all of these cases, it is necessary to be able to decide, with some degree of accuracy, which E_T is the translation of the E_S in question. The notion of alignment thus plays a central role; either a fine-grained alignment at the level of lexical units (which might correspond to one or more graphological words), or a coarser alignment of more extended regions of text: sentences or even sequences of sentences.

Another type of translation error, and one which is in some ways far more problematic, is the unintended omission. As we show below, it is a relatively straightforward matter to identify source-language words for which no obvious translation exists within the target text. What is far more difficult is to determine the reason for each such case. For, as Melamed (1996) and many others have observed, a host of factors other than translator error underlie omissions. These range from the relatively linguistically tractable phenomena of anaphora and ellipsis, through the use of idioms and metaphors, to the translator's conscious decision not to render source-text material judged to be somehow redundant. The naive view of translation as an exact, meaning-preserving correspondence may be valid under some exceptional conditions, but it is of little help in the design of practical and general tools.

The central problem posed by this task, then, is how to distinguish erroneous, accidental, omissions such as might arise from inattention or faulty manipulation of word-processor controls from acceptable manifestations of "translator's licence". It is useful to compare the situation that holds with respect to detection of spelling errors. Here, fully automatic methods achieve high accuracy, in large part due to the fact that the correct way to write a given word is (i) well-defined, in the sense that it can be stored in a dictionary and is subject to finitely specifiable rules, and (ii) typically unique. Moreover, it is possible to construct a reasonably accurate model of the system's user, in terms of intention, common phonologically conditioned confusions, the likelihood of certain typing errors given keyboard layout, and so on. Adequate handling of false friends and terminology appears to lie rather close to this end of the complexity spectrum; omissions, however, rapidly lead us in the direction of general NLP.

Our conclusion, perhaps slightly surprisingly, is that fully accurate automatic detection of unwanted omissions is at least as hard as FAHQMT. For an MT system need not aim to generate all possible correct translations (and normally, should not even try). But detecting omissions requires the ability to know when the target text is not a possible translation of the source text as a whole, and this is a notoriously "AI-complete" problem, compounded by the intentional factors alluded to above.

So what can we realistically hope to achieve in this area, and will it be of any practical use? Clearly, simplifications will be necessary: first, a certain amount of noise and silence must be tolerated, ideally under some measure of user control; second, we shall restrict our attention to a subset of omissions permitting a relatively simple characterization. Melamed (1996), for example, suggests using the length of an omitted sequence as a cue to its error status, on the assumption that "[i]ntended omissions are seldom longer than a few words, while accidental omissions are often on the order of a sentence or more." This is certainly true, albeit far from wholly reliable: while in many cases it is safe to overlook a missing word and a flag a missing paragraph as an error, it is quite legitimate for a translator to intentionally suppress an entire paragraph where it is not believed to contain information of interest to the reader. Conversely, a single omitted word may crucially modify the required meaning.

In the remainder of this paper, we describe a system currently under development which inspects an aligned bitext, essentially attempting to answer two questions:

(i) Which words in the source-text component of a bitext region can be expected to have translation counterparts in the target-text component?

(ii) Of these, which fail to meet that expectation?

According to the answers obtained, the system presents the user with potential omissions. Since the questions are couched in terms of word occurrences, the system makes use of a type of translation model which states for certain source-language words which target-language words are most typically associated with them.[1] In principle, these associations might be derived from a bilingual dictionary. However, the current system acquires them by examining a previously aligned corpus. In the following section, we describe the alignment model employed in constructing the translation model and in processing the bitext to be checked, and the manner in which word associations are calculated.

2 A Framework for Studying Translation Omissions

2.1 Alignment

The type of alignment assumed here is one in which the parallel texts (bitext) are viewed as a sequence of **regions** spanning the two texts, each of which has the property that the content of its target-text component has been determined to correspond to that of its source-text component. Source (S) and target (T) texts are viewed as sequence of **segments** $\langle S_0^s, \ldots, S_m^s \rangle$, $\langle S_0^t, \ldots, S_n^t \rangle$; the precise definition of 'segment' is relatively unimportant, but is typically given in terms of the sentence or some other typographically delimitable unit. Note that in general $m \neq n$. The bitext is a sequence of regions, $\langle R_0, \ldots, R_p \rangle$, each of which has a source-text (R_i^s) and target-text (R_i^t) component. Each of these in turn consists of zero or more segments: $\langle S_i^s, \ldots S_{i+k}^s \rangle$, $\langle S_j^t, \ldots S_{j+l}^t \rangle$.

This model permits N-to-N mappings of segments within regions, but excludes dependencies between regions and crossing alignments.

2.2 Word Associations

What is required from the word-association component of the system is the knowledge of which pairs of words from the source and target languages typically cooccur with sufficient regularity to be considered as reliable cues to potential omissions. A bilingual dictionary might appear to be an obvious choice here, but this approach has a number of drawbacks: dictionaries are difficult to obtain, they tend to contain pairings drawn from the general language that may be inapplicable to specific kinds of text, and, most importantly for our purposes, they fail to convey any notion of strength of equivalence or preferred translation. Rather than acquiring word associations from a bilingual dictionary, then, the system learns from an aligned corpus, calculating them in a manner similar to that proposed by Melamed (1998b): cooccurrences of words within the same region of the training text are counted, and used as the basis for a statistical calculation of the strength of their association.

[1] This is not quite the same thing as a translation lexicon, for reasons given below.

The measure employed in the current system is the likelihood ratio described by Dunning (1993); this reflects the difference between the number of cooccurrences actually observed in the training text and the number that would have been expected, given the individual frequencies of the two words. Higher scores indicate that the words in question cooccur more frequently. In many cases, a high-scoring pair can be regarded as mutual translations, but the phenomenon of 'indirect association' complicates matters. For example, according to the word-association model employed in the experiments described below, the English verb *faxed* corresponds to French *télécopieur* "fax NOUN" as well as to *envoyé* "sent", since a frequent translation of the English is *envoyé par télécopieur* "sent by fax". If the aim were to produce a bilingual dictionary rather than characterize potential omissions, these indirect relations might be problematic, but in the present context they generally do no harm. Indeed, they provide a means of circumventing difficulties arising from some elliptical or incomplete translations: *faxed it yesterday* contains no omission with respect to *l'a envoyé hier* "sent it yesterday", even though *télécopieur* does not appear in the French region.

Cooccurrences are counted in the simplest way possible. Starting from a tokenized bitext, the count $C(x, y)$ for two words x and y is just the number of regions R_i in which $x \in R_i^s$ and $y \in R_i^t$. More accurate counts are of course possible: multiple occurrences within a given region could be taken into account in various ways. Melamed (1998a) discusses counting methods in more detail.

The word-association model is refined in a further stage of processing, essentially the first iteration of Melamed's (1998b) 'competitive linking' algorithm. Given the likelihood-ratio association score $A(x, y)$ for word pairs $\langle x, y\rangle \in S \times T$, the number of occasions is counted on which x and y cooccur within a region and neither x nor y enters into a higher-scoring association with any other word in that region:

$$\begin{aligned} L(x,y) \quad = \quad |R_i| : \quad & x \in R_i^s, \text{ and } y \in R_i^t, \text{ and} \\ & \neg\exists x' \in R_i^s : A(x',y) > A(x,y), \text{ and} \\ & \neg\exists y' \in R_i^t : A(x,y') > A(x,y). \end{aligned}$$

Comparing $L(x, y)$ with the cooccurrence count then provides their link ratio, $B(x, y)$, a more precise indication of how strongly x and y are associated:

$$B(x,y) = \frac{L(x,y)}{C(x,y)}$$

The initial word-correspondence model contained well over 2.5 million pairs; after removing pairs $\langle x, y\rangle$ for which $B(x, y) < 0.5$, approximately 131,000 remained.

A notable effect of filtering the word-correspondence model in this way is the elimination of most of the noise arising from chance cooccurrences with frequent words; items such as English *the*, *of* or French *de* "of", *la* "the FEM SNG" are so common that they figure prominently in the initial model as spurious partners of many other items. However, since the value of $A(la, the)$, $A(de, of)$ etc. tends to be very high, most occurrences of pairs such as $\langle ministre, the\rangle$, $\langle la, minister\rangle$ are preempted by the presence within the same region of $\langle la, the\rangle$ or $\langle de, of\rangle$. Interestingly, many of the more valuable indirect associations like $\langle faxed, télécopieur\rangle$ survive.

The training text employed in this study was composed of approximately 32 million words of English and 34 million words of French drawn from the Canadian Hansard (daily record of proceedings in the federal parliament).

2.3 Omission Detection

The omission detector takes as its input a bitext and the set of word-pairs with their association scores which survived the link-ratio filter described in the preceding section. The current version of the system accepts bitexts encoded in the CES "cesalign" format (Ide 1998), although nothing essential in its operation depends on this. For each potential omission, it writes out the putatively untranslated source text material, together with the region in question and the identifier of the segment where the omission is located.[2]

2.3.1 Resolved and unresolved tokens

The detection process resembles the link calculation in so far as it involves for each region R considering the score $A(x, y)$ of each word-pair $\langle x, y\rangle \in R^s \times R^t$. We refer to $\langle x, y\rangle$ as a **resolved pair** if it is the highest-scoring association of any involving x or y in that region:

$$\begin{aligned} \text{resolved}(x,y) \quad \equiv \quad & x \in R^s \text{ and } x \in R^t \text{ and} \\ & \neg\exists x' \in R^s : A(x', y) > A(x, y), \text{ and} \\ & \neg\exists y' \in R^t : A(x, y') > A(x, y). \end{aligned}$$

An **unresolved** token is a token in R^s for which one or more partners exist in the word-correspondence model, but which has not been resolved against any word in R^t:

$$\begin{aligned} \text{unresolved}(x) \quad \equiv \quad & \neg\exists y \in R^t : \text{resolved}(x, y), \text{ and} \\ & \exists z : A(x, z) \text{ is defined.} \end{aligned}$$

The intuition here is that a resolved pair represents a "true" translational equivalence within the current region, while an unresolved token occurs for one of three reasons:

(i) a deficiency in the word-correspondence model — a valid translation exists within the target region, but either does not occur in the model (being absent from the training text or failing to survive the filtering), or has been assigned too low an association score to be resolved;

(ii) a correct translation which is sufficiently novel or non-literal to exceed the capabilities of the model;

(iii) a genuine omission, either deliberate or accidental.

We refer to the remaining source-text tokens as 'dummies'; they are the tokens for which no pair exists in the word-correspondence model, and play no role in this component of the detection process.[3]

[2] Clearly, this last point is a matter of convenience; the only reason for creating such a display is to aid the human user, and it is not hard to imagine ways in which another program could make use of the same information expressed in a different form.

[3] However, they are employed in the first heuristic mentioned in section 3.3.

2.3.2 Region score

The detection of potential omissions within a given region R is based on the ratio of the sum of the **resolved scores** to the sum of the **unresolved scores** in R. In effect, the two sums are used as indicators of the amount and significance of the translated and untranslated material in R; the assumption is that a region is more likely to contain an omission if the total 'weight' of the tokens lacking an expected translation is relatively large in comparison with that of the tokens for which a counterpart has been found.

The resolved score $\mathrm{Res}(x)$ of a token x is just the score of the resolved pair of which it forms the first element, if any:

$$\mathrm{Res}(x) = \begin{cases} A(x,y) & \text{if } \exists y \in R^t : \text{resolved}(x,y), \\ 0 & \text{otherwise} \end{cases}$$

The unresolved score $\mathrm{Unres}(x)$ of a token x is a less concrete notion. What is required is a value which reflects the importance of the fact that no counterpart for x has been found in the current region. It seems reasonable to take this as being a function of the set of association scores $A_x = \{A(x,y)\}$ recorded for x in the word correspondence model. The possibilities include:

Max: the greatest value in A_x

Min: the least value in A_x

Mean: $\frac{\sum_{i \in A_x} i}{|A_x|}$

Median: Min + 1/2(Max − Min)

Effects of the choice are illustrated in section 3 below; here it is sufficient to note that Max will tend to lead the system to detect more omissions, at the expense of increased noise, while Min will have the opposite effect. The intention underlying Median and Mean is to achieve a compromise between the high and low values.

2.3.3 The decision

We are now in a position to state the criterion governing the operation of the system: a region R is identified as containing a potential omission if the following inequation is true:

$$w \left(\frac{\sum_{x \in R^s} \mathrm{Res}(x)}{\sum_{x \in R^s} \mathrm{Unres}(x)} \right) < 1$$

Here, w is a weighting coefficient employed in order to permit adjustments to the sensitivity of the detector. Again, its effect is illustrated in the results given in section 3.

3 Evaluation

3.1 Test Corpus

The test data was derived from the BAF corpus (Simard 1998) in the following manner. From each French-language text, a sample of approximately 10% of the regions was taken, subject to the requirement that they contain no omissions. The corresponding regions were then extracted from the corresponding English-language text. Since the resulting 'complete' subcorpus is free of omissions, it provides the basis for estimating the system's rate of overdetection, or 'false positives': none of its regions will be identified by a perfect detector.

In order to examine cases where the system misses true omissions, a copy of the first subcorpus was made in which some material was deleted from at least one segment in the target-language portion of each region. Possible deletions were 25%, 50%, 75% or 100% of the segment in question. Each region of this second 'sparse' subcorpus is therefore known to contain at least one omission; accordingly, a perfect detector should identify every one of its regions.

3.2 Performance Measurement

Raw coverage figures for the two test corpora can be obtained by counting the number of correct and incorrect identifications made for the two test corpora. However, similar identification and retrieval tasks are frequently evaluated in terms of slightly different two ntions: 'recall' and 'precision'. In general, the former indicates what proportion of the desired results are actually obtained, while the latter indicates what proportion of the results obtained are correct. In the present context, recall is the proportion of regions containing omissions which the system correctly identifies as such, and precision is the proportion of the regions identified by the system which actually contain omissions. The two values are customarily combined to produce a single 'F-score' as follows:

$$F = 2\left(\frac{p \cdot r}{p + r}\right)$$

3.3 Results and implications

In this section, we show some results obtained with differing values of the weighting parameter w and the unresolved-token scoring function Unres. Recall that the purpose of w is to permit control of the balance between recall and precision, and that values of w over 1 tend to privilege recall over precision (so relatively more true omissions are identified, along with more false positives) while values below 1 have the opposite effect.

We first show an example broken down by document (table 1). Here, w has been set to 1 so that that resolved and unresolved tokens have equal weight in the detection process. The column headed 'regions' give the sample size, those headed 'complete' and 'sparse' give the number of correct results for the corresponding subcorpus, while r, p and F are recall, precision and F-score respectively. Note that performance is relatively strong with the Hansard, CITI1 and TAO2 texts; this pattern holds with most other values of w and Unres. Table 2 summarizes results on the test corpus for three settings of Unres and four of w; the effect of varying the value of w is to alter the trade-off between recall and precision.

It should be borne in mind that the results shown here are obtained from testing on an artificial corpus which may be considered deviant both in the high proportion (effectively 50%) of omissions it contains, and in the random nature of the omissions

text	regions	complete	sparse	r	p	F
Citi1	55	40	49	0·73	0·87	0·79
Citi2	138	76	122	0·55	0·83	0·66
Cour	136	91	123	0·67	0·88	0·76
Hans	286	198	262	0·69	0·89	0·78
ILO	709	330	675	0·47	0·91	0·62
ONU	255	118	241	0·46	0·89	0·61
TAO1	35	21	31	0·60	0·84	0·70
TAO2	30	20	29	0·67	0·95	0·78
TAO3	17	9	15	0·53	0·82	0·64
Verne	228	141	192	0·62	0·80	0·70
Xerox	340	175	309	0·51	0·85	0·64
Total	2229	1219	2048	0·55	0·87	0·67

Table 1: Results with $w = 1, \text{Unres} = \mathsf{Max}$.

Unres	w	r	p	F
Max	1·25	0·48	0·89	0·62
	1·00	0·55	0·87	0·67
	0·50	0·67	0·75	0·71
	0·10	0·80	0·54	0·64
	0·05	0·81	0·52	0·64
Mean	1·25	0·35	0·89	0·48
	1·00	0·38	0·88	0·53
	0·50	0·56	0·83	0·67
	0·10	0·76	0·58	0·66
	0·05	0·79	0·54	0·64
Median	1·25	0·36	0·89	0·52
	1·00	0·42	0·88	0·57
	0·50	0·60	0·82	0·69
	0·10	0·78	0·57	0·65
	0·05	0·80	0·53	0·64

Table 2: Summary of results with varying w and Unres.

themselves. As might be expected, trials against more realistic texts show noise to be a problem; the system is led astray by its lack of knowledge concerning typical patterns of omission. Precision can be improved by adding a number of simple heuristics reflecting such patterns.

One heuristic is based on the assumption that true omissions are continuous, and of a certain minimum length K. This constraint can be implemented by restricting the system's attention to cases where the source text contains a sequence of at least K unresolved tokens, possibly interspersed with 'dummies' (see section 2.3.1). In effect, this amounts to searching for a string matching the regular expression

```
u(d*u){K-1}
```

where `u` denotes an unresolved token, `d*` zero or more non-links, and `K-1` the lower bound on the number of repetitions of the parenthesized subexpression, subject to the additional constraint that the sequence lies within a single segment containing no resolved link.

Another heuristic is based on the assumption that boundaries of true omissions coincide with certain easily identifiable markers within the source text: conjunctions, punctuation such as commas, parentheses and so on. The rationale for this is that certain omissions correspond to semantic units which have simply slipped the translator's mind, and that these units tend to be delimited by such markers. Again, this heuristic can be implemented by means of a constraint on candidate sequences.

Incorporating these two modifications into the detector enables us to reduce noise to the point where results such as those in table 3 are obtained. Note that these results are expressed in terms of correct and incorrect identifications and do not indicate the rate of recall.

Text	Proposed	Correct	Incorrect
Citi1	6	5	1
Citi2	6	4	2
Cour	1	0	1
Hans	7	6	1
ILO	8	5	3
ONU	1	1	0
TAO1	9	9	0
TAO2	2	1	1
TAO3	1	1	1
Verne	369	367	2
Xerox	70	62	8

Table 3: Results obtained using heuristics.

4 Discussion

The work described in this paper has been carried out as part of the RALI's machine-assisted translation program TransCheck. Development work is continuing in a number of areas mentioned below.

Unlike Melamed's (1996) proposal, which is based on a different model of alignment and which searches for character mismatches, the present system makes heavy use of word-association information derived from an aligned corpus. The choice of alignment method is therefore significant; those of Catizone et al. (1989) and Kay & Röscheisen (1993) are based on word distributions within the parallel text, and the potential for interference is clear: if regions are defined in terms of word-cooccurrences, then the strength of associations based on cooccurrence within those regions will be overstated. The method used here is a variant of that presented by Simard et al. (1992), which employs a much weaker type of lexical correspondence, namely string-identity of (prefixes of) the words being compared, in conjunction with length-based alignment as developed by Gale & Church (1993).

As indicated in section 1, we have not attempted to incorporate information from a conventional bilingual dictionary into the system's knowledge of word-associations. Nevertheless, it may be beneficial to do so, provided that some principled decision can be reached on how to merge it with the model. One possibility would be to regard pairs listed in the dictionary as more strongly associated than any acquired from parallel texts, so privileging them in the detection process. Note that many users of the system can be expected to possess or have access to specialized glossaries, termbanks, etc.; it is important to allow information from such sources to override that obtained from the training corpus when appropriate.

Finally, it is worth pointing out that the general approach presented here has wider application. A false move with a word processor can just as easily insert or copy a portion of text as delete it; in this case, rather than the source containing material with no equivalent in the target, the direction of the asymmetry is reversed. It is a simple matter to invert the sense of the operations described above so as to detect spurious insertions of this kind. More interestingly, the same technique may be employed to improve the performance of automatic alignment programs. One of the most common types of error with some popular algorithms is the displacement of a region boundary so that, for example, the translation of part of the content of R_i^s appears in an adjacent region R_{i+1}^t. This configuration is equivalent to that which would arise from the simultaneous omission of the corresponding material from R_i^t and its insertion into R_{i+1}^t.

Acknowledgments

I am grateful to my colleagues at the RALI for their comments on this work, in particular Elliott Macklovitch and George Foster, who also supplied an efficient program for extracting, counting, and storing word-pairs.

References

Bashkansky, Guy and Uzzi Ornan: 1998, 'Monolingual Translator Workstation', in D. Farwell, L. Gerber and E. Hovy (eds.) *Machine Translation and the Information Soup: Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, Lecture Notes in Artificial Intelligence no. 1529, Berlin: Springer, pp. 136–149.

Catizone, Roberta, Graham Russell and Susan Warwick: 1989, 'Deriving Translation Data from Bilingual Texts', in U. Zernik (ed.) *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, August 21st, 1989.

Dunning, Ted: 1993, 'Accurate Methods for the Statistics of Surprise and Coincidence', *Computational Linguistics* 19(1), pp. 61–74.

Gale, William A., and Kenneth W. Church: 1993, 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics* 19(1), pp. 75–102.

Ide, Nancy: 1998, 'Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora', in *First International Conference on Language Resources and Evaluation*, vol. 1, Granada, pp. 463–469.

Kay, Martin and Martin Röscheisen: 1993, 'Text-Translation Alignment', in *Computational Linguistics* 19(1), pp. 121–142.

Macklovitch, Elliott: 1993, 'PTT-3: A Third Version of the CITI's Workstation for Translators', Techical Report, CITI

Melamed, I. Dan: 1996, 'Automatic Detection of Omissions in Translation', in *16th International Conference on Computational Linguistics*, Copenhagen, pp. 764–769.

Melamed, I. Dan: 1998a, 'Models of Co-occurrence', ICRS Technical Report 98-05, University of Pennsylvania.

Melamed, I. Dan: 1998b, 'Word-to-Word Models of Translational Equivalence', ICRS Technical Report 98-08, University of Pennsylvania.

Picchi, Eugenio, Carol Peters and Elisabetta Marina: 1992, 'A Translator's Workstation', in *14th International Conference on Computational Linguistics* vol. 3, Nantes, pp. 972–976.

Simard, Michel: 1998, 'The BAF: A Corpus of English-French Bitext', in *First International Conference on Language Resources and Evaluation* vol. 1, Granada, pp. 489–494.

Simard, Michel, George F. Foster and Pierre Isabelle: 1992, 'Using Cognates to Align Sentences in Bilingual Corpora', in *TMI'92*, Montreal, pp. 67–81.