Identifying Word Correspondences in Parallel Texts

William A. Gale
Kenneth W. Church[1]

AT&T Bell Laboratories
Murray Hill, N.J., 07974
gale@research.att.com

1. Introduction

Researchers in both machine translation (e.g., Brown *et al*, 1990) and bilingual lexicography (e.g., Klavans and Tzoukermann, 1990) have recently become interested in studying *parallel texts* (also known as *bilingual corpora*), bodies of text such as the Canadian Hansards (parliamentary debates) which are available in multiple languages (such as French and English). Much of the current excitement surrounding parallel texts was initiated by Brown *et al.* (1990), who outline a self-organizing method for using these parallel texts to build a machine translation system.

Brown *et al.* begin by aligning the parallel texts at the sentence level. In our experience, 90% of the English sentences match exactly one French sentence, but other possibilities, especially two sentences matching one or one matching two, are not uncommon. There has been quite a bit of recent work on sentence alignment, e.g., (Brown, Lai and Mercer, 1991), (Kay and Röscheisen, 1988), (Catizone, Russell, and Warwick, to appear); we use a method described in (Gale and Church, 1991) which makes use of the fact that the length of a text (in characters) is highly correlated (0.991) with the length of its translation. A probabilistic score is assigned to each proposed match, based on the lengths of the two regions and some simple assumptions about the distributions of these two lengths. This probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences.

After sentences have been aligned, the second step is to identify correspondences at the word level. That is, we would like to know which words in the English text correspond to which words in the French text. The identification of word level correspondences is the main topic of this paper.

We wish to distinguish the terms *alignment* and *correspondence*. The term *alignment* will be used when order constraints must be preserved and the term *correspondence* will be used when order constraints need not be preserved and crossing dependencies are permitted. We refer to the matching problem at the word level as a correspondence problem because it is important to model crossing dependencies (e.g., *sales volume* and *volume des ventes*). In contrast, we refer to the matching problem at the sentence level as an alignment problem because we believe that it is not necessary to model crossing dependencies at the sentence level as they are quite rare and can be ignored for now.

Here is an example of our word correspondence program. Given the input English and French sentences:

> *English*
> we took the initiative in assessing and amending current legislation and policies to ensure that they reflect a broad interpretation of the charter .
>
> *French*
> nous avons pris l' initiative d' évaluer et de modifier des lois et des politiques en vigueur afin qu' elles correspondent à une interprétation généreuse de la charte .

The program would produce the following correspondences:

> *Output:*
> we/nous took/0 the/0 initiative/initiative in/0 assessing/évaluer and/et ammending/modifier current/0 legislation/0 and/et policies/politiques to/à ensure/0 that/qu' they/elles reflect/0 a/une broad/0 interpretation/interprétation of/de the/la charter/charte ./.

In this example, 15 out of the 23 (65%) English words were matched with a French word (with *to/à* in error), and 8 of the English words were left unmatched (paired with "0"). Throughout this work, we have focused our attention on robust statistics that tend to avoid making hard decisions when there isn't much confidence. In other words, we favor methods with relatively high

1. The second author is visiting USC/ISI, 4676 Admiralty Way, Marina del Rey, CA 90292, USA until 9/1/91.

precision and possibly low recall. For now, we are more concerned with errors of commission than errors of omission. Based on a sample of 800 sentences, we estimate that our word matching procedure matches 61% of the English words with some French word, and about 95% of these pairs match the English word with the appropriate French word.

After word correspondences have been identified, it is possible to estimate a probabilistic transfer dictionary. The entry for "the" found in (Brown *et al.*) includes the estimates of Prob(*le* | *the*)=.61 and Prob(*la* | *the*)=.18. Brown *et al.* show how this probabilistic transfer dictionary can be combined with a trigram grammar in order to produce a machine translation system. Since this paper is primarily concerned with the identification of word correspondences, we will not go into these other very interesting issues here.

2. Applications Beyond MT

As mentioned above, MT is not the only motivation for sentence alignment and word correspondence. Computational linguists (e.g., Klavans and Tzoukermann, 1990) have recently become interested in bilingual concordances. Table 1, for example, shows a bilingual concordance contrasting the uses of *bank* that are translated as *banque* with those that are translated as *banc*. Of course it is well-know that sense disambiguation is important for many natural language applications (including MT as well as many others). In the past, this fact has been seen as a serious obstacle blocking progress in natural language research since sense disambiguation is a very tricky unsolved problem, and it is unlikely that it will be solved in the near future.

However, we prefer to view these same facts in a more optimistic light. In many cases, the French text can be used to disambiguate the English text, so that the French can be used to generate a corpus of (partially) sense-disambiguated English text. Such a sense-disambiguated corpus would be a valuable resource for all kinds of natural language applications. In particular, the corpus could be used to develop and test sense-disambiguation algorithms. For example, if you have an algorithm that is intended to distinguish the "money" sense of bank from the "place" sense of bank, then you might apply your algorithm to all of the uses of *bank* in the English portion of the parallel corpus and use the French text to grade the results. That is, you would say that your program was correct if it identified a use of *bank* as a "money" sense and it was translated as *banque*, and you would say that the program was incorrect if the program identified the use as a "money" sense and it was translated as *banc*. Thus, the availability of the French text provides a valuable research opportunity, for both monolingual and bilingual applications. The French text can be used to help clarify distinctions in the English text that may not be obvious to a dumb computer.

Table 1: A Bilingual Concordance Based on Aligned Sentences

bank/ banque ("money" sense)

f finance (mr . wilson) and the governor of the	bank of canada have frequently on
es finances (m . wilson) et le gouverneur de la	banque du canada ont fréquemme
reduced by over 800 per cent in one week through	bank action . SENT there was a h
us de 800 p . 100 en une semaine à cause d' une	banque . SENT voilà un chemisie

bank/ banc ("place" sense)

h a forum . SENT such was the case in the georges	bank issue which was settled betw
entre les états-unis et le canada à propos du	banc de george . SENT > c' est da
han i did . SENT he said the nose and tail of the	bank were surrendered by this gov
gouvernement avait cédé les extrémités du	banc . SENT en fait , lors des nég

3. Using Word Correspondances Rather than Sentence Alignments

Most bilingual concordance programs such as ISSCO's BCP program mentioned in footnote 1 of (Warwick and Russel, 1990) and a similar program mentioned on page 20 of (Klavans and Tzoukermann, 1990) are based on aligned sentences rather than word correspondences. Table 1 shows an example of such a sentence-based concordance program. These sentence-based programs require the user to supply the program with both an English and a French word (e.g., *bank* and *banque*). In contrast, a word-based concordance program is given just *bank* and finds the French translations by making use of the word correspondences.

The advantage of the word-based approach becomes important for complicated words like *take*, where it is difficult for users to generate many of the possible translations. *take* is often used in complex idiomatic expressions, and consequently, there are many uses of *take* that should not be translated with *prendre*. In fact, most uses of *take* are not translated with *prendre* (or any of its morphological variants). The word-based bilingual concordances show this fairly clearly. We find that only 23% of the uses of *take* are translated with a form of *prendre*, a figure is fairly consistent with IBM's estimate of 28% (Brown, personal communication). The striking absence of *prendre* is consistent with the observation in the Cobuild dictionary (Sinclair *et al.*, 1987, p. 1488) that "[t]he most frequent use of **take** is in expressions where it does not have a very distinct meaning of its own, but where most of the meaning is in ... the direct object."

4. Two Possible Problems with the EM Algorithm

This paper is primarily concerned with the task of identifying word correspondences. There is relatively little discussion of this topic in Brown *et al.* (1990), although a brief mention of the EM algorithm is made. We decided to look for an alternative estimation algorithm for two reasons.

First, their procedure appears to require a prohibitive amount of memory. We observed that they limited the sizes of the English and French vocabularies, V_E and V_F, respectively, to just 9000 words each. Having constrained the vocabularies in this way, there were a mere 81 million parameters to estimate, all of which could be squeezed into memory at the same time. However, if the two vocabularies are increased to a more realistic size of 10^6 words, then there are 10^{12} parameters to estimate, and it is no longer practical to store all of them in memory. (Apparently, in some more recent unpublished work (Brown, personal communication), they have also found a way to scale up the size of the vocabulary).

Secondly, we were concerned that their estimates might lack robustness (at least in some cases):

> "This algorithm leads to a local maximum of the probability of the observed pairs as a function of the parameters of the model. There may be many such local maxima. The particular one at which we arrive will, in general, depend on the initial choice of parameters." (Brown *et al.*, p. 82)

In particular, we looked at their estimates for the word *hear*, which is surprisingly often translated as *bravo* (especially, *Hear, hear!* → *Bravo!*), though it is not clear just how common this is. Brown *et al.* reported that more than 99% of the uses of *hear* were translated with *bravo*, whereas we estimate the fraction to be much closer to 60% (which is fairly consistent with their more recent estimates (Brown, personal communication)). The fact that estimates can vary so widely from 99% to 60% indicates that there might be a serious problem with robustness. It became clear after more private discussions that our methods were coming up with substantially different probability estimates for quite a number of words. It is not clear that the maximum likelihood methods are robust enough to produce estimates that can be reliably replicated in other laboratories.

5. Contingency Tables

Because of the memory and robustness questions, we decided to explore an alternative to the EM algorithm. Table 2 illustrates a two-by-two contingency table for the English word *house* and the French word *chambre*. Cell a (upper-left) counts the number of sentences (aligned regions) that contain both *house* and *chambre*. Cell b (upper-right) counts the number of regions that contain *house* but not *chambre*. Cells c and d fill out the pattern in the obvious way.

The table can be computed from $freq(house, chambre)$, $freq(house)$ and $freq(chambre)$, the number of aligned regions that contain one or both these words, and from $N = 897{,}077$, the total number of regions.

$a = freq(house, chambre)$
$b = freq(house) - freq(house, chambre)$
$c = freq(chambre) - freq(house, chambre)$
$d = N - a - b - c$

Table 2: A Contingency Table

	chambre	
house	31,950	12,004
	4,793	848,330

We can now measure the association between *house* and *chambre* by making use of any one of a number of association measures such as mutual information. ϕ^2, a χ^2-like statistic, seems to be a particularly good choice because it makes good use of the off-diagonal cells b and c.

$$\phi^2 = \frac{(ad - bc)^2}{(a + b)\ (a + c)\ (b + d)\ (c + d)}$$

ϕ^2 is bounded between 0 and 1. In this case, ϕ^2 is 0.62, a relatively high value, indicating the two words are strongly associated, and that they may be translations of one another. One can make this argument more rigorous by measuring the confidence that ϕ^2 is different from chance (zero). In this case, the variance of ϕ^2 is estimated to be 2.5×10^{-5} (see the section "Calculation of Variances"), and hence $t = \phi^2/\sqrt{(var(\phi^2))} = 0.62/\sqrt{2.5\times10^{-5}} = 123$. With such a large t, we can very confidently reject the null hypothesis and assume that there is very likely to be an association between *house* and *chambre*.

6. A Near Miss

i.e. how do you avoid believing
House = Commons

One might contrast *house/chambre* with a near miss such as *house/communes* (see Table 3).

Table 3: A Near Miss

	communes	
house	4,974	38,980
	441	852,682

Unfortunately, this pair is also significantly different from zero ($t = 31$) because there are many references in the Canadian Hansard to the English phrase *House of Commons* and its French equivalent *Chambre des Communes*. How do we know that *house* is more associated with *chambre* than with *communes*? Note that mutual information does not distinguish these two pairs. Recall the mutual information $I(x;y)$ is computed by

$$\log_2 \frac{Prob(x,y)}{Prob(x)Prob(y)}$$

where $Prob(x,y) = a/N$, $Prob(x) = (a + b)/N$, and $Prob(y) = (a + c)/N$. If we plug these numbers into the formulas, we find that *house* and *chambre* actually have a lower mutual information value than *house* and *communes*: $I(house;chambre) = 4.1$ while $I(house;communes) = 4.2$.

Mutual information picks up the fact that there are strong associations in both cases. Unfortunately, it is not very good at deciding which association is stronger. Crucially, it does not make very good use of the off-diagonal cells b and c, which are often better estimated than cell a since the counts in b and c are often larger than those in a.

In this case, the crucial difference is that cell b is much smaller in Table 2 than in Table 3. ϕ^2 picks up this difference; Table 3 has a ϕ^2 of 0.098, significantly less than Table 2's ϕ^2 of 0.62:

$$t = \frac{\phi^2(h,ch) - \phi^2(h,co)}{\sqrt{var^2(\phi^2(h,ch)) + var^2(\phi^2(h,co))}} = \frac{0.62 - 0.099}{\sqrt{2.5\times10^{-5} + 9.9\times10^{-6}}} = 88$$

Thus, we can very confidently say that *house* (h)is more associated with *chambre* (ch) than with *communes* (co).

7. Calculation of Variances

The estimate of $var(\phi^2)$ is very important to this argument. We use the following reasoning:

$$var(\phi^2) \approx \left[\frac{\partial\phi^2}{\partial a}\right]^2 var(a) + \left[\frac{\partial\phi^2}{\partial b}\right]^2 var(b) + \left[\frac{\partial\phi^2}{\partial c}\right]^2 var(c) + \left[\frac{\partial\phi^2}{\partial d}\right]^2 var(d)$$

where $var(a) \approx a$, $var(b) \approx b$, $var(c) \approx c$ and $var(d) \approx a + b + c$. A direct calculation of this is valid when ϕ^2 is small:

$$var_{small}(\phi^2) \approx \phi^2(4 \frac{a^2 var(d) + d^2 a + b^2 c + c^2 b}{(a + b)(c + d)(a + c)(b + d)} + \phi^2(\frac{1}{a + b} + \frac{c + var(d)}{(c + d)^2} + \frac{1}{a + c} + \frac{b + var(d)}{(b + d)^2}))$$

As ϕ^2 approaches 1, $var(\phi^2)$ decreases to 0, which makes the equation for var_{small} unsuitable as an estimate of the variance. We calculate a variance for this case by assuming that $bc << ad$, which implies that $\phi^2 \approx 1 - (b + c)/a$. With this assumption, we obtain

$$var_{large}(\phi 2) \approx a^{-2}(b + c)(1 + \frac{b + c}{a})$$

We do not have an exact relation to specify when ϕ^2 is large and when it is small. Rather, we observe that each estimate produces a value that is small in its domain, so we estimate the variance of ϕ^2 by the minimum of the two cases: $var(\phi^2) \approx \min(var_{small}, var_{large})$.

8. Selecting Pairs

We have now seen how we could decide that *house* and *chambre* are more associated than *house* and *communes*. But why did we decide to look at these pairs of words and not some others? As we mentioned before, we probably can't afford to look at all $V_E V_F$ pairs unless we limit the vocabulary sizes down to something like the 9000 word limit in Brown *et al.* And even then, there would be 81 million pairs to consider. If the training corpus is not too large (e.g., 50,000 regions), then it is possible to consider all pairs of words that actually co-occur in at least one region (i.e., $a \neq 0$). Unfortunately, with a training corpus of $N = 890{,}000$ regions, we have found that there are too many such pairs and it becomes necessary to be more selective (heuristic).

We have had fairly good success with a progressive deepening strategy. That is, select a small set of regions (e.g., 10,000) and use all of the training material to compute ϕ^2 for all pairs of words that appear in any of these 10,000 regions. Select the best pairs. That is, take a pair (x, y) if it has a ϕ^2 significantly better than any other pair of the form (x, z) or (w, y). This procedure would take *house/chambre* but not *house/communes*. Repeat this operation, using larger and larger samples of the training corpus to suggest possibly interesting pairs. On each iteration, remove pairs of words from the training corpus that have already been selected so that other alternatives can be identified. We have completed four passes of this algorithm, and selected more than a thousand pairs on each iteration.

Iteration	Sample Size	Number of Pairs Selected
0	10,000	1223
1	30,000	1537
2	50,000	1692
3	220,000	1967

A few of the selected pairs are shown below. The first column indicates the iteration that the pair was selected on. The second column indicates the number of sentences (aligned regions) that the pair appears in. Note that the most frequent pairs are usually selected first, leaving less important pairs to be picked up on subsequent iterations. Thus, for example, *accept/ accepter* is selected before *accept/ accepte*. Based on a sample of 1000 pairs, about 98% of the selected pairs of words are translations. Here, as elsewhere, we act to keep our errors of commission low.

Iteration	Freq	English	French
2	278	accept	accepte
0	1335	accept	accepter
3	111	accept	acceptons
1	165	acceptable	acceptables
2	101	acceptable	inacceptable

1	90	acceptance	acceptation
1	596	accepted	accepté
1	55	accepting	acceptant
3	130	accepting	accepter
0	62	accepts	accepte

After a few iterations, it became clear that many of the pairs that were being selected were morphologically related to pairs that had already been selected on a previous iteration. A remarkably simple heuristic seemed to work fairly well to incorporate this observation. That is, assume that two pairs are morphologically related if both words start with the same first 5 characters. Then, select a pair if it is morphologically related to a pair that is already selected and it appears "significantly often" (in many more sentences than you would expect by chance) on any iteration. This very simple heuristic more than doubled the number of pairs that had been selected on the first four iterations, from 6419 to 13,466. As we will see in the next section, these 13 thousand pairs cover more than half of the words in the text. Again, the error rate for pairs selected by this procedure was low, less than two percent.

9. Returing to the Sentence Context

It is now time to try to put these pairs back into their sentence context. Consider the pair of sentences mentioned previously.

> *English:*
> we took the initiative in assessing and amending current legislation and policies to ensure that they reflect a broad interpretation of the charter .
>
> *French:*
> nous avons pris l' initiative d' évaluer et de modifier des lois et des politiques en vigueur afin qu' elles correspondent à une interprétation généreuse de la charte .

The matching procedure attempts to match English and French words using the selected pairs. When there are several possibilities, the procedure uses a slope condition to select the best pair. Thus, for example, there are two instances of the word *and* in the English sentence and two instances of the word *et* in the French sentence. We prefer to match the first *and* to the first *et* and the second *and* to the second *et*, as illustrated below. (The *i* and *j* columns give the positions into the English and French sentences, respectively. The column labeled *slope* indicates the difference between the *j* values for the current French word and the last previous non-NULL French word.

i	English	j	French	slope	score
1	we	1	nous	1	–0.5
2	took		NULL		–5.5
3	the		NULL		–10.5
4	initiative	5	initiative	4	–14.2
5	in		NULL		–19.2
6	assessing	7	évaluer	2	–21.5
7	and	8	et	1	–22.0
8	amending	10	modifier	2	–24.2
9	current		NULL		–29.2
10	legislation		NULL		–34.2
11	and	13	et	3	–37.3
12	policies	15	politiques	2	–39.6
13	to	22	à	7	–44.5
14	ensure		NULL		–49.5
15	that	19	qu'	–3	–54.1
16	they	20	elles	1	–54.6
17	reflect		NULL		–59.6
18	a	23	une	3	–62.7
19	broad		NULL		–67.7
20	interpretation	24	interprétation	1	–68.2
21	of	26	de	2	–70.4
22	the	27	la	1	–70.9
23	charter	28	charte	1	–71.4
24	.	29	.	1	–71.9

The matching procedure uses a dynamic programming optimization to find the sequence of *j* values with the best score. A sequence of *j* values is scored with

$$\sum_j \log prob(match \mid slope_j)$$

Using Bayes rule, the $prob(match \mid slope_j)$ is rewritten as $prob(slope_j \mid match)\ prob(match)$. Both terms were estimated empirically.

The second term is determined by the fan-in, the number of possible matches that a particular *j* value might play a role in. In this example, most of the *j* values had a fan-in of 1. However, the two instances of *et* had a fan-in of 2 because they could match either of the two instances of *and*. The score is smaller for both of these uses of *et* because there is more uncertainty. We considered three cases: the fan-in is 1, 2 or many. The log $prob(match)$ in each of these three cases is –0.05, –0.34 and –0.43, respectively.

The first term is also determined empirically. The score is maximized for a slope of 1. In this case, log $prob(slope \mid match)$ is –0.46. The score falls off rapidly with larger or smaller slopes.

The dynamic programming optimization is also given the choice to match an English word to NULL. If the procedure elects this option, then a constant, log $prob(NULL)$, is added to the score. This value is set so that the matching procedure will avoid making hard decisions when it isn't sure. For example, the 5^{th} English word (*in*) could have been matched with 16^{th} French word (*en*), but it didn't do so because log $prob(NULL)$ was more than the score of such a radical reordering. We have found that –5 is a good setting for log *prob*(*match*). If we set the value much higher, then the matching procedure attempts to reorder the text too much. If we set the value much lower, then the matching procedure does not attempt to reorder the text enough.

This matching procedure works remarkably well. As mentioned above, based on a sample of 800 sentences, we estimate that the procedure matches 61% of the English words with some French word, and about 95% of these pairs match the English word with the appropriate French word. All but one of these errors of commission involved a function word, usually one surrounded on both sides by words that could not be matched.

10. Conclusions

We have been studying how to find corresponding words in parallel texts given aligned regions. We have introduced several novel techniques that make substantial progress toward this goal. The philosophy underlying all our techniques is to keep errors of commission low. Whatever words *are* matched by these robust techniques should almost always be correct. Then, at any stage, the words that are matched can be used confidently for further research.

The first technique we have introduced is the measurement of association of pairs of words by ϕ^2, based on a two by two contingency table. This measure does better than mutual information at showing which pairs of words are translations, because it accounts for the cases in which one of the words occurs and the other does not. We apply this measure iteratively. Our caution is expressed by selecting at most one pair of words containing a given word on each iteration. The ϕ^2 measure for a selected pair must be significantly greater than the ϕ^2 measures for each of the words of the pair and any other suggested translation.

The iteration is accompanied by a progressive enlargement of possibly interesting pairs. We could not study all pairs of words, or even all occurring pairs of words. Rather we take all the occuring pairs in a progressively enlarged sample of regions. This does propose the most frequently cooccurring pairs first. On each iteration we delete the pairs of words that have already been selected, thereby reducing the confusion among collocates. Our caution was expressed by hand checking the accuracy of selected pairs after each iteration. We chose techniques which could give 98 percent accuracy on the selected pairs. This has not been a blind automatic procedure, but one controlled at each step by human expertise.

When we observed that many of the pairs considered contained morphological variants of a pair selected, we allowed such pairs to be accepted if they also had a ϕ^2 significantly greater than chance.

Several of our tests acknowledge that any function, such as ϕ^2, of noisy data, such as frequencies, is itself a noisy measure. Therefore our caution is to require not just that one measure be greater than another, but that it be significantly greater. This calculation is made using an estimate of the variance of ϕ^2.

We then used the selected word pairs to suggest word correspondences within a given aligned region. The alignment was done by a dynamic programming technique with a parameter that controlled how certain we should be before accepting a specific pair of words as corresponding. We set the parameter to give results that are quite likely to be correct. Currently we suggest correspondences for about 60 percent of the words, and when we do suggest a correspondence we are correct in about 95 percent of cases.

This is work in progress. We expect that in the future the coverage can be increased substantially above 60% while errors can be decreased somewhat from 5%. We believe that errors of omission are much less important than errors of commission and expect to continue choosing techniques accordingly.

References

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin (1990) "A Statistical Approach to Machine Translation," *Computational Linguistics*, v 16, pp 79-85.

Brown, P, J. Lai, R. Mercer (1991) "Aligning Sentences in Parallel Corpora," IBM Report submitted to 29th Annual Meeting of the Association for Computational Linguistics.

Catizone, R., G. Russell, and S. Warwick (to appear) "Deriving Translation Data from Bilingual Texts," in Zernik (ed), *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum.

Church, K., (1988) "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," Second Conference on Applied Natural Language Processing, Austin, Texas.

Gale, W. and K. Church (1990) "A Program for Aligning Sentences in Bilingual Corpora," unpublished ms., submitted to 29th Annual Meeting of the Association for Computational Linguistics.

Kay, M. and M. Röscheisen (1988) "Text-Translation Alignment," unpublished ms., Xerox Palo Alto Research Center.

Klavans, J., and E. Tzoukermann (1990) "The BICORD System," *COLING-90*, pp 174-179.

Warwick, S. and G. Russell (1990) "Bilingual Concordancing and Bilingual Lexicography," Euralex 1990.