Identifying Word Correspondances in Parallel Texts

William A. Gale
Kenneth W. Church

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. This approach is extremely controversial among researchers in machine translation because it questions many of the basic assumptions that have dominated the field since the 1950s when Chomsky and others pointed out that statistical ngram methods are incapable of modeling certain syntactic constraints such as agreement over long distances. It is possible that the Brown *et al.* approach is deeply flawed for many of the reasons that were discussed in the 1950s. Nevertheless, there is a growing community of researchers in corpus-based linguistics who are becoming convinced that the approach is worth pursuing because there is a very good chance that it will produce a number of lexical resources that could be of great value to their research.

Brown *et al.* begin by aligning the parallel texts at the sentence level. That is, they input a pair of texts as illustrated in Table 1, and output a sequence of aligned regions as illustrated in Table 2. 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. Table 2 illustrates a particularly hard paragraph; only two of the sentences line up one for one.

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.

We define the alignment task so that the order of the two texts must be preserved; crossed dependencies are not allowed. That is, if we match the i^{th} English sentence with the j^{th} French sentence, then we do not permit the next English sentence $i + 1$ to correspond to the previous French sentence $j - 1$.

Table 1: Input to Alignment Program

English	**French**
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Table 2: Output from Alignment Program

English	**French**
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

After sentences have been aligned, the second step is to identify correspondances 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. For example, the tenth and eleventh English words in the second region in Table 2 (i.e., the word *in* and the word *the* in the sentence *The higher turnover was largely due to an increase in the sales volume*) correspond to the twelfth French word (i.e., *du* in *La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes*). The identification of word level correspondances is the main topic of this paper.

We wish to distinguish the terms *alignment* and *correspondance*. The term *alignment* will be used when order constraints must be preserved and the term *correspondance* 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 correspondance 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 (at least for now).

Here is an example of our word correspondance 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 alignment:

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’ that/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, 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 precision and possibly low recall. For now, we are more concerned with errors of comission 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 correspondances have been identified, it is possible to estimate a probabilistic transfer dictionary. The entry for ‘‘the’’ found in (Brown *et al.*) is:

Table 3: An Entry in a Probabilistic Transfer Dictionary

English	**French**	**Prob(French | English)**
the	le	0.610
the	la	0.178
the	l’	0.083
the	les	0.023
the	ce	0.013
the	il	0.012
the	de	0.009
the	à	0.007
the	que	0.007

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 correspondances, we will not go into these other very interesting issues here.

2. ApplicationsBeyond

As mentioned above, MT is not the only motivation for sentence alignment and word correspondance. Computational linguists (e.g., Klavans and Tzoukermann, 1990) have recently become interested in bilingual concordances. Table 4, 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. At least in some 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 4: A Bilingual Concordance Based on Aligned Sentences

bank/ banque (''money'' sense)

65020: > it could also be a place where we would have a	bank of experts . SENT i know several people who a
20707373: ftre le lieu où se retrouverait une espèce de	banque d' experts . SENT je connais plusieurs pers
65988: f finance (mr . wilson) and the governor of the	bank of canada have frequently on behalf of the ca
20708447: es finances (m . wilson) et le gouverneur de la	banque du canada ont fréquemment utilisé au co
66551: reduced by over 800 per cent in one week through	bank action . SENT there was a haberdasher who wou
20709109: us de 800 p . 100 en une semaine à cause d' une	banque . SENT voilà un chemisier qui aurait appr

bank/ banc (''place'' sense)

4863261: h a forum . SENT such was the case in the georges	bank issue which was settled between canada and th
25799194: entre les états-unis et le canada à propos du	banc de george . SENT > c' est dans le but de ré
4866340: han i did . SENT he said the nose and tail of the	bank were surrendered by this government . SENT th
25802354: gouvernement avait cédé les extrémités du	banc . SENT en fait , lors des négociations de l

3. Using Word Correspondances Rather than Sentence Alignments

Most bilingual concordance programs such as ISSCO's BCP program mentioned in footnote 1 of (Warwick *et al*., 1990) and a similar program mentioned on page 20 of (Klavans and Tzoukermann, 1990) are based on aligned sentences rather than word correspondances. Table 4 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 correspondances.

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 one of its morphological variants). The word-based bilingual concordances (Table 5) show this fairly clearly. Only one of the first ten uses of *take* is translated with *prendre*. 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.''

Table 5: A Sample of Bilingual Concordance for ''take''

347: ise above the conventional political options and	take a new and broader view . SENT [english] S
376: ons politiques traditionnelles pour nous ouvrir à	de nouvelles perspectives . SENT [traduction] S
637: nslation] SENT > mr . speaker , as we	take further steps to promote equal op
685: ur le président , au fur et à mesure que nous	prendrons d' autres initiatives visant à favoris
1966: anel for the court challenges program . SENT i	take this opportunity to call upon the secretary
2074: rogramme de contestation judiciaire . SENT je	profite de l' occasion pour demander au secrétaire d
3744: > i would not mind if the government wanted to	take the $ 500,000 penalty out of clause 6 . SENT
4069: aquité au travail . SENT > cela me serait égal	que le gouvernement veuille retirer l' amende de
8206: oing what the law says and no one will be able to	take us to court " . SENT > i have a ques
8491: pas faire ce que la loi exige , et personne ne pourra	nous traîner devant les tribunaux " . SENT
10277: along with the amendment . SENT the vote was to	take place . SENT it is my understanding that ther
10810: ble avec le sous-amendement . SENT il devait y avoir	un vote . SENT je crois comprendre qu' il y a e
10731: rights because they know it is good business to	take advantage of the skills and abilities
11273: x de tirer parti des talents et des compétences	de l' ensemble de la main-d'oeuvre . SENT il est r
12333: an be traced back to the employer who failed to	take the necessary safety measures . SENT > the sa
13086: en souvent la faute de l' employeur qui n' a pas vu	à mettre en place les mesures de sécurité
12770: evious amendments , today we can say , and we	take great pride in this , that at the federal
13566: arieurs , nous sommes capables aujourd'hui de dire ,	et assez fièrement , qu' au niveau du gouver
16540: . speaker : | i declare the motion lost . SENT i	take it there is an agreement . SENT > | mr . dean
17615: ' examen de la motion no 31a . SENT serait-il possible	d' y remédier maintenant ? cela permettrait

After looking at more data, we find that only 23% of the uses of *take* are translated with a form of *prendre*. This figure is fairly consistent with IBM's estimate of 28% (Brown, personal communication). These numbers shows a nice contrast between a word like *take* and a more transparent word like *risk*, which is much more likely (61%) to be translated with one of its more salient translations (e.g., *risque* or *danger*).

It is fairly easy to make these estimates after word correspondances have been identified. We believe the word-based concordance provides a fairly useful tool for determining just how common various translations really are. The word-based concordance program is more appropriate than a sentence based program for this purpose. THis is especially true for a word like *take* because most users would find it difficult to enumerate the possible French translations for *take*. One might follow (Klavans and Tzoukermann, 1990) and turn to a bilingual dictionary to suggest some of the possibilities, but we suspect that currently available bilingual dictionaries would not cover the possibilities adequately for a complex word like *take*.

4. Two Possible Problems with the EM Algorithm

This paper is primarily concerned with the task of identifying word correspondances. 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 prohibative amount of memory. We observed that they limited the size 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* (e.g., *Hear, hear!* → *Bravo!*), though it isn't 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 faction 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 priviate discussions that our methods were coming up with radically 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 repliably 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 6 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 6: 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. In this case, the variance of ϕ^2 is estimated to be 2.5×10^{-5} (see section 6), 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

One might contrast *house*/*chambre* with a near miss such as *house*/*communes* (see Table 7).

Table 7: A Near Miss

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

Unfortunately, this pair is also highly significant ($t = 31$) 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 $Prob(x,y) = a/N$, $Prob(x) = (a + b)/N$, $Prob(y) = (a + c)/N$ and $N = a + b + c + d$. 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*.

$$I(house;chambre) = \log_2 \frac{31{,}950N}{(31{,}950 + 12{,}004)(31{,}950 + 4793)} = 4.1$$

$$I(house;communes) = \log_2 \frac{4974N}{(4974 + 38{,}980)(441 + 852{,}682)} = 4.2$$

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. In addition, 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 6 than in Table 7. ϕ^2 picks up this difference; Table 7 has a ϕ^2 of 0.098, significantly less than Table 6's ϕ^2 of 0.62.

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

Thus, we can very confidently say that *house* is more associated with *chambre* than with *communes*.

7. Two by Two by Two Contingency Tables

This is the method we currently use. However, a theoretically more attractive approach would make use of a two by two by two contingency table (see Table 8). The three dimensions correspond to the presence or absence of each of the three words *chambre*, *house* and *communes*. These cells can be determined by knowing the number of sentences (aligned regions) that contain one, two or three of these words: $freq(chambre) = 36{,}743$; $freq(house) = 43{,}954$; $freq(communes) = 5415$; $freq(chambre,house) = 31{,}950$; $freq(chambre,communes) = 5052$; $freq(house,communes) = 4974$; $freq(chambre,house,communes) = 4760$.

There are then six ways to project this 3-dimensional cube down to two dimensions. We can look at each of the six resulting two by two contingency tables. It turns out that all six have significant ϕ^2 values since all three of these words are so highly associated with each other. This example was selected because all of the associations are so strong that it is difficult to distinguish the desired pair (*house/ chambre*) from the near miss (*house/ communes*). Fortunately, the desired result can be obtained by comparing the –chambre case with the the –communes case. Table 8 shows that –communes has a much larger ϕ^2 (0.58) than –chambre (0.10), and this difference is very significant ($t = 79$) since the variances are quite small. We therefore conclude that *communes* is less associated with *house* than *chambre* is.

Table 8: Six projections of a two by two by two contingency table

	a	b	c	d	i
+ chambre	4760	292	27,190	4501	0.0074
– chambre	214	149	11,790	857,701	0.10
+ house	4760	27,190	214	11,790	0.034
– house	292	4501	149	857,701	0.040
+ communes	4760	292	214	149	0.10
– communes	27,190	4501	11,790	857,701	0.58

8. Calculation of Variances

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

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

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_{small} decreases to 0, which makes this equation unsuitable as an estimate of the variance. We calculate a variance for this case by assuming that $bc << ad$, which imples that $\phi^2 \approx 1 - (b + c)/a$. With this assuption, 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: $var(\phi^2) \approx \min(var_{small}, var_{large})$.

9. 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).

Fortunately, it is fairly easy to identify many candidate pairs after looking at a relatively small subset of the training corpus. The *house*/*chambre* pair, for example, is so frequent that one might suspect that there is a relationship after seeing just a few examples. After looking at just a handful of examples, we might already suspect that *house* and *chambre* deserve further investigation.

7: SENT %c.l. tabling of documents SENT %10xbf	house of commons SENT * date thursday , april 1
5: SENT %c.l. ṕeatitions SENT %10xbf	chambre des communes SENT * date le jeudi 17 avril
172: would like to bring to the attention of the	house that today , as hon . memb
179: ident , je voudrais porter`a l’ attention de la	chambre que nous ćeĺeabrons aujourd’hui , comme
936: d i want to join with all members of this	house in proclaiming this happy occasion . SEN
1004: ftre fiers , et je me joins`a tous les ďeputés	pour ćeĺebrer cette heureuse occasion . SEN
1474: party and join the other members of this	house in celebrating the anniversary of the procl
1579: ďemocratique et de me joindre aux autres ďeputés	en ćeĺebrant l’ anniversaire de la procl

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 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 (z, y). This procedure would take *house*/*chambre* but not *house*/*communes*. Repeat this operation with larger and larger samples of the training corpus. 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*.

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. For example, the following pairs were considered on iteration 3, but only the first one was selected, because the conservative select-best procedure will select no more than one pair with the same English word. There ought to be a way to take advantage of the obvious morphological generalization and select all of the morphologically related pairs at the the same time.

Pairs Considered on Iteration 3

i	English	French
0.19	accept	acceptons
0.19	accept	accepteront
0.16	accept	acceptera
0.15	accept	acceptent
0.11	accept	accepterait
0.085	accept	accepteraient
0.065	accept	acceptera-t-il
0.065	accept	accepterions
0.052	accept	accepte-t-il

A remarkably simple heuristic seemed to work fairly well. 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.

10. 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’ ´evaluer et de modifier des lois et des politiques en vigueur afin qu’ elles correspondent`a une interpŕetation ǵeńereuse de la charte .

The matching procedure attempts to match English and French words using 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 labelled *slope* indicates the difference between the previous two *j* values.)

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	eˆavaluer	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	aˆg	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	interpreˆatation	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 score. A sequence of j values is scored with

$$\underset{slope_j}{product}\; prob(match \mid slope_j)$$

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 much larger or much 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 $\log\ prob(match)$ 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 procedcure 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 comission involved a function word, usually one surronded on both sides by words that could not be matched.

11. Conclusions

In the future we plan to line up the words in their sentence context and see if some of the remaining gaps can be filled in using word-order constraints. Consider, for example, the two phrases (below). The system ought to be able to figure out that *placed* and *imposeès* are translations of one another given that it knows all of the other word pairs in this example. Brown *et al.* use the EM algorithm to fill in these sorts of gaps. We may adopt this option, too, but we tend to prefer methods that avoid making hard decisions unless they can be confident, unlike maximum likelihood methods which always make some decision, even when there aren't any good decisions to be made. For now, we are less concerned with errors of omission than errors of commission.

Eng:
restrictions placed on the metropolitan police
Fr:
restrictions imposèes à la police mètropolitaine

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,'' *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp 169-176.
- 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 the ACL.
- 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.