Aligning the Scania Corpus

Erik F. Tjong Kim Sang
Department of Linguistics
University of Uppsala
erik.tjong@ling.uu.se

April 10, 1996

Abstract

The Scania corpus is a collection of truck manuals available in eight European languages. We have applied the *GC-align* program that has been presented in [GC93] for aligning the sentences in eight language versions of one document of this corpus marked up with SGML. The error distribution of the program was similar to the results reported by Gale and Curch. Nearly all the errors of the program were caused by two paragraph errors. After having removed those *GC-align* came very close to perfect alignment for all seven language pairs. The overall performance of the program supports the claim of Gale and Church that *GC-align* can be applied successfully to other European language pairs than English-French and English-German.

1 Introduction

The Scania corpus is a multilingual collection of truck maintenance manuals of the Swedish company Scania CV AB. The Department of Linguistics of the University of Uppsala in cooperation with Scania will use this corpus for developing translation support tools. The corpus was delivered to Uppsala in Framemaker format. It has now been converted to TEI-compliant SGML [TKS96]. The corpus contains documents in eight languages and has a total size of 1.6 million words.

At this moment the Swedish part of the corpus is being used for developing a controlled language, ScaniaSwedish, which will be used by Scania as the standard source language for the manuals [ASH96]. However, we also want to use the corpus for extracting translation equivalents. In order to do this we need to align all seven translated document versions with the version in the source language, Swedish.

There are different automatic sentence alignment techniques available: methods based on character n-grams [Chu93], sentence lengths [GC93] and approximated bilingual dictionaries [FC94]. To start with we have chosen the sentence length method of Gale and Church: the

language	files	words
Swedish	95	220248
Dutch	75	201289
English	79	222211
German	75	184588
Finnish	79	143381
French	74	234467
Italian	81	233791
Spanish	75	220631
multilingual	37	2659
total	670	1663265

Figure 1: The Scania corpus at March 18, 1996. The corpus consists of documents in eight languages and contains in total 1.6 million words.

align program hence called *GC-align*. Our prime reason for chosing this approach was availability. The core of *GC-align* has been made available to the research community in 1993.

The alignment method used by Gale and Curch relies on a statistical model for translated sentence lengths. The authors have applied this alignment technique for aligning English and French texts and for aligning English and German texts. For both language pairs they have used the same statistical model. We would like to know if this model can also be used for aligning the seven language pairs that are present in our corpus.

In this paper we will give a brief introduction to sentence alignment with the technique presented in [GC93]. We will apply this technique to a document in our corpus and we will present the results. We will show that the program generates reasonable results. After this we will discuss the cause for the errors and give some concluding remarks.

2 Sentence Alignment with *GC-align*

The sentence alignment method presented in [GC93] is based on the fact that there is a correlation between the length of a sentence and the length of its translation. The method consists of two steps. In the first step the text will be divided in so-called hard regions (typically paragraphs) and in the second step the sentences within each hard region will be aligned. The algorithm used in the first step is similar to the one used in the second step. We will will presume that paragraph alignment has been performed successfully and concentrate on sentence alignment.

Sentence alignment is not trivial. A translator may have removed, inserted, split or combined sentences. An example of this is shown is figure 2 where the first Swedish sentence has been split in two English sentences. The *GC-align* program is able to recognize the following sentence alignment categories: 1-1 (the most frequent one-to-one sentence translation), 1-0

Swedish	**English**
Vanligaste enheten är kg/dm^3 (vattens densitet är 1 kg/dm^3).	The most common unit is kg/dm^3. The density of water is 1 kg/dm^3.
Specifik vikt är en annan benämning	Specific gravity is another term for density.

Figure 2: A Swedish paragraph in document 000103 of the Scania corpus with its English translation. The first sentence in the source text corresponds with two sentences in the target text.

language	pages	paragraphs	sentences
Swedish	12	76	357
Dutch	12	77	366
English	12	75	364
German	12	75	367
Finnish	12	77	362
French	12	75	360
Italian	12	75	364
Spanish	12	77	361

Figure 3: The internal structure of the eight language versions of document 000103. All documents have the same number of pages but the number of paragraphs and the number of sentences are different.

(sentence deletion), 0-1 (insertion), 2-1 (combination), 1-2 (splitting) and 2-2 (combination of two source language sentences in two target language sentences). The other categories are considered to occur infrequently enough to be ignored.

GC-align computes the probability of every possible sentence alignment category within a hard region. Its basic assumption is that a sentence of n characters in the source document will lead to a sentence of m characters in the target document. Here m is a normally distributed variable with mean n and variance 6.8 (computed for the language pairs English-French and English-German). The probability of a certain alignment category is dependent on the frequency of the alignment categories as observed in some standard text and a sentence length distribution function based on *align*'s basic assumption of translated sentence length. After computing the probabilities of all alignment categories within a hard region, the program will use a dynamic programming algorithm for computing the optimal set of alignment categories [GC93].

3 Experiments with *GC-align*

We have used *GC-align* for aligning the seven translated versions of one document of the Scania corpus with the original Swedish document. The document versions contain about 360 sentences and about 76 paragraphs. (see figure 3). For *GC-align* it is necessary that

	Swedish-Dutch			Swedish-English			Swedish-Finnish			Swedish-French		
	N	errors	%	N	errors	%	N	errors	%	N	errors	%
0-1	6	6	0	2	2	0	4	4	0	6	6	0
1-0	2	2	0	0	-	-	0	-	-	2	2	0
1-1	350	21	94	350	5	99	356	6	98	353	26	93
1-2	2	0	100	3	0	100	1	0	100	0	-	-
2-1	1	1	0	0	-	-	0	-	-	1	0	100
3-1	0	-	-	1	1	0	0	-	-	0	-	-
1-5	1	1	0	1	1	0	0	-	-	0	-	-
total	362	31	91	357	9	97	361	10	97	362	34	91

	Swedish-German			Swedish-Italian			Swedish-Spanish			total		
	N	errors	%	N	errors	%	N	errors	%	N	errors	%
0-1	6	6	0	6	6	0	6	6	0	36	36	0
1-0	2	2	0	2	2	0	2	2	0	10	10	0
1-1	346	27	92	346	23	93	343	24	93	2444	132	95
1-2	4	1	75	2	1	50	1	0	100	13	2	85
2-1	2	0	100	3	0	100	5	1	80	12	2	83
3-1	0	-	-	0	-	-	0	-	-	1	1	0
1-5	1	1	0	1	1	0	1	1	0	5	5	0
total	361	37	90	360	33	91	358	34	91	2521	188	93

Figure 4: The alignment results for the language pairs present in document 000103. *GC-align* makes few errors for the alignment categories 1-1, 1-2 and 2-1 but fails to recognize any other category.

the texts are divided in hard regions. Gale and Church have suggested paragraphs as hard regions. However, because the language versions of this document have different numbers of paragraphs we have chosen page boundaries as hard region separators.

Apart from some minor changes the version of *GC-align* that we have used is exactly the same as published in [GC93]. We have made no changes to the language model of the program because we wanted to know if Gale and Church's statistical description for the English-French and English-German language pairs would also fit other European language pairs.

We have applied *GC-align* to the seven language pairs for document 000103 of our corpus and have checked the results manually. The results of the alignment process can be found in figure 4. We have achieved good results for the language pairs Swedish-English and Swedish-Finnish (97% correct alignments) and reasonable results for the other five language pairs (90-91%). Overall there were 188 errors for 2521 alignments (93% correct alignments).

The distribution of the errors confirms the results reported in [GC93]. The easiest align category for the program is obviously 1-1 followed by 2-1 and 1-2 (the 2-2 category was not present in this document). *GC-align* was unable to recognize any other alignment category than these three. Gale and Curch reported a 0% correct score for the alignment categories

1-0/0-1, 3-1/1-3 and 3-2/2-3 and we got the same figure for the categories 1-0, 0-1, 3-1 and 1-5. The cause for the result for the latter two is obvious: they do not fit in *GC-align*'s error model which only contains categories up to 2-2. However the first two categories need improvement.

4 Improving the results

[Dah95] reports an alignment success rate of 96% for the Swedish and English versions of a different document of the corpus. The alignment was performed by using different techniques and by using ASCII versions of the document. The current version of the Scania corpus has been marked up with TEI-compliant SGML. [TKS96] has claimed that the SGML markup would improve the quality of the alignment software output. *GC-align* was able to improve on the 96% score for the language pair Swedish-English but for five language pairs it only achieved a score of about 90%.

We were interested in what caused this difference so we have inspected *GC-align*'s errors more carefully. We have found that nearly all errors (180 of 188) were caused by two paragraph errors in the document: a duplicate paragraph which was present in all language versions except for Swedish and a misplaced paragraph which was present in all language versions except for Swedish, Finnish and English.

By getting rid of these paragraph errors with additional preprocessing, we should be able to achieve alignment scores which are close to 100%. Removing paragraph errors can be achieved by applying *GC-align* to paragraphs using page boundaries as hard region separators. We have attempted to use this approach but we were unable to achieve satisfactory results. One paragraph error required deletion of a paragraph which corresponds with the alignment category 1-0. This is one of the problematic alignment categories for *GC-align*. The other paragraph error required cross links and these have been excluded from the alignment model presented in [GC93].

However, applying *GC-align* to paragraphs can be used for signaling problematic cases (non-1-1-alignments). We have used the program exactly in this way and have used its pointers to problematic paragraphs for correcting the paragraph order manually. After this we have repeated the sentence alignment process. In order to be able to compare the alignment results with the previous attempt we have used page boundaries as hard region boundaries again. The results can be found in figure 5.

In the previous section we have made the remark that 180 of the 188 errors were caused by the two paragraph errors in the language versions of the document. All these errors disappear when the paragraphs have been aligned correctly. The eight errors that remained are caused by alignment categories which were not present in *GC-align*'s error model (3-1 and 1-5). The program achieved scores of over 99% correct alignments for all language pairs and aligns two language pairs, Swedish-Finnish and Swedish-French perfectly.

	Swedish-Dutch			Swedish-English			Swedish-Finnish			Swedish-French		
	N	errors	%	N	errors	%	N	errors	%	N	errors	%
0-1	0	-	-	0	-	-	0	-	0	0	-	-
1-0	0	-	-	0	-	-	0	-	-	0	-	-
1-1	351	0	100	350	2	99.4	356	0	100	354	0	100
1-2	3	0	100	3	0	100	1	0	100	1	0	100
2-1	1	0	100	0	-	-	0	-	-	1	0	100
3-1	0	-	-	1	1	0	0	-	-	0	-	-
1-5	1	1	0	1	1	0	0	-	-	0	-	-
total	356	1	99.7	355	3	99.2	357	0	100	356	0	100

	Swedish-German			Swedish-Italian			Swedish-Spanish			total		
	N	errors	%	N	errors	%	N	errors	%	N	errors	%
0-1	0	-	-	0	-	-	0	-	-	0	-	-
1-0	0	-	-	0	-	-	0	-	-	0	-	-
1-1	347	0	100	347	0	100	344	0	100	2449	2	99.9
1-2	5	0	100	3	0	100	2	0	100	18	0	100
2-1	2	0	100	3	0	100	5	0	100	12	0	100
3-1	0	-	-	0	-	-	0	-	-	1	1	0
1-5	1	1	0	1	1	0	1	1	0	5	5	0
total	355	1	99.7	354	1	99.7	352	1	99.7	2485	8	99.7

Figure 5: The alignment results for the language pairs present in document 000103 after manually aligning the paragraphs. *GC-align* achieves scores above 99% correct alignments for all language pairs and aligns Swedish-Finnish and Swedish-French perfectly.

5 Concluding remarks

We have applied the *GC-align* program which was presented in [GC93] for aligning a document of the Scania corpus. This required aligning the language pairs Swedish-Dutch, Swedish-English, Swedish-Finnish, Swedish-French, Swedish-German, Swedish-Italian and Swedish-Spanish. In our first attempt *GC-align* has achieved an overall score of 93% correct alignments. The distribution of the errors was similar to the one reported in [GC93]: the program has few problems with the alignment categories 1-1, 2-1 and 1-2 but fails to recognize any other category.

We have argued that the majority of the errors was caused by two paragraph errors in the document. These errors could not be removed by applying *GC-align* to paragraphs. One of them required a paragraph deletion and the other required cross links. We have removed the paragraph errors manually and after that applied the alignment program to the seven language pairs again. It achieved scores of over 99% correct sentence alignments for all language pairs and aligned Swedish-Finnish and Swedish-French perfectly.

[Dah95] reported 96% correct alignments for the Swedish and English ASCII versions of a different document of our corpus. The performance of *GC-align* on SGML marked up documents supports the claim of [TKS96] that marking up documents will improve the performance of the alignment software.

[GC93] have used the same statistical model for aligning English and French texts as for aligning English and German texts. They have assumed that this model can also be used for aligning other European language pairs. We have used the same statistical model for aligning seven different language pairs. The results we have presented here support Gale and Church's assumption that European language pairs are relatively independent when it comes to describing them with a sentence length based alignment model.

The reported results of *GC-align* applied to our corpus were much better than we had expected in advance. We believe that the prime reason for this is the lack of complexity of our data. After removing the paragraph errors, the document that we considered required 98.5% 1-1 alignments compared with 88.7% 1-1 required alignments in the texts used in [GC93]. The SGML markup of the document was helpful but we do not consider it to be the most important cause for our results. Actually we discovered that the markup, which had been performed automatically, contained some sentence boundary errors. Most of these errors were correctly handled by *GC-align*.

In the near future we will apply *GC-align* for aligning the sentences of all language versions of the documents in our corpus. Our experiments with aligning one document indicate that fixing the paragraphs errors in our corpus will require more work than we had expected. Another problem might be performing the final manual checks of the results. We expect to be able to minimize the amount of work involved there by applying a different alignment technique in parallel with *GC-align* and then only check those parts of the documents that were treated differently by the two methods.

References

[ASH96] Ingrid Almqvist and Anna Sagvall-Hein. Defining scaniaswedish - a controlled language for truck maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications*, pages 159–165, 1996.

[Chu93] Kenneth Church. Char align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8, 1993.

[Dah95] Bengt Dahlqvist. *Aligning the Chassis Text Document of the Scania Corpus.* Technical Report, Department of Linguistics, University of Uppsala (in Swedish), 1995.

[FC94] Pascale Fung and Kenneth Church. K-vec: A new approach for aligning parallel texts. In *Proceedings COLING-94*, 1994.

[GC93] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1), 1993.

[TKS96] Erik F. Tjong Kim Sang. *Converting the Scania Framemaker Documents to SGML.* Internal report, Department of Linguistics, University of Uppsala, 1996.