HLT-NAACL 2003 Workshop: Building and Using Parallel Texts
Data Driven Machine Translation and Beyond , pp. 31-35
Edmonton, May-June 2003

Phrase-based Evaluation of Word-to-Word Alignments

Michael Carl and Sisay Fissaha
Institut für Angewandte Informationsforschung
66111 Saarbrücken, Germany
{carl;sisay}@iai.uni-sb.de

Abstract

We evaluate the English—French word alignment data of the shared tasks from a phrase alignment perspective. We discuss peculiarities of the submitted data and the test data. We show that phrase-based evaluation is closely related to word-based evaluation. We show examples of phrases which are easy to align and also phrases which are difficult to align.

1 Introduction

We describe a phrase-based evaluation of the 16 English-French alignment submissions for the shared task on Parallel Texts. The task was to indicate which word token in an English alignment sample corresponds to which word token in the French alignment sample. Two types of submission were permitted: for restricted submissions were allowed a "sentence" aligned segment of the Canadian Hansards to train the systems while unrestricted submission would be allowed to use additional resources. The performance of the systems was compared for a set of 447 English—French hand-aligned test samples which were also taken from the Canadian Hansards.

Five institutes participated in the English—French alignment task, submitting a total of 16 sets of alignment data. To evaluate the submitted data, we extracted bilingual phrase dictionaries from the word-alignment data. The extracted dictionaries of the submitted data were compared with the extracted dictionary of the test data.

We first discuss word-to-word and phrase-to-phrase alignment format. We present two different methods for extracting bilingual dictionaries from the word alignment data: a minimal dictionary contains the least number of unambiguous phrase-to-phrase translations while an exhaustive dictionary contains all possible unambiguous translations. We examine the test data (i.e. the "golden standard") and the submitted alignment data. We discuss their peculiarities and give examples of phrases easy and difficult to align.

2 Types of Alignment

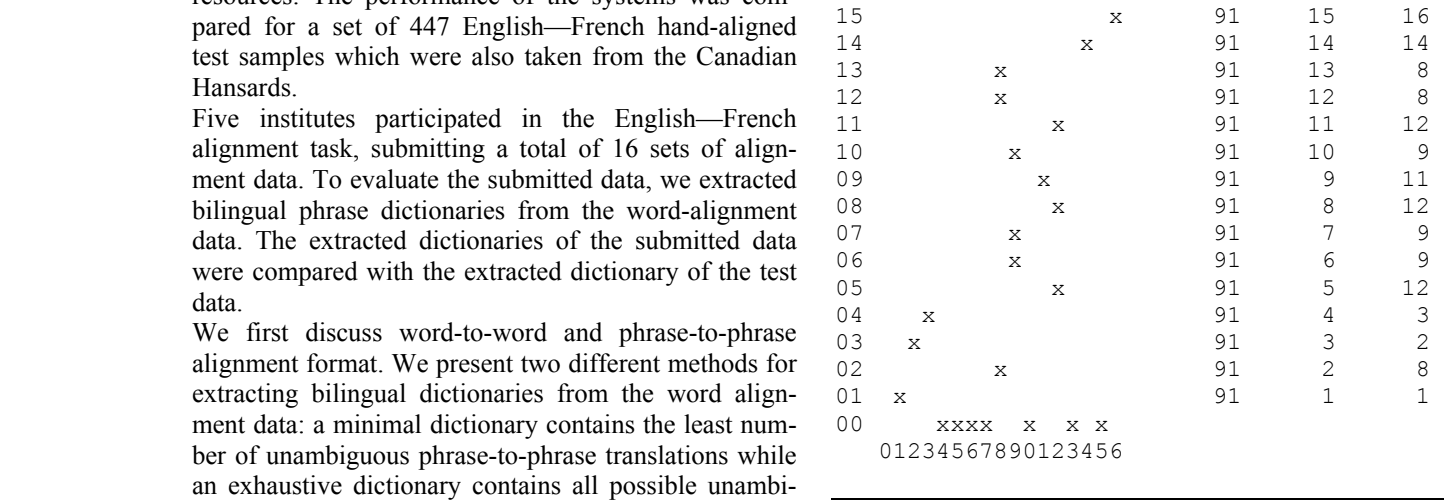
The test set consists of 447 alignment samples from the Canadian Hansards which were pre-tokenized. A three-tuple containing the alignment number, an English word offset and a French word offset would indicate an exact word-to-word translation[1]. The submitted data was supposed to comply with this word-to-word alignment format. In example 1 the English sentence has 15 tokens while the French sentence has 16 tokens. Example 1 shows the word-to-word alignment data of sample 91 for submission 12 and a plot of the data.

Example 1: Alignment sample 91:

```
English (vertical):
i was not asking for a detailed explana-
tion as to what he was doing .

French (horizontal):
je ne lui ai pas demandé de me fournir de
telles explications sur ces activités .
```

Plot and word alignment data for submission 12:



[1] There was also an optional slot to indicate whether this alignment would be [S]ure or [P]robable. We ignore this information in our evaluation.

2.1 Word-to-word alignment

There are two underlying assumptions in word-to-word alignment:

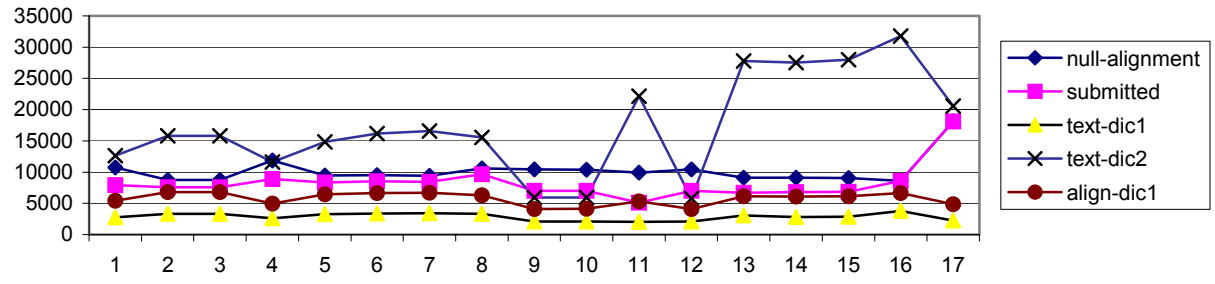
(i) each word token on the English side can have any number of word correspondences -- including zero -- on the French side and vice versa. Word alignments may have crossing and ambiguous branches. For instance in example 1, the French word "me" on position 8 has the translations "was", and "he", while "ai" has no connection to the English side.

(ii) words (English or French) for which no alignments are given in the submitted data are assigned a null-alignment.

Example 1 has 22 word alignment points, where the evaluators inserted 7 null-alignments. In some cases (i.e. submission 11) this insertion accounts for almost 50% of the alignment data. In figure 1, "null-alignment" plots the union of the submitted alignment data and the inserted null-alignments. Null-alignments were not added to submission 16 as it provides alignment information for every word. The last data point on the x-axis (i.e. 17) represents the test data.

As outlined in Melamed (1998), a sequence of words which translates in a non-compositional fashion into a target sequence is exhaustively linked (see example 2).

Figure1: Number of word alignment points and size of extracted dictionaries



2.2 Phrase-to-phrase alignment

Phrase-to-phrase alignment is represented by intervals indicating the starting and ending words of the phrases. In phrase-to-phrase alignment:

(i) a sequence of English word tokens (i.e. a phrase) are mutually linked with sequences of French word tokens (i.e. a French phrase)[2].

(ii) an English phrase may only be *unambiguously* linked to exactly one French phrase and vice versa.

Phrase-to-phrase alignments can be nested. For instance, the shorter English—French phrase translation `9-9 <-> 11-11` is included in the longer phrase translation `5-11 <-> 9-12`:

```
5-11 9-12:   for a detailed explanation
      as to what
      <-> fournir de telles explications
9-9 11-11:   as <-> telles
```

In this way structural information can be stored. On the other hand, we do not allow ambiguous phrase alignments as e.g.:

```
 8-8   12-12  explanation <-> explications
11-11  12-12        what <-> explications
```

When extracting phrase-to-phrase translations from the word-to-word alignment data we include a sufficient context which disambiguates the phrases. Given the word alignment data in example 1, the minimum context required to disambiguate the French word "explications" is the phrase `5-11 <-> 9-12`.

From the word alignment data we generate bilingual dictionaries in two different ways: a *minimal dictionary* contains only the shortest unambiguous phrase-to-phrase translations. For instance, from the alignment data in example 1, the following 8 entries are generated as a minimal dictionary:[3]

En	Fr
1-1	1-1
2-13	2-12
3-3	2-2
4-4	3-3
5-11	9-12
9-9	11-11
14-14	14-14
15-15	16-16

[2] We do not use the term "phrase" here in its linguistic sense: a phrase in this paper can be any sequence of words, even if they are not a linguistic constituent.

[3] As shorthand notation we use here the offset numbers. In the generated dictionary, we have extracted the sequences of words instead of the offset numbers.

In an *exhaustive dictionary* all possible unambiguous phrase translations are extracted. An exhaustive dictionary is a superset of the minimal dictionary. For example 1, seven additional entries are generated:

```
En     Fr
1-13   1-12
1-14   1-14
1-15   1-16
2-14   2-14
2-15   2-16
3-4    2-3
14-15  14-16
```

Note that these additional phrase translations can be compositionally generated with the minimal dictionary. To evaluate the word alignment data through phrasal alignments, we generated three types of dictionaries for all 16 submissions and the test data:

(i) an alignment-based minimal dictionary, align-dic1; actually 447 small dictionaries for each sample alignment.

(ii) a text-based minimal dictionary (text-dic1)which is the union of the align-dic1.

(iii) an exhaustive text-based dictionary (text-dic2) which is the union of exhaustive alignment dictionaries.

As can be seen from figure 1, the size of the exhaustive dictionary (text-dic2) is in most cases much bigger than those of the minimal dictionaries align-dic1 and text-dic1. The reason is due to the way the data has been aligned.

3 The word alignment data

In this section we show that the test alignment data is structurally different from the submitted data. The hand aligned test data reflects the phrasal nature of the alignments, while the submissions are to a greater extent compositional.

The test data (see set 17 in figure 1) has about twice to three times as many word-alignment points than the submissions. While this often leads to high precision and lower recall for word alignment, the reverse is true for the extracted phrasal dictionaries (also figure 3). The test alignment data of sample 91 contains 68 word-to-word alignment points shown in example 2; about four times the average number of word alignment points for this sample. Comparing example 2 with the submitted data of submission 16 (example 3) brings to light the phrasal nature of the test set.

Extracting a minimal phrase dictionary from test set word alignment data in example 2 produces the following three entries

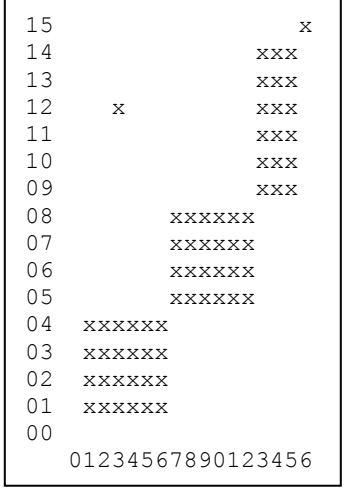
```
1-14   1-15
5-8    7-12
15-15  16-16
```

The following additional entry is generated in the exhaustive dictionary: `1-15 <-> 1-16.` Note that more word alignments could have been possible here, for instance:

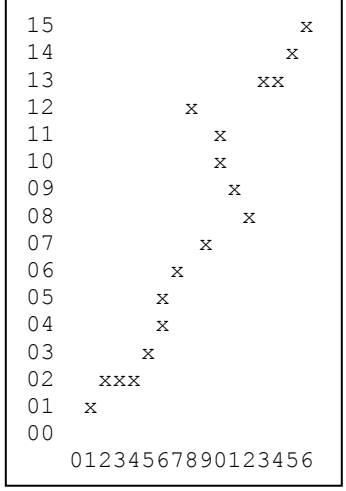
```
i <-> je
asking <-> demandé
explanation <-> explications
not <-> ne , pas
```

Despite the existence of some fine-grained word-to-word correspondences in the test data, human aligners tend to mark phrasal translations. In contrast to the phrasal nature of the test alignments, most submissions show a more compositional alignment structure. For instance, alignment data of sample 91 for submission 16 has 18 word-to-word alignment points (example 3). When the alignments are more compositional, more coherent phrasal translations can be extracted. Thus, the minimal phrase dictionary extracted from example 3

Example 2: sample 91 of test set:



Example 3: sample 91 of submission 16

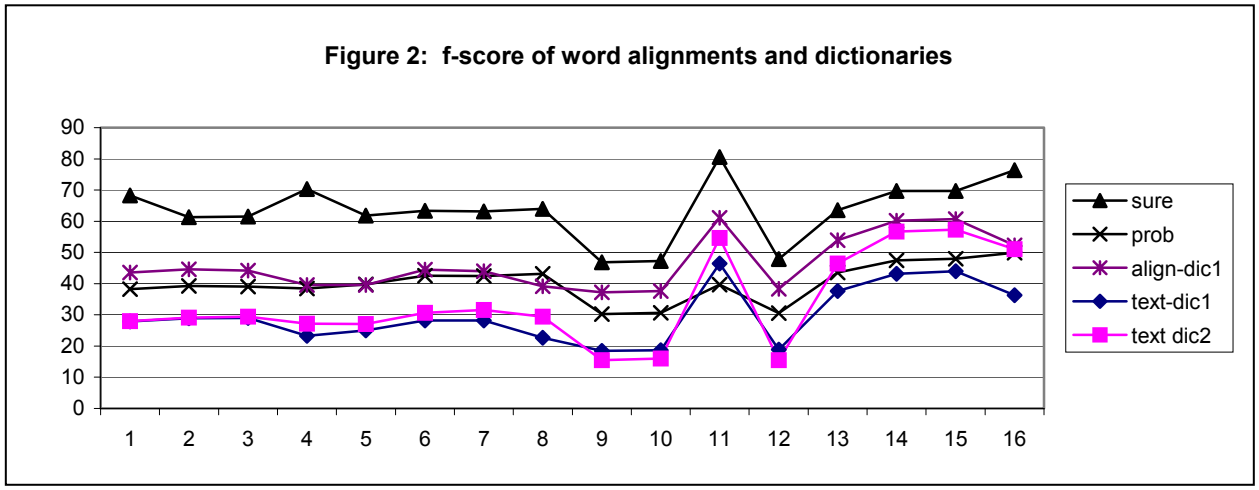


contains 13 entries, while the exhaustive dictionary contains 54 entries. Therefore, we expect that mapping the phrase dictionaries on the test dictionaries would result in high recall and low precision while mapping the submitted word alignment data on the test data would result in high precision and low recall.

4 Phrase-based Evaluation

Figure 2 shows the correlation of the f-score of the three extracted dictionaries and the word alignment data (both sure and probable). For each submission the f-score was calculated as 2*precision*recall/precision+recall. The curves for the alignment-based dictionaries (align-dic1) show the mean f-score computed over all 447 samples. Roughly all submissions show a similar pattern for word-to-word alignment and phrase dictionaries.

We wanted to see what factors influence precision and recall. A correlation between the length of the sample alignment and its average recall and precision is shown



in figure 3. As one would have expected, the graph shows a tendency that shorter samples are easier to align (higher precision and recall) than longer samples. However, there is higher variation among shorter alignments than among longer sample alignments which indicates unpredictability of shorter samples. As an example consider alignment sample 7 (length 3):

```
hear, hear ! <-> bravo !
```

The extracted minimal test dictionary contains the two entries:

```
hear,hear <-> bravo
! <->!
```

While most of the minimal dictionaries extracted from the submitted data contain the entries:

```
hear <-> bravo
! <-> !
```

This leads to a value of 50 for recall and precision for both word and phrase alignments. The average recall and precision of sample 7 (length 3) is 53,1 and 57,8.

Figure 3 also shows that PD-recall (phrasal dictionary) is higher than PD-precision as the samples become longer. For example, sample 91 (length 15,5) has PD-recall and PD-precision values of 65, and 50,2 respectively. For word-to-word evaluation, however, WD-precision is higher than WD-recall. For sample 91 (length 15,5) the WD-recall and WD-precision values are 18,03 and 53,86 respectively.

Next we wanted to see which parts in the sample alignments would be easy and which parts would be difficult to align. We assume that correct translations which appear in all submissions would be easy to find while translations which occur only in the test set but in none of the submissions would be difficult to find. Finally, the same noisy translations produced by all submissions would indicate mistakes in the test data. The cardinality of these sets is shown in the table below.

Intersection of	text-dic1	text-dic2
correct	150	434
missing	837	1949
noise	11	22

There were 150 one-word entries in the intersection of the correct translations contained in all 16 dictionaries text-dic1. These translations include transfer rules which are easy to discover such as numbers, function words, pronouns, frequent content words and also domain specific translations:

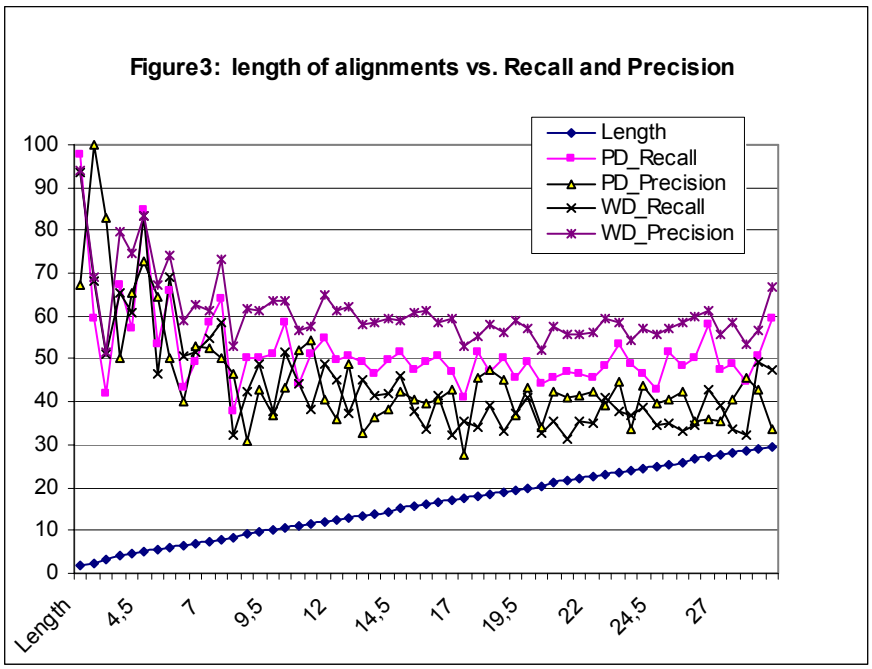
1) pronouns

```
he <-> il
it <-> il
there <-> il
```

2) frequent content words

```
women <-> femmes
work <-> travaillent
compulsory <--> obligatoire
say <-> dire
says <-> dit
```

3) function words



```
such <-> tel
to <-> de
to <-> pour
```

5) Text typical translations:

```
House <-> Chambre
```

The set of translation equivalences missing in all submissions was much larger. There were only the following five one-word equivalences:

1) on-word translations:

```
and <-> puisque
balance <-> niveau
do <-> fait
per <-> le
very <-> fondamentalement
```

Most of the missing entries were multi-word translations, such as idioms, compound words etc.

1) idiomatic expressions

```
A buck is a buck is a buck <-> une
piastre est toujours une piastre
thank you very much <-> je vous re-
mercie
```

2) compound:

```
Canadian Wheat Board <->
  Commission canadienne de le blé
```

3) complex prepositions

```
as for <-> en ce qui concerne
```

4) complex verbs and negation

```
does not like <-> ne aime pas
will be <-> feront
```

5) adverbs and adjective phrases

```
previous <-> qui me a précédé
a good thing <-> intéressant
```

6) unresolved pronouns

```
the government <-> il
```

There were also 11 noisy entries which occurred in all generated submissions dictionaries but not in the test data dictionary. The obvious explanation for this is, again, the phrasal nature of the test data: single word translations would be hidden in phrase translations and not extracted as separated word translations:

```
before <-> avant
believe <-> crois
days <-> jours
every <-> chacune
facilities <-> installations
jobs <-> emploi
positive <-> positifs
public <-> public
representations <-> instances
why <-> comment
will <-> servira
```

5 Submissions

This section lists the origin of the submitted data. A more detailed description can be found in the system description contained in these proceedings.

1 BiBr.EF.7
Limited Resources 7. intersection of 1 & 3
2 BiBr.EF.1
Limited Resources 1. Baseline of Bi-lingual Bracketing
3 BiBr.EF.2
Unlimited Resources 2. Baseline of Bi-lingual Bracketing + POS (Brill's POS tagger for English only)
4 BiBr.EF.8
Unlimited Resources 8. intersection of 3 & 6
5 BiBr.EF.3
Unlimited Resources 3. Baseline of Bi-lingual Bracketing + POS (Brill's POS tagger for English only) + English_Chunker.
6 BiBr.EF.4
Limited Resources 4. reverse direction of (1)
7 BiBr.EF.5
Unlimited Resources 4. reverse direction of (2)
8 BiBr.EF.6
Unlimited Resources 4. reverse direction of (3)
9 data withdrawn
10 UMD.EF.
Limited Resources Trained on House and Senate Data
11 ProAlign.EF.1
Unlimited Resources ProAlign uses the cohesion between the source and target languages to constrain the search for the most probable alignment (based on a novel probability model). The extra resources include: An English parser A distributional similarity database for English words.
12 data withdrawn
13 XRCE.Base.EF.1
Limited Resources GIZA++ with English and French lemmatizer (no trinity lexicon)
14 XRCE.Nolem.EF.2
Limited Resources GIZA++ only (no lemmatizer, no trinity lexicon), Corpus used: Quarter
15 XRCE.Nolem.EF.3
Limited Resources GIZA++ only (no lemmatizer, no trinity lexicon), Corpus used: Half
16 ralign.EF.1
Limited Resources Recursive parallel segmentation of texts; scoring based on IBM-2
17 test data (golden standard)

6 References

Ulrich Germann, editor (2001). Aligned Hansards of the 36th Parliamentof Canada. http://www.isi.edu/natural-language/download/hansard/index.html

I. Dan Melamed (1998). Annotation Style Guide for the Blinker Project, IRCS Technical Report #98-06, http://www.cs.nyu.edu/~melamed/ftp/papers/styleguide.ps.gz

Franz Josef Och, Hermann Ney (2000). A Comparison of Alignment Models for Statistical Machine Translation.. COLING 2000..http://www-i6.informatik.rwth-aachen.de/Colleagues/och/COLING00.ps