Comparing the Sentence Alignment Yield from Two News Corpora Using a Dictionary-Based Alignment System

Stephen Nightingale
ATR, Kyoto, JAPAN
Stephen.Nightingale@atr.co.jp

Hideki Tanaka
ATR, Kyoto, JAPAN
Hideki.Tanaka@atr.co.jp

Abstract

Corpus-based MT requires the input of large sentence aligned bilingual corpora, but these are hard to find for Japanese. Bilingual news corpora seem to offer a useful resource for Machine Translation, but their quality is variable. Sentence alignments produced by filtering literal word translations from the NHK corpus yield disappointing results, though correlating NP translations performs better. Using this method gets even better results from the Nikkei corpus. This paper reports sentence alignment results from 2 corpora, in a 2-pass dictionary based alignment system.

1 Credits

The research for this paper was done under a grant from the Telecommunications Advanced Organization (TAO) of Japan.

2 Introduction

Large, sentence-aligned bilingual corpora suitable for input to corpus-based Machine Translation are not a natural product of the translator's art. They are specifically crafted and selected for a high degree of word and phrase alignment. There is a healthy body of literature attesting to this fact, with simple length-based methods (Gale and Church, 1991), dictionary-based alignments such as (Haruno and Yamazaki, 1996), and more sophisticated bags of words approaches, suitable for noisy corpora, such as Kvec (Fung and Church, 1993). All these methods are intended to produce aligned corpora which can be input to Statistical Machine Translation (SMT) training. With noisy corpora such as news article translation pairs, some are more free translations, some summarizations. Occasionally too, there are close literal translations. The NHK and Nikkei are two different news corpora which have been reported in ICSLP (Tanaka et al, 2002). Attempts to sentence-align the NHK corpus using numerically derived single word correspondences have been disappointing (Nightingale and Tanaka, 2002). We therefore redeveloped the alignment method to give the dictionary check primacy over the statistical comparator, and modified its scope to find Noun Phrase translations first (Nightingale and Tanaka, 2003). The method is applied in a second pass to find sentence alignments, taking account of the NP alignments already discovered, but still the proportion of incommensurable sentence pairs is very high. More recently, a part of the Nikkei corpus became available, and we applied the same method so that: (a) the two corpora could be compared for their yield, and (b) so that we could realize a larger aligned corpus for MT investigations. The Nikkei corpus yields better results using the same method. Both the alignment method and the results are described in this paper. After some brief description of the corpora in Section 3, the alignment method is shown in Section 4, in three parts, with the Many-to-1 tokenization, translation candidate generation, and the dictionary filtering separately described. The alignment results for NPs and for sentences are given in Section 5, with some concluding remarks in 6.

3 The Corpora

The NHK and Nikkei are news article corpora, having typically 4-10 sentences in each article, average lengths 20+ words, with no guarantee of 1-for-1 sentence alignment. NHK comprises 41.1K article pairs covering 1995-2000, which we split into two parts according to news genre, with 29.6K and 11.5K; Nikkei has 1929 article pairs from July 2000. To extract and correlate word groups such as Noun Phrases, we parse the corpora using (Charniak, 2001) for English, and CaboCha (Kudoh and Matsumoto, 2000) for Japanese. A Charniak parse is shown in the example below.

(S1 (S (NP (DT the) (NN team))
(VP (MD will)
(ADVP (RB also))
(VP (VB compare)
(NP (NN risk) (NN capital) (NNS allocations))
(PP (IN with) (NP (NN profitability)))))

We find the phrasal structures in the sentence using a lexicalization program which extracts the words or POS's according to parameter, as in this list:
(a) NP = the team
(b) NP = risk capital allocations
(c) NP = profitability
(d) PP = with profitability
(e) VP = compare NP
(f) VP = will ADVP VP
(g) S = the team will also compare risk capital allocations with profitability

The sentence contains 3 NPs extracted fully lexicalized, a PP, 2 VPs extracted partially lexicalized and a sentence, S fully lexicalized. The corpora contain many thousands of NPs, VPs and sentences which can be thus extracted. A separate program lexicalizes the dependency chunks of Japanese. In the case of investigating NP alignments, all other structures are filtered out. The sentence counts and unique NP counts are given in Table 1. The Nikkei shows a much higher proportion of Unique NPs than the NHK, with similar numbers of NPs (38K E/55K J) from a much smaller corpus.

Corpus	**Articles**	**Sents**	**Unq NPs**
NHK1	29.6K	218K/180K	57K/32K
NHK2	11.5K	86K/81K	37K/28K
Nikkei	1929	21K/15K	38K/55K

Table 1: NHK and Nikkei Corpus Statistics

English Compound	**Frequency**
civilians	110
chemical weapons	110
foreign ministry spokesman	109
Japanese Compound	**Frequency**
合意 文書	205
ポル・ポト 派	204
青木 官房 長官	204

Table 2: Noun Phrases

4 Extracting Alignments

Investigations of the EGYPT tools (Al Onaizan et al, 2000; Och and Ney, 2000) for SMT show that the surface details of the two languages are abstracted away from the problem of identifying translation candidates (or token alignments). EGYPT's Whittle corpus pre-processor tokenizes the vocabulary in source and target corpora and maps words to tokens 1-to-1. An example of input is given below, with a sentence from the Japanese input file in 1(a) and its translation from the English input file in 1(b). Whittle tokenizes the vocabulary and maps words to tokens as in (2) (a) and (b), where each unique word form associates with a single value.

1(a) 冷やし た ミネラルウォーター の 小 ビン を 持っ てき てください 。
1(b) bring me one small bottle of chilled mineral water please .
2(a) 5899 14 1303 6 1383 4031 8 97 143 16 2
2(b) 195 18 39 231 547 24 3857 1101 169 9 2

EGYPT's Giza, the SMT training tool engine takes these token vector pairs and performs the IBM Model 1-5 transformations (Brown, et al, 1993) to generate a translation model. Because of the abstraction, Giza only knows about these token vectors and not about words and sentences. It is perfectly feasible therefore to map more complex word groups onto single tokens, in this instance parsed phrasal structures. Although we ultimately departed from EGYPT, this was one of the inspirations for our modularization discussed below. Because we extracted phrasal chunks from the corpus, these are tokenized with many words mapped to one token (4.1). Candidate generation proceeds with these token vector pairs (4.2), followed by dictionary filtering (4.3).

4.1 Flexible Tokenization

The tokenizer takes as input pairs of word strings, which may be phrases or sentences, articles or paragraphs: the surface form can be identical to that taken by Whittle. These need not be complete sentences. The primary sources are filtered for salient features such as Noun Phrases, discarding verbal phenomena and function words. The tokenizer also takes as input lists of word groups, such as NPs, to be identified in the source strings, and mapped onto single tokens. Examples (not translation pairs) of such NP lists are given in Table 2, with determiners and postpositions pruned. Nouns occurring singly (e.g *civilians*) are not excluded.

The strategy is to first tokenize individual words, and separately tokenize the word group lists, then map words to tokens 1-for-1, to create token vectors. In a second pass, the token sequences from the word group lists are substituted into the token vectors, thus compressing them. As example, the sentence pair in 1(a) and (b) can be filtered for Noun Phrases, yielding the result in 3 (a) and (b), and mapped onto tokens in 4(a) and (b). The tokenized word groups are then remapped to give the tokenized NP sequences in 5(a) and (b), where for example the string "one small bottle" maps to the unique token 9991.

3(a) (冷やし た ミネラルウォーター) (小 ビン)
3(b) (me) (one small bottle) (chilled mineral water)
4(a) (5899 14 1303) (1383 4031)
4(b) (18) (39 231 547) (3857 1101 169)
5(a) 9990 9999
5(b) 18 9991 9992

Following translation candidate generation we expect to see the pairings: 小_ビン = *one_small_bottle*, and 冷やし_た_ミネラルウォーター = *chilled_mineral_water* identified as translation pairs.

4.2 Translation Candidate Generation

Japanese	English
インターナショナル スクール	foreign schools
外国 人 学校	foreign schools
制度 改正	foreign schools
教育	foreign schools
要件	foreign schools

Table 3: Raw Translation Candidates

Previous experiments (Nightingale and Tanaka, 2003) compared the results of translation candidate generation and statistical filtering using various numeric comparators, then compared the filtered results against dictionary translations. For example applying a Mutual Information value and selecting the higher values as plausible translations; or applying the EM algorithm over word counts and selecting the higher probabilities as plausible translations. When we compare these against exhaustive generation/selection by dictionary matching (Section 4.3), we find that the exhaustive method yields more plausible translation candidates. For this exercise we use the exhaustive method, generating M * N token pairs over all sentence pairs. This yields 6.4M candidates for NHK1, 2M for NHK2 and 342K for Nikkei. This generation creates candidates which are token pairs. These are relexicalized to produce entries such as those in Table 3, with no statistical comparator. There are many inappropriate candidates (Nos. 3, 4, 5 and more) and a few appropriate (Nos. 1, 2). Selection is performed entirely by the dictionary filter, next.

4.3 The Dictionary Filter

When the volume of good translation results depends on the magnitude and word sense variation of the dictionary it behooves you to get a big dictionary. We use the combination of EDICT, ENAMDICT and EIJIRO with a total of 1.2M entries. ENAMDICT with 200K names is particularly important for matching news articles, which contain many such references. The method of accounting for dictionary equivalence is fully described in (Nightingale and Tanaka, 2003). Briefly, we count word matches in translation candidate pairs and for each unique English NP take the pair with the highest number of matches. In Table 3, the second entry with 2 overlaps is accepted (foreign schools = 外国 人 学校 [foreign person school]). Plurals are dealt with by lemmatization heuristics, so schools=school. When checking for sentence alignment we add a prior check of the known NP pairs to yield a total word+phrase match count.

5 Results

First Pass: NP Alignments: In NHK1 with 218K English and 180K Japanese sentences there are 57K and 32K, NPs. The alignment finds 24504 unique translation candidates, or 43% of the English NPs. Results from the second part of the NHK corpus are in the same range, with 12.2K found from 37K or 33%. The results of processing the Nikkei corpus are significantly better. Of 38K English NPs the extraction process yields 24.3K or 64%. Training to establish verb frame translations was less successful since our corpora generally have a low literal translation rate: because Verbs are not always translated as Verbs, and actual verb translations used in practice differ from forms which can be found in the dictionary.

Second Pass: Sentence Alignments: Aligned sentence pairs are discovered by lexicalizing sentential clauses extracted from the Charniak and Cabocha parses. From the Nikkei corpus with 21K/15K English and Japanese sentence pairs, we find 2825 pairs containing 6 or more matching word or phrases. An example with 10 matches is given in Table 4. 2825/21K gives us about a 13.5% recovery rate. This means that no adequate literal translation can be found among the Japanese sentences for 18K of the English sentences. There are a very few existing Japanese English corpora of around 150K sentence pairs: the ATR Basic Travel Expres-

Counts	word overlap=10, jlength=52, elength=36
Japanese	日本 貿易 振興 会 （ ジェトロ ） が 日本 市場 の 開放 状況 を 調べ た 「 対 日 アクセス 実態 調査 」 に よる と 海外 メーカー が 日本 に 進出 する 場合 の 初期 投資 コスト は 欧米 の 四 ー 十一 倍 に 達し て いる こと が 分かっ た
English	it costs foreign companies between four and 11 times more to set up manufacturing operations in japan than in europe and the u s according to a survey by the japan external trade organization jetro .
Matches	日本 =japan (japan) 貿易 =trade (trade) ジェトロ =external (japanese external trade organization) 対 =set (set) 調査 =survey (survey) 海外 =foreign (foreign) コスト =cost (cost) 欧米 =europe (europe & america) 四 =four (four) 倍 =times (times)

Table 4: Results of Sentence Alignment

sions Corpus, and the Yomiuri News Corpus, both of which have been in great part hand-aligned. In order to generate a parallel corpus automatically of this size, we would need to start with a raw corpus of about 1 million sentence pairs, assuming the same translation quality holds throughout. In contrast the NHK alignments are much poorer using the 6 or more dictionary matches criterion, with 1265 aligned sentences from 200K, or 0.63%, and 656 aligned sentences from 86K, or 0.76%. Most of the translation candidates have match counts lower than 6, probably too low to merit their inclusion as aligned sentences.

6 Conclusions

Content-aligned news corpora should be suitable raw material for training corpora, but different translation strategies seem to yield different proportions of literal versus stylistic translation candidates. The two news corpora we sentence-aligned yield differing results, with a higher proportion extracted from the Nikkei than the NHK. In both cases there is a big reduction from the content-aligned source corpus to sentence alignments we hope are more suitable as MT input. Perhaps higher correlations can be achieved using richer linguistic means such as expression and idiom matching, and perhaps Corpus Based methods can be applied iteratively to help achieve these.

References

Al Onaizan. Yaser et al. 2000. *Statistical Machine Translation, Final Report.* Johns Hopkins University, Baltimore, MD.

Brown. Peter et al. 1993. *The Mathematics of Machine Translation: Parameter Estimation. Computational Linguistics*, vol 19, number 2, pp 263-311, 1993.

Charniak, Eugene.: 2001. *Immediate Head Parsing for Language Models.* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, 2001

Fung, Pascale. and Church, Kenneth. W. 1993. *K-Vec: A New Approach for Aligning Parallel Texts.* In Proceedings 15th COLING pp 1096-1102 (1994)

Gale, William .A. and Church, Kenneth .W. 1991. *A Program for Aligning Sentences in Bilingual Corpora.* ACL 1991 pp177-184.

Haruno. Masahiko. and Yamazaki. Takefumi. 1991. *High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information.* ACL 1996 pp131-138.

Kudoh, Taku. and Matsumoto, Yuji. 2000. *Japanese Dependency Structure Analysis Based on Support Vector Machines.* In Empirical Methods in Natural Language processing and Very Large Corpora, Pages 18–25, 2000

Nightingale, Stephen. and Tanaka, Hideki. 2002. *Aligning for SMT: Results from Real World Corpora.* Presented at the Forum on Information Technology, FIT, Tokyo.

Nightingale, Stephen. and Tanaka, Hideki. 2003. *The Word is Mightier than the Count: Accumulating Translation Resources from Parsed Parallel Corpora.* in Proceedings CICLing 2003, Springer-Verlag LNCS 2588 pp420-431, Alexander Gelbukh (Ed).

Och, Franz Josef. and Ney, Hermann. 2000. *Improved Statistical Alignment Models.* in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.

Tanaka, Hideki. et al. 2002. *Speech to Speech Translation System for Monologues – Data Driven Approach* ICSLP, Denver Colorado, 2002