Annotated Corpora for Word Alignment Between Japanese and English and its Evaluation with MAP-based Word Aligner

Tsuyoshi Okita

Dublin City University, School of Computing
Glasnevin, Dublin 9, Ireland
tokita@computing.dcu.ie

Abstract

This paper presents two annotated corpora for word alignment between Japanese and English. We annotated on top of the IWSLT-2006 and the NTCIR-8 corpora. The IWSLT-2006 corpus is in the domain of travel conversation while the NTCIR-8 corpus is in the domain of patent. We annotated the first 500 sentence pairs from the IWSLT-2006 corpus and the first 100 sentence pairs from the NTCIR-8 corpus. After mentioned the annotation guideline, we present two evaluation algorithms how to use such hand-annotated corpora: although one is a well-known algorithm for word alignment researchers, one is novel which intends to evaluate a MAP-based word aligner of Okita et al. (2010b).

Keywords: Annotated Corpus for Word Alignment, Statistical Machine Translation, Evaluation

1. Introduction

Word alignment based on IBM and HMM Models (Brown et al., 1993; Vogel et al., 1996) is an important first step in Statistical Machine Translation (Koehn, 2010). This paper provides annotated corpora for word alignment between Japanese and English. We have two intentions.

Our first intention is to supply annotated corpora for word alignment between Japanese and English since such corpora do not exist. Unfortunately, the unavailability of such corpora is common in many language pairs due to the cost of annotation for word alignment. First of all, we need bilingual speakers. Secondly, we need to disentangle the inherent difficulties around non-literal translations and non-compound words (idiomatic expressions and Multi-Word Expressions) among others. Furthermore, people have been discouraged recently to build a new corpus due to the fact that the improvement of performance on word alignment may not lead to the improvement in terms of BLEU (which is the end-to-end translation quality) (Fraser and Marcu, 2007a). [1] Currently publicly available resources include English-French (Och and Ney, 2003), Romanian-English (Mihalcea and Pedersen, 2003), Chinese-English and Arabic-English (Consortium, 2006b; Consortium, 2006a), several European languages (Portuguese-English / Portuguese-French / Portuguese-Spanish / English-Spanish / English-French / French-Spanish) (Graca et al., 2008). A restricted resource includes German-English parallel corpus of Verbmobil project (Callison-Burch et al., 2004). Unavailability of such corpora creates two obstacles. The first obstacle is that we cannot evaluate the performance of word alignment by Alignment Error Rate (AER) (Och and Ney, 2003). The second obstacle is that without a hand-annotated corpus we cannot use several word alignment algorithms, such as a discriminative word aligner (Moore et al., 2006) and a semi-supervised word aligner (Fraser and Marcu, 2007b). Some word aligner, such as an MAP-based word aligner (Okita et al., 2010a; Okita and Way, 2011; Okita, 2011b), expects some knowledge about alignment links which can be partly supplied by such corpora.

Our second intention is to use these corpora as a benchmark for a word aligner which considers semantic knowledge. Incorporation of additional linguistic resource has been often prohibited in order to focus on the basic mechanism of word alignment in many machine translation evaluation campaigns. However, recent trend is to obtain better performance when we can incorporate linguistic resource (Okita et al., 2010b; Okita et al., 2010a; Okita and Way, 2011; Okita, 2011a): the demand is increasing to compare the standard word alignment algorithms with those algorithms which consider semantic knowledge. One way to compare these in a fairly manner would be not to leave the task of semantic annotation open for users, but to embed semantic annotation in a corpus: it is often the case that those who extracted linguistic knowledge better tend to obtain better overall performance compared to those who did not. Otherwise, despite that what we want to compare is word alignment algorithm, we compare extraction algorithm. Up until now, there have been not many semantically-informed word aligners proposed so far. Hence, the semantic annotation scheme proposed here may not satisfy the semantically-informed word aligner which will be proposed in future. In this sense, our corpus will be changed according to their requests.

The remainder of this paper is organized as follows. Section 2 describes statistics of hand-annotated corpora. Sec-

[1] There were two successful word alignment workshops at HLT-NAACL 2003 and at ACL 2005. The title of workshops was "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond."

tion 3 describes the guideline for hand annotation, and the semantic annotation is mentioned in Section 4. In Section 5, the two kinds of usage are presented: the first one is an established algorithm for generative / discriminative word aligners while the second one is a new algorithm intended for the MAP-based word aligner. We conclude in Section 6.

2. Statistics of Hand-Annotated Corpora

We provide annotation for 600 sentence pairs between EN-JP through two corora. The first corpus is the IWSLT-2006 sub-corpus (Paul, 2006) consisting of 500 sentence pairs. The second corpus is the NTCIR-8 sub-corpus (Fujii et al., 2010) of 100 sentence pairs. We use the alignment process of Lambert et al. (Lambert et al., 2006). However, we take the approach not to give the average of several persons, but rather to adopt one annotation which is consistent throughout the corpus.

3. Hand Annotation Guideline for Japanese and English

We built our hand annotation guideline based on the Chinese and English annotation guildine by the Gale project (Consortium, 2006b). Their guideline tend to examine extensively about 'not-translated' objects whose unit is often small (or at most word-level) while our approach tends to admit many-to-many mapping objects and translational noise (Okita et al., 2010b; Okita et al., 2010a) whose unit is slightly bigger than this (or up to chunk-level).

The result of word alignment will be different if the input text is segmented by the different morphological analyzers. It is known that different segmentation will yield different translation models whose performance are different: a lattice-based decoding (Dyer et al., 2008) is to take advantage of these different performances in order to choose an optimal path. The IWSLT-2006 corpus is provided with morphological analysis, while the NTCIR-8 corpus is provided without morphological analysis (hence, a user has to do morphological analysis). We follow the segmentation as is for the IWSLT-2006 corpus, while we did the morphological analysis on the NTCIR-8 corpus using Mecab / Cabocha (Kudo and Matsumoto, 2002).

As a notation, we use an extended A3 final file format used in GIZA++ (Och and Ney, 2003) to show the alignment links conveniently although this is unidirectional. The alignment links are shown as the index number according to the first sentence (English side). Hence, the word in the second sentence (Japanese side) which attaches index in its right shows an alignment link.

3.1. Anaphora (Pronoun)

A subject in Japanese or Chinese is often dropped after its first mention. However, a subject in English is often not dropped in the subsequent mention. Although Chinese-English annotated corpus by Gale project (Consortium, 2006b) attaches the pronoun or subject to its referents, our annotated corpus do not add such pronoun or subject. In the following example, a subject on the Japanese side (in this case 'わたしは') is omitted.

i 've never heard of this address around here .
NULL ({ 1 5 }) この({ 6 }) 住所({ 7 }) は({ }) この({ 8 }) 辺({ 9 }) で({ }) 聞い({ 4 }) た({ 2 }) こと({ }) ない({ 3 }) です({ }) ね({ }) 。({ 10 })
(Omitted: わたしは ({ 1 }))

3.2. Demonstrative Words

Demonstratives refer to 'this', 'that', 'these', 'those', 'here', and 'there'. If the both sides include demonstratives which corresponds together, they are aligned. If one side is referent, demonstratives can be linked to referent. In the following example, 'this' and 'この' are both demonstrative words, and they are corresponding.

does this bus stop at stoner avenue ?
NULL ({ 1 }) この ({ 2 }) バス({ 3 }) は({ }) ストーナー街({ 6 7 }) に({ }) 止まり({ 4 5 }) ます({ }) か({ 8 }) 。({ })

3.3. Measure Words

If the both sides include measure words, they can be linked. However, this is rare. Extra measure words on the Japanese side are quite common.Extra measure words can be glued to their head numbers, ordinal numbers or demonstratives. In the following example, '袋' and '個' show extra measure words where '一' means number. '一袋' corresponds to 'a bag of' while '一個' corresponds to 'a'.

a bag of cookies and a lemon bar .
NULL ({ 3 }) クッキー({ 4 }) 一 ({ 1 }) 袋 ({ 2 }) と({ 5 }) レモンバー({ 7 8 }) 一 ({ 6 }) 個 ({ 6 }) 。({ 9 })

3.4. Case Particles, Prepositions, and Passive Sentences

In Chinese, a link verb 'to be' (am, is, are, be, been, being) informs about the properties of its argument. This can separate an adjective from its noun. Japanese case particles can be thought of as this extension. Japanese case particles can indicate several meanings and functions, such as speaker affect and assertiveness. Japanese case particles are often appeared after noun, while the English side can be simply noun or can be noun with prepositions. Assumed that the Japanese side is morphologically separated (hence, case particles are separated from noun), we first align content words and then search the correspondence of case particles. Due to this order, case particles on the Japanese side can often be 'not-translated'. In the following example, there is no alignment links to 'は'. There is no alignment links to 'に' either.

does this bus stop at stoner avenue ? (NULL)
NULL ({ 1 }) この({ 2 }) バス({ 3 }) は ({ }) ストーナー街({ 6 7 }) に ({ }) 止まり({ 4 5 }) ます({ }) か({ 8 }) 。({ })

When the source and target sides take different voices, i.e. the active and the passive voices, these differences are often absorbed by the combination of the case particles and the main verbs.

3.5. Proper Noun

When the proper noun is compositional we take the minimum-match approach (Consortium, 2006b). When the proper noun is non-compositional we consider it as one entity when aligning it. For names, the first and the last names are aligned separately. A country name can be considered as a non-separate unit and can be aligned as a many-to-many mapping object. Acronyms of proper nouns can be treated as a non-separate unit. In the following example, '日本列島' and '太平洋' on the Japanese side are treated as non-separate units.

the japanese islands run northeast to southwest in the northwestern part of the pacific ocean .
NULL ({ 1 9 13 }) 日本列島 ({ 2 3 }) は({ }) 太平洋 ({ 14 15 }) の({ 12 }) 北西({ 10 11 }) に({ 8 }) 、({ }) 北東({ 5 }) から({ 6 }) 南西({ 7 }) の({ }) 方向({ }) に({ }) 伸び({ 4 }) てい({ }) ます({ }) 。({ 16 })

3.6. Determiners

Since there is no determiner in Japanese, determiners are only existed on the English side. The determiners on the English side can be either 1) omitted or 2) translated into other words than a determiner. In the following example, 'the' is not aligned while 'light' and '信号' are aligned.

the light was red .
NULL ({ 1 }) 信号({ 2 }) は({ }) 赤({ 4 }) でし({ 3 }) た({ }) 。({ 5 })

3.7. Auxiliary Verbs

The auxiliary verbs whose function can be passive, progressive, model, etc, can be added both on the Japanese and the English sides. When they appear on both sides, they are simply linked. When they appear on the one side, the extra auxiliary verbs can be glued to the main verb. In the following example, 'could' is an auxiliary verb. Literal translation would be '預かることができますか' where '預かる' corresponds to 'keep'. Hence, we align 'could' to 'て下さい'.

could you keep this baggage ?
NULL ({ 2}) この({ 4 }) 荷物({ 5 }) を({ }) 預かっ({ 3 }) て下さい ({ 1 }) 。({ 6 })

3.8. Expletives

Expletives refer to the words which have a syntactic role but contribute nothing to the meaning. Examples are 'it', 'there', and 'here'. If it has equivalent counterpart, we can align them. In the following example, 'there' is expletive. In this case, 'there are' corresponds to 'ある' which is inflected to 'あり'. Rather than considering 'there' as not-translated, we align 'there are' with 'あり'.

are there any baseball games today ?
NULL ({ 3 }) 今日({ 6 }) 、({ }) 野球({ 4 }) の({ }) 試合({ 5 }) は({ }) あり ({ 1 2 }) ます({ }) か({ 7 }) 。({ })

3.9. Conjunctions

Conjunctions such as 'and' can be corresponded either to 1) 'and', 2) ',' and 3) omitted. In the case of 2) ',' can be aligned to 'and'. In the following example, 'and' and 'と' are aligned.

i give you his telephone number and address .
NULL ({ }) 彼({ 4 }) の({ }) 電話番号({ 5 6 }) と ({ 7 }) 住所({ 8 }) を({ }) 教え({ 2 3 }) てあげよ({ }) う({ }) 。({ 9 })

3.10. Verb Particles

In English, a verb combined with a preposition, or an adverb, or an adverbial particle is called a verb particle. Verb particles are often inseparable from their verbs which have a fixed meaning. We treat this as a many-to-many mapping alignment. In the following example, 'wrap up' is a verb particle which aligns to '包ま'.

no worry about that . i 'll take it and you need not wrap it up .
NULL ({ 6 10 11 }) 結構({ 1 2 3 4 }) です({ 1 2 3 4 }) 。({ 5 }) それ({ 9 }) を({ }) 頂き({ 7 8 }) ましょ({ }) う({ }) 。({ }) 包ま ({ 14 15 16 }) なく({ 13 }) て({ }) も({ }) 構い({ 12 }) ません({ }) 。({ 17 })

3.11. Possessives

Possessives can be appeared either ''s' or 'of' in English, and appeared some equivalent forms or omitted in Japanese. If there is no counterparts, they can be marked as 'not-translated'. In the following example, possessive ''s' corresponds to '用の'. However, we align 'children 's' with '子供用の' since we could expect that the correspondece between ''s' and '用の' is quite rare.

i 'd like a children 's sweater .
NULL ({ 1 4 }) 子供({ 5 6 }) 用 ({ 5 6 }) の({ 5 6 }) セーター({ 7 }) が({ }) 欲しい({ 2 3 }) の({ }) です({ 6 }) が({ }) 。({ 8 })

3.12. Subordinate Clauses

In English, subordinate conjunctions refer to 'after', 'although', 'as', 'because', 'before', 'even if', 'in order that', 'once', 'provided that', 'rather than', 'since', 'so that', 'than', 'that', 'though', 'unless', 'when', "whenever', 'where', 'whereas', 'whenever', 'whether', 'while', 'why', and so forth, while relative pronouns refer to 'that', 'which', 'whichever", 'who', 'whoever', 'whom', 'whose', and so forth.

Subordinate conjunctions on the English side has often lexical counterparts on the Japanese side, while for relative pronouns on the English side are often omitted or taken other forms including comma on the Japanese side. Hence, the

relative pronouns can be aligned to comma on the Japanese side. In the following example, ‘when the fluid pressure cylinder 31 is used’ is a subordinate clause. A subordinate conjunction ‘when’ is in this case translated into the lexical counterpart ‘場合’.

when the fluid pressure cylinder 31 is used , fluid is gradually applied .
NULL ({ 2 7 8 9 11 }) 流体({ 3 }) 圧({ 4 }) シ リ ン ダ({ 5 }) 3 ({ 6 }) 1 ({ 6 }) の({ }) 場合 ({ 1 }) は({ }) 流体({ 10 }) が({ }) 徐々に({ 12 }) 排出({ 11 13 }) さ({ 11 13 }) れる({ 11 13 }) こと({ }) と({ }) なる({ }) . ({ 14 })

3.13. Punctuations

The punctuations, such as a period and a comma, on the English side often have their equivalent counterparts on the Japanese side. They can be often aligned. However, ‘!’ and ‘?’ often do not have their equivalent counterparts as punctuations but some lexical counterparts on the Japanese side. In the following, ‘?’ is aligned to ‘か’.

do you do alterations ?
NULL ({ 1 2 }) 直し({ 4 }) は({ }) し({ 3 }) てい({ }) ます({ }) か ({ 5 }) 。({ })

3.14. Translational Noise or Contextually Attached Words / Sentences

Translators may add contextual words for better understanding. Although these extra words can be considered as words associated or related when the length of extra words are short, if they are long such extra words are considered as translational noise. Although there is no example in the IWSLT-2006 and the NTCIR-8 corpora, it may be possible that some sentences are ’not-translated’ or contextually attached (Okita, 2009). In the following example, ‘ものである’ does not correspond to any words.

in other words , in the embodiments shown in figs . 32 and 33 , running is effected in such a manner that the wheels 2 straddle the guides 5 in a manner similar to that shown in fig . 30 . (NULL)
NULL ({ 5 6 15 21 23 24 28 32 }) つまり({ 1 2 3 }) , ({ 4 }) 図({ 10 11 }) 3 ({ 12 }) 2 ({ 12 }) , ({ 13 }) 図({ }) 3 ({ 14 }) 3 ({ 14 }) に({ }) 示す({ 8 9 }) 実施({ 7 }) 例({ 7 }) は({ }) , ({ 6 }) 図({ 39 40 }) 3 ({ 41 }) 0 ({ 41 }) に({ }) 示す({ 37 38 }) 実施({ 36 }) 例({ 36 }) と({ 35 }) 同様({ 31 32 33 34 35 }) に({ 31 32 33 34 35 }) 車輪({ 25 }) 2 ({ 26 }) が({ }) ガイド({ 29 }) 5 ({ 30 }) を({ }) 跨ぐ({ 27 }) よう({ 19 20 21 22 }) に({ 19 20 21 22 }) し({ 19 20 21 22 }) て({ 19 20 21 22 }) 走行({ 16 17 18 }) する({ 16 17 18 }) もの ({ }) で ({ }) ある ({ }) . ({ 42 })

3.15. Rhetorically Attached Words

When the repetition structure let omit the extra element only on the one side (for example, some noun are shared only on the one side), the extra elements can be ‘not-translated’. In the following example, ‘figs . 32 and 33’ is translated into ‘図32, 図33’ where ‘figs’ corresponds to ‘図’. That is, the Japanese side is ‘fig . 32 and fig . 33’ and there is no counterparts of the second ‘fig’.

in other words , in the embodiments shown in figs . 32 and 33 , running is effected in such a manner that the wheels 2 straddle the guides 5 in a manner similar to that shown in fig . 30 .
NULL ({ 5 6 15 21 23 24 28 32 }) つまり({ 1 2 3 }) , ({ 4 }) 図 ({ 10 11 }) 3 ({ 12 }) 2 ({ 12 }) , ({ 13 }) 図 ({ }) 3 ({ 14 }) 3 ({ 14 }) に({ }) 示す({ 8 9 }) 実施({ 7 }) 例({ 7 }) は({ }) , ({ 6 }) 図({ 39 40 }) 3 ({ 41 }) 0 ({ 41 }) に({ }) 示す({ 37 38 }) 実施({ 36 }) 例({ 36 }) と({ 35 }) 同様({ 31 32 33 34 35 }) に({ 31 32 33 34 35 }) 車輪({ 25 }) 2 ({ 26 }) が({ }) ガイド({ 29 }) 5 ({ 30 }) を({ }) 跨ぐ({ 27 }) よう({ 19 20 21 22 }) に({ 19 20 21 22 }) し({ 19 20 21 22 }) て({ 19 20 21 22 }) 走行({ 16 17 18 }) する({ 16 17 18 }) もの({ }) で({ }) ある({ }) . ({ 42 })

3.16. Unmatched / Unattached Words

Words which help smoothing of the sentence often do not carry meaning. In the following example, ‘ね’ helps smoothing of a sentence without carrying a meaning. This word is considered to be ‘not-translated’.

i ’ve never heard of this address around here . (NULL)
NULL ({ 1 5 }) この({ 6 }) 住所({ 7 }) は({ }) この({ 8 }) 辺({ 9 }) で({ }) 聞い({ 4 }) た({ 2 }) こと({ }) ない({ 3 }) です({ }) ね ({ }) 。({ 10 })

3.17. Register Modes

Japanese has three types of honorific speeches, such as polite, respectful, and humble languages. Since these appear only in colloquial modes, this is related only to the IWSLT-2006 corpus. In the following example, ‘御’ falls among humble language. In this case, we consider this as ‘not-translated’.

we want to have a table near the window . (NULL)
NULL ({ 1 5 }) 窓際({ 7 8 9 }) の({ 7 8 9 }) 席({ 6 }) を({ }) 御 ({ }) 願い({ 2 3 4 }) し({ 2 3 4 }) ます({ }) 。({ 10 })

4. Semantic Annotation

The hand-annotated alignment links are purely for the purpose of evaluation while the semantic annotation is not necessarily restricted within tiny corpus but also for training corpus. As is mentioned in Section 1, since the semantically-informed word aligner has not been established yet, the details of semantic annotation will be changed. In this sense, semantic annotation below is quite experimental (Okita, 2011a) where we annotated first by tools and correct them by hands (only apparent mistakes).

- MWEs: MWEs are extracted by the statistical MWE extraction method (Kupiec, 1993; Okita and Way, 2011),
- Lexical semantics: lexical semantics in the form of Senseval-2 / 3 data (Snyder and Palmer, 2004) was done using Japanese WordNet (Bond et al., 2009) complemented by human beings,
- Dependency structures: other semantic structures, such as coordination structure and dependency structure, are extracted by dependency parser (Kurohashi and Nagao, 1998),
- Translational noise: translational noise is extracted by the method which is mostly statistical but with human beings.

5. Usage

In this section, we presents two different usage of these corpora. The first algorithm is an established method for evaluating Bayesian / discriminative word aligner by AER, precision, or recall. The second algorithm is a new method for evaluating a MAP-based word aligner (Okita et al., 2010b) by AER, precision, or recall.

Alignment error rate (Och and Ney, 2003) is defined via the set of sure alignments S, possible alignments P, and whole alignments A. Recall is defined on S while precision is defined on P where $P \supset S$. These definitions are shown as in (1):

$$\begin{cases} \text{Precision}(A,P) &= \dfrac{|A \bigcap P|}{|A|}, \\ \text{Recall}(A,S) &= \dfrac{|A \bigcap S|}{|S|}, \\ \text{AER}(A,P,S) &= 1 - \dfrac{|A \bigcap S| + |A \bigcap P|}{|A| + |S|} \end{cases} \quad (1)$$

5.1. Evaluation of Generative / Discriminative Word Aligner

The standard usage of these hand-annotated corpora is for the evaluation for word aligner measured by Alignment Error Rate, precision, or recall (Och and Ney, 2003). We prepare the hand-annotated parallel corpus C (e'=$\{e'_1, \ldots, e'_{|C|}\}$, f'=$\{f'_1, \ldots, f'_{|C|}\}$, a'=$\{a'[1]_1^{|e'|}, \ldots, a'[|C|]_{|1|}^{|e'|}\}$) which is typically in small size and the parallel corpus D (e,f). By definition, C has alignment information for whole of sentence pairs while D does not have.

The standard algorithm is shown in Algorithm 1 (Och and Ney, 2003; Moore et al., 2006; Blunsom and Cohn, 2006; Graca et al., 2008). Note that in order to obtain the Viterbi alignments a_E for IBM Models 1 and 2, the practical method would be to run additional 1 cycle of Model 4 in the case of GIZA++.[2] Note that between EN-FR using Och's hand-annotated corpus, there is several version which divide 447 / 484 sentence pairs depending on papers. Figure 1 shows the performance of GIZA++ on EN-FR Hansard datasets [3] in this way.

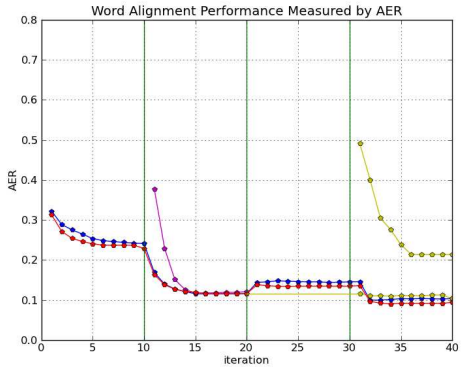

Figure 1: AER performance on EN-FR. The line which starts from the 11th iteration shows that this is only trained by the HMM model. The other line which starts from the 31st iteration shows that this is only trained by IBM Model 4. The red and blue lines show that they are different translation directions. The size of D is 1.1 million sentence pairs (which is the whole set of ISI version of Hansard training corpus) while that of C is 488 sentence pairs.

Algorithm 1 Evaluation of Generative / Discriminative Word Aligner (Uni-directional)

Given: Word aligner M, parallel corpus D=(e, f), hand-annotated parallel corpus C=(e', f', a').
Step 1: Concatenate D and C in the source and the target sides respectively to make E=($e_{new} = \{e', e\}$, $f_{new} = \{f', f\}$).
Step 2: For the given E=(e_{new}, f_{new}), we run a word aligner M which outputs the Viterbi alignments a_E.
Step 3: Among the Viterbi alignments a_E, we extract the Viterbi alignments a_C which corresponds to the alignment between e' and f'.
Step 4: Compare the results of alignment a_C and the hand-annotated alignment a'. Standard criteria are AER, precision, and recall.

Figure 2 shows the performance of GIZA++ on the EN-JP corpus consisting of 10 iterations of IBM Model 1, 10

[2] That is, GIZA++ commands are follows and we will get the Viterbi alignent in a file called *IBM1dictionary.A3.final*: *plain2snt.out train.en train.fr; GIZA++ -S train.en.vcb -T train.fr.vcb -C train.en_train.fr.snt -p0 0.98 -model1iterations 10 -model2iterations 0 -model3iterations 0 -model4iterations 1 -model5iterations 0 -hmmiterations 0 -o IBM1dictionary;*

[3] There are two versions of Hansard corpora: one is the version of USC / ISI edited by Ulrich Germann, and the other is the LDC catalogue LDC95T20 edited by Salim Roukos, et al.. We used the former one.

iterations of HMM Model, 10 iterations of IBM Model 3 and 10 iterations of Model 4.

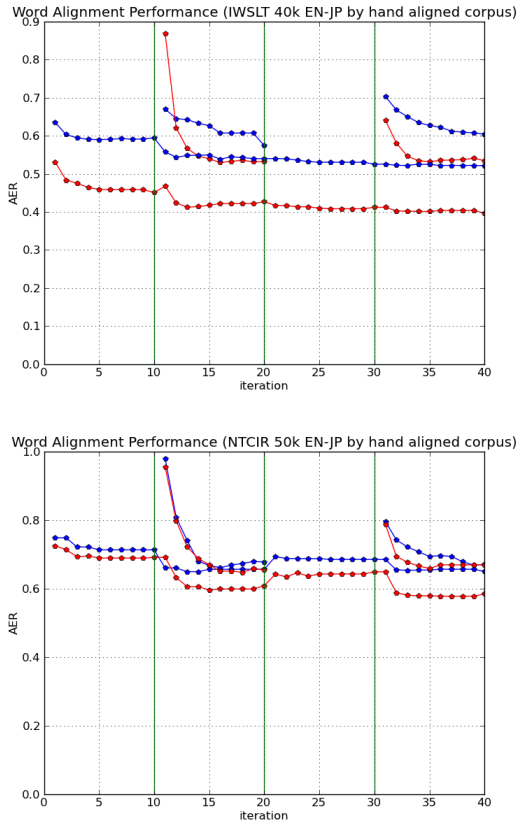

Figure 2: AER performance on EN-JP. The upper figure shows the performance on the IWSLT-2006 corpus, and the lower figure shows the performance on the NTCIR-8 corpus. The size of D is 40k sentence pairs for the IWSLT-2006 corpus and 200k sentence pairs for the NTCIR-8 corpus, while that of C is 500 and 100 sentence pairs, respectively.

The first observation is that the performance on EN-JP is considerably worse than the performance on EN-FR. (In the case of EN-FR, it achieves 0.11 AER between 15 to 20 iterations and 0.09 AER between 34 to 40 iterations.) In the case of EN-JP, it achieves 0.52 on the IWSLT-2006 corpus and 0.62 on the NTCIR-8 corpus. The second observation is the variability for different translation directions. In the case of EN-FR, the red and blue lines are quite identical, while in the case of EN-JP, the red line is always better than the blue line.[4]

5.2. Evaluation of MAP-based Word Aligner

The alternative usage is to evaluate the MAP-based word aligner (Okita et al., 2010b). The difference between generative / discriminative word aligner and this MAP-based word aligner is whether the input to the word aligner includes information about the alignment links or not. In the MAP-based word aligner, the alignment links, which work as prior knowledge, will guide the word alignment. Note that the alignment links are the result of generative /discriminative word aligner while they are supplied to the MAP-based word aligner. In this sense, one of the performance measure of this kind of word aligner is the performance when the given prior knowledge is, say, 60% (In Figure 4, Point 4 in the x-axis shows 66%). Our interests in this word aligner are 1) to measure the performance which will achieve the performance near 100%, and 2) to measure the performance of the practical level within which we can supply the prior knowledge. Note that in the latter case, the practical level depends on statistics of MWEs, links via lexical semantics, translational noise, and so forth. However, since it is not practically possible in many cases to get the comparable performance in 1) and 2), Algorithm 2 describes a method to measure the performance in various points at $0\%, 10\%, 20\%, \ldots, 100\%$. As is shown in Figure 4, the results will be varied depending on which links we provide as prior knowledge. When we only provide alignment links via hand-annotated corpus D, we can provide the evaluation which is limited from 0 up to D / $C + D$ in the x-axis. In this sense, if the rate of D / $C + D$ is small,[5] it would be better to provide prior knowledge about alignment links additionary via other sources, such as MWEs, lexical semantics, translational noise, and so forth.

Algorithm 2 Evaluation of MAP-based Word Aligner (Uni-directional)

Given: Word aligner M^{MAP}, parallel corpus D=(e,f,(a)), hand-annotated parallel corpus C=(e',f',a').
Step 1: To make A=($a_0, a_{10}, a_{20}, \ldots, a_{100}$) by sampling randomly $0\%, 10\%, 20\%, \ldots, 100\%$ of a'(or $a' \cup a$).
Step 2: To make E=($e_{new} = \{e', e\}$, $f_{new} = \{f', f\}$) by concatenating D and C in the source and the target sides respectively.
Step 3: For the given E=(e_{new}, f_{new}) and a_i of A ($i = 0, \ldots, 100$), we run a MAP-based word aligner M^{MAP} which outputs the Viterbi alignments a_E.
Step 4: Among the Viterbi alignments a_E, we extract the Viterbi alignments a_C which corresponds to the alignment between e' and f'.
Step 5: Compare the results of alignment a_C and the hand-annotated alignment a'. Standard criteria are AER, precision, and recall.

6. Conclusion and Further Study

This paper presents two annotated corpora for word alignment between Japanese and English. We annotated corpora

[4] Our results in terms of BLEU for NTCIR-8 can be available in (Okita et al., 2010c).

[5] The case of Hansard corpus in Figure 1, this rate is at most 0.49%.

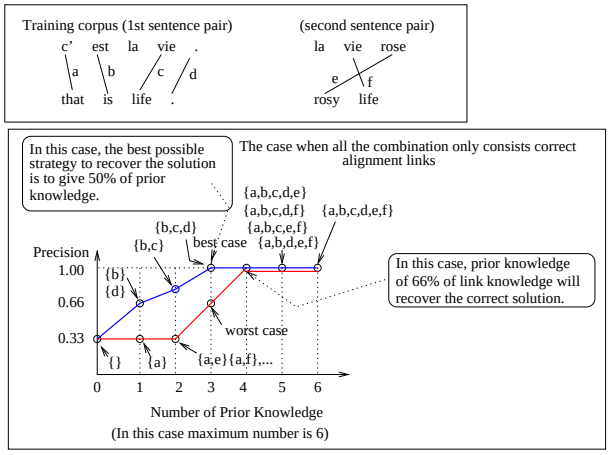

Figure 3: We show how much information about alignment links was required to recover the specified precision shown in the y-axis. If we gave more than four correct alignment links, the MAP-based aligner was able to obtain the precision 1.0 (i.e. correct alignment). If we gave three correct alignment links, the solution was correct in the case of $\{b, c, d\}$. However, for other cases such as $\{a, e, f\}$, the precision was 0.66. The point at 0 in the x-axis indicates the performance of a traditional word aligner where no prior knowledge was provided. The precision was 0.33.

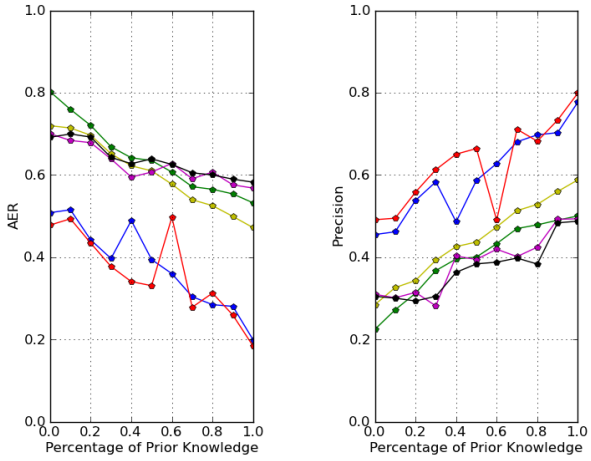

Figure 4: AER performance (left) and precision (right) for the MAP-based word aligner based on the IBM Model 1 for Hansard EN-FR (blue) / FR-EN (red), for IWSLT-2006 JP-EN (yellow) / EN-JP (green), and for NTCIR-8 JP-EN (violet) / EN-JP (black). The size of D=C is 484 (Hansard), 500 (IWSLT-2006) and 100 (NTCIR-8).

based on two existing parallel corpora: the IWSLT-2006 and the NTCIR-8 corpora. We briefly mentioned the annotation guideline when we used building these corpora. Then, we presented two algorithms for evaluation.

One avenue for further research is about the input structure of semantic knowledge. It may be straight forward to construct network structure of semantic knowledge (We refer to the semantic knowledge incorporating into the system by T4ME project WP2 (Federmann, 2011; Okita and van Genabith, 2012)). However, such network structure will not be an input of the word aligner as it is, but will possibly lose its structural information.

7. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`) at Dublin City University. We would also like to thank the Irish Centre for High-End Computing.

8. References

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with Conditional Random Fields. *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-06)*, pages 65–72.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. *In Proceedings of the 7th Workshop on Asian Language Resources (in conjunction with ACL-IJCNLP 2009)*, pages 1–8.

Peter F. Brown, Vincent J.D Pietra, Stephen A.D.Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, Vol.19, Issue 2*, pages 263–311.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. *In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 175–182.

Linguistic Data Consortium. 2006a. Guidelines for arabic-english word alignment. *http://www.ldc.upenn.edu/Project/GALE*.

Linguistic Data Consortium. 2006b. Guidelines for chinese-english word alignment. *http://www.ldc.upenn.edu/Project/GALE*.

C. Dyer, S. Muresan, and P. Resnik. 2008. Generalizing word lattice translation. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 1012–1020.

Christian Federmann. 2011. Results from the ml4hmt shared task on applying machine learning techniques to optimise the division of labour in hybrid mt. *In Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation LIHMT and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation ML4HMT*, pages 110–117.

Alexander Fraser and Daniel Marcu. 2007a. Getting the structure right for word alignment: Leaf. *In Proceedings*

of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 51–60.

Alexander Fraser and Daniel Marcu. 2007b. Measuring word alignment quality for statistical machine translation. *Computational Linguistics, Squibs and Discussion*, 33(3):293–303.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 293–302.

Joao Graca, Joana Paulo Pardal, Luisa Coheur, and Diamantino Antonio Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. *In The 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 986–993.

Philipp Koehn. 2010. Statistical machine translation. *Cambridge University Press*.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis unsing cascaded chunking. *In Proceedings of the tenth Conference on Computational Natural Language Learning (CoNLL-2002)*, pages 63–69.

Julian Kupiec. 1993. An algorithm for finding Noun phrase correspondences in bilingual corpora. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22.

Sadao Kurohashi and Makoto Nagao. 1998. Building a japanese parsed corpus while improving the parsing system. *In Proceedings of The First International Conference on Language Resources & Evaluation*, pages 719–724.

Patrik Lambert, Adria de Gispert, Rafael Banchs, and Jose B. Marino. 2006. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation (Springer)*, 39(4):267–285.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. *In Proceedings of the 38th In Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, collocated with Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics (HLT / NAACL 2003)*, pages 1–10.

Robert C. Moore, Yih Wen-tau, and Andreas Bode. 2006. Improved discriminative bilingual word alignment. *In Proceedings of International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL2006)*, pages 513–520.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Tsuyoshi Okita and Josef van Genabith. 2012. Minimum bayes risk decoding with enlarged hypothesis space in system combination. *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012). LNCS 7182 Part II. A. Gelbukh (Ed.)*, pages 40–51.

Tsuyoshi Okita and Andy Way. 2011. Given bilingual terminology in statistical machine translation: Mwe-sensitve word alignment and hierarchical pitman-yor process-based translation model smoothing. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 269–274.

Tsuyoshi Okita, Yvette Graham, and Andy Way. 2010a. Gap between theory and practice: Noise sensitive word alignment in machine translation. *In Proceedings of the Workshop on Applications of Pattern Analysis (WAPA2010). Cumberland Lodge, England.*, pages 119–126.

Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010b. Multi-Word Expression sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA2010, collocated with COLING2010), Beijing, China.*, pages 1–8.

Tsuyoshi Okita, Jie Jiang, Rejwanul Haque, Hala Al-Maghout, Jinhua Du, Sudip Kumar Naskar, and Andy Way. 2010c. MaTrEx: the DCU MT System for NTCIR-8. *In Proceedings of the MII Test Collection for IR Systems-8 Meeting (NTCIR-8), Tokyo.*, pages 377–383.

Tsuyoshi Okita. 2009. Data cleaning for word alignment. *In Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) Student Research Workshop*, pages 72–80.

Tsuyoshi Okita. 2011a. Meaning representations in statistical word alignment. *Learning Semantics Workshop (collocated with NIPS 2011)*.

Tsuyoshi Okita. 2011b. Word alignment and smoothing method in statistical machine translation: Noise, prior knowledge and overfitting. *PhD thesis Dublin City University*.

Michael Paul. 2006. Overview of the iwslt 2006 evaluation campaign. *In Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2006)*, pages 1–15.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. *In Proceedings of ACL / SIGLEX Senseval-3*, pages 41–43.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-Based word alignment in statistical translation. *In Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 836–841.