Online Search Interface for the *Sejong* Korean-Japanese Bilingual Corpus and Auto-interpolation of Phrase Alignment

Sanghoun Song
Korea Univ.
Anam-dong, Sungbuk-gu, Seoul,
South Korea
sanghoun@gmail.com

Francis Bond
NICT Language Infrastructure Group
2-2-2 Hikaridai, Seika-cho,
Soraku-gun, Kyoto, Japan
bond@ieee.org

Abstract

A user-friendly interface to search bilingual resources is of great help to NLP developers as well as pure-linguists. Using bilingual resources is difficult for linguists who are unfamiliar with computation, which hampers capabilities of bilingual resources. NLP developers sometimes need a kind of workbench to check their resources. The online interface this paper introduces can satisfy these needs. In order to implement the interface, this research deals with how to align Korean and Japanese phrases and interpolates them into the original bilingual corpus in an automatic way.

1 Introduction

Bilingual or multilingual corpora are significant language resources in various language studies, such as language education, comparative linguistics, in particular, NLP. What holds the key position in bilingual resources is how to align linguistic units between two languages. In this context, three fundamental questions about how to harness bilingual resources can be raised; (i) which linguistic unit or level should correspond to those in the corresponding language? (ii) which method should be employed for alignment? (iii) which environments should be prepared for users?

This paper covers these matters related to bilingual resources and their use. The language resource that this paper handles is the *Sejong Korean-Japanese Bilingual Corpus* (henceforth SKJBC).[1] The original version of the SKJBC, constructed in a XML format, aligns sentence by sentence or paragraph by paragraph. This research re-organizes and re-aligns the original version using GIZA++ (Och and Ney, 2003) and *Moses* (Koehn et al. 2007), and interpolates the aligning information into each original file automatically. Turning to the interface, this research converts the whole data into a database system (MySQL) to guarantee data integrity. Building on the database, this research implements an online search system accessible without any restrictions; dubbed NARA[2].

2 The SKJBC

The SKJBC had been constructed as a subset of the *Sejong* project[3] which had been carried out from 1998 to 2007, sponsored by the Korean government. The SKJBC is divided into two parts; one is the raw corpus annotated only with sentence aligning indices, the other is the POS-tagged corpus, in which the tag set for Korean complies with the POS-tagging guideline of the *Sejong* project, and the morphological analysis for Japanese is based on *ChaSen* (Matsumoto et al., 1999). This paper is exclusively concerned with the latter, because it is highly necessary for the phrase alignment to make use of well-segmented and well-refined data. Table 1 illustrates the basic configuration of the SKJBC.

Since the prime purpose of the *Sejong* project was to build up balanced corpora, the SKJBC consists of various genres, as shown in Figure 1. This makes a clear difference from other bilingual resources where the data-type is normally homogeneous (e.g. newspapers). Moreover, since it had been strictly prohibited to manipulate the

[1] The SKJBC is readily available for academic and research purposes only. For information on license conditions and others, please contact the *National Academy of Korean Language* (http://www.korean.go.kr/eng/index.jsp).

[2] The name, NARA, has meanings in both Korean and Japanese. It is a local name in Japan; it also means 'a country' in Korean. Since the name can properly stands for this research's goal, the name has been used as a project title.

[3] http://www.sejong.or.kr/eindex.php

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 146–149,
Suntec, Singapore, 6-7 August 2009.

original source for any reasons, the data in SKJBC fully reflect on the real usage of Korean. These characteristics, however, sometimes work against computational implementation. Bi-texts do not always correspond to each other sentence by sentence; we can find out that there are a number of cases that a sentence matches two or more sentences in the other language or the corresponding sentences might be non-existent. In other words, it is almost impossible to align all the sentences only one-to-one. These cases eventually produce the multiple-to-multiple alignment, unless annotators discard or separate them artificially. No artificial manipulation was allowed under construction, the SKJBC contains quite a few pairs in a multiple-to-multiple relation.

	Korean		Japanese	
	type	token	type	token
document	50 (KoJa : 38, JaKo : 12)			
sentence		4,030		4,038
word	21,734	43,534	10,452	93,395
morpheme	9,483	101,266	10,223	

Table 1. Configuration of the SKJBC

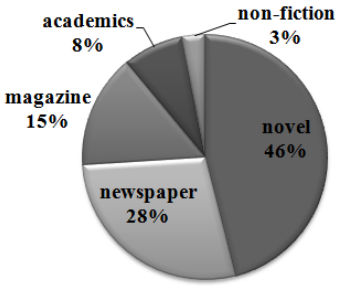

Figure 1. Composition of the SKJBC

3 Alignment

This section is connected with the first question raised in section 1; the proper level of alignment. Most bilingual corpora, including the SKJBC, have been constructed sentence by sentence despite shortcomings, because it costs too much time and effort to annotate word or phrase correspondence by hand (Abeillé, 2003). To annotate more specified alignment between two languages is to enhance the utility value of the resource; this research, first of all, considers how to align at the level of word and phrase.

Multiple Alignments: Because of the problem mentioned in the previous section, the pairs which do not match in a one-to-one relation were excluded from the target of alignment. Throughout a preliminary experiment, it was born out that, if they remained, they led to a worse result. After casting them away, the number of target sentences is 3,776, which account for about 86 percent of the whole data.

Word vs. Phrase: To make the units equivalent as far as possible is the crucial factor in aligning as accurately as possible. One of the main issues that should be taken into account in aligning Korean and Japanese phrases is word boundary. Though Korean and Japanese share lots of features, the boundary of word or phrase is inconsistent with each other. The general concept to segment words in Korean is the so-called *ejeol*, tantamount to word-spacing, whereas that in Japanese is *bunsetsu*, what we say. The difference stems from the different writing style; Korean inserts spacing between words, while Japanese seldom uses spacing. Consequently, each word in Korean is virtually equivalent to a phrasal unit in Japanese, as given in (1-2).

(1) 웃었다 웃/VV+었/EP+다/EF
wus-ess-ta
laugh-PAST-DC 'laughed'

(2) 笑っ 笑う/VIN
た た/AU
warat-ta
laugh-PAST 'langhed'

The first line (i.e. *ejeol*) in (1) for Korean corresponds to the first and second line (i.e. *bunsetsu*) in (2). Hence, it is the most suitable choice to align Korean morphemes (e.g. 웃 *wus*) and Japanese *bunsetsu* (e.g. 笑っ *warat*).

On the other hand, there is a clear cut between lemmatized lexical forms and surface forms in Japanese, (e.g. 笑う and 笑っ in the above, respectively), whereas there is none in Korean. In order to prevent the result from being biased, this paper establishes two training sets (i.e. lemmatized and surface forms) for alignment.

Word Sense Disambiguation (WSD): Other than the above issues, it is also needed to consider WSD. For example, a Korean word 삶 *salm* has two meanings; one is 'life' as a nominal expression, the other is 'boil' as a verbal lexeme, which correspond to 生 *sei*, 煮る *niru*, respectively. This research, therefore, makes training data composed of each morpheme plus its POS tag, such as '삶/NNG' and '生/NCPV'.

4 Auto-interpolation

Turning to the second question, this part covers how to align and annotate. Were it not for automatic processing, it would be painstaking work to construct bilingual resources even line by line. One popular toolkit to align linguistic units be-

tween two languages in an unsupervised way is GIZA++.

Even though GIZA++ yields fairly good 'word' alignment, much remains still to be done. For instance, those who want to study two or more languages from a comparative stance are certain to need syntactic data which offer more information about language diversity than plain word-mapping. Besides, Statistical Machine Translation (SMT) commonly runs under the phrase-based model. This research employs the *Moses* toolkit to establish phrase tables. The baseline of this research is the factorless one with a five-gram language model.

In order to measure the accuracy of alignment, this research uses the BLEU scoring (Papineni et al., 2002) which has been widely used for evaluating SMT, under the hypothesis that the BLEU score denotes how well-established the phrase table is. For the evaluation purpose, 500 sentences were selected from the SKJBC at random, and tested within each SMT baseline, as given in Table 2.

	KoJa	JaKo
lemmatized	72.72	71.37
surface	72.98	72.83
surface + tag	70.55	68.26

Table 2. BLEU Score

(3)
```
<Link xtargets="1.1.p8.s4 ; 1.1.p14.s3">
<phr xtargets="w1 w2 w3 w4 ; w1 w2 w3 w4">
<wrd xtargets="w3 ; w1">
<wrd xtargets="w5 ; w5">
</Link>
```

(4)
```
<s id=1.1.p8.s4>
그래야   kulaya          'then'
  <w id=w1>그러/VV</w>   kule
  <w id=w2>ㅓ야/EC</w>   yeya
자유롭지.  caywulop-ci    'be free'
  <w id=w3>자유롭/VA</w>  caywulop 'free'
  <w id=w4>지/EF</w>      ci
  <w id=w5>./SF</w>
</s>
```

(5)
```
<s id=1.1.p14.s3>
<w id=w1>自由</w>   自由/NG  jiyuu   'freedom'
<w id=w2>だ</w>     だ/AU    da
<w id=w3>から</w>   から/PJC kara
<w id=w4>ね</w>     ね/PEN   ne
<w id=w5>。</w>     。/SYF
</s>
```

Korean and Japanese are typologically very similar. In particular, they have very similar word order, which makes them easy to align using *GIZA++* and *Moses*. Therefore, we could expect the baselines to perform well, and Table 2 proves it. Table 2 indicates the baselines using Japanese surface forms are slightly better than those using lemmatized forms. The next step is to confirm whether or not the baselines with POS tags decrease performance. The last line in Table 2 implies it is not the case, there is a slight decline.

Building on the last baselines, this research interpolates word and phrase aligning information into the original XML files as presented in (3-5), which means 'Then, you will be free'. Figure 2 represents how the online interface this paper proposes handles (3-5).

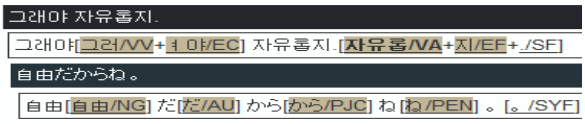

Figure 2. Sample of Online Interface

5 Online Search Interface

Last but maybe the most important is a user-friendly interface. Those who have a solid background in computation could take advantage of computational surroundings (e.g. *Moses*). Most linguists, however, are not aware of how to use bilingual data so well. It might look uneasy or even vague for them to harness bilingual resources for their current research. That means, no matter how good the bilingual resource is or no matter how well-trained the word or phrase table is, unless there is an available interface, the resource becomes no more than a very restricted one for a small number of people. Bilingual resources are not NLP-dominated ones, admitting NLP developers employ them most widely. They are also useful in doing comparative language research, making learning materials, or even human translation. Since one of the easiest interface in these days would be web-browsers, this research provide a web-interface; NARA (*ver*. 2).[4]

The interface of NARA system looks like a common search site (e.g. Google). A simple search option takes a position on the front side, assuming most of users are unfamiliar with linguistic terms. On the other hand, advanced search mode, as given in Figure 3, offers more specialized search options. One can search by tag, morpheme, or word with specific sub-options, such as matching type. One can also select the result format such as word, sentence, or span. In order to compare the search result in various ways, there are some configuration options, such as search direction (i.e. KoJa or JaKo), genre, source language, etc.

[4] http://corpus.mireene.com/nara.php

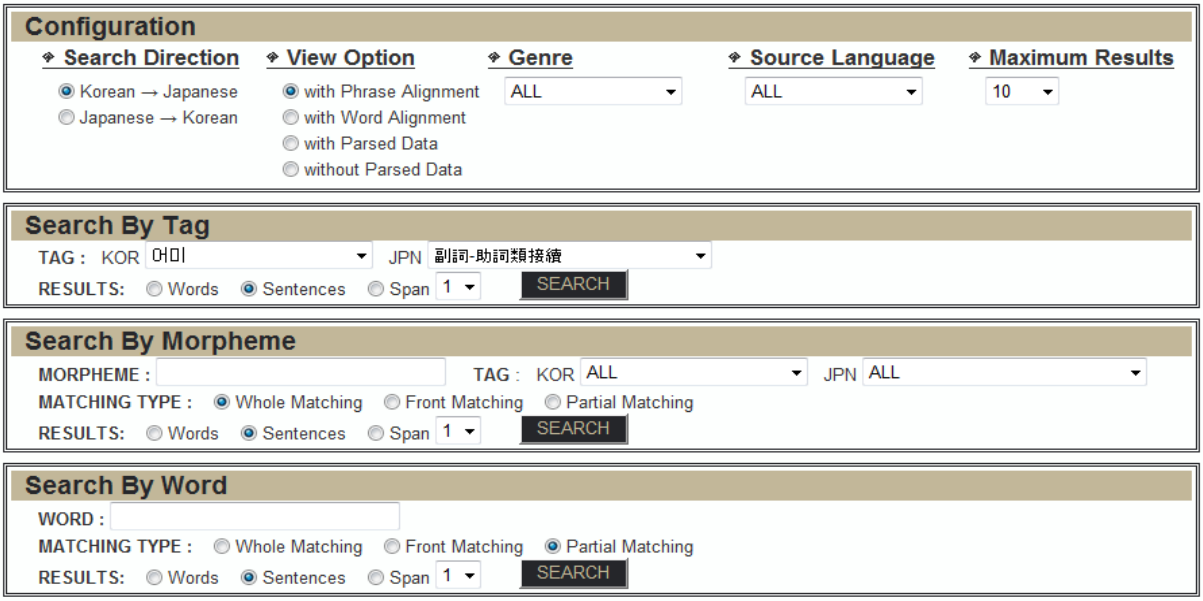

Figure 3. Screenshot of Advanced Search Mode

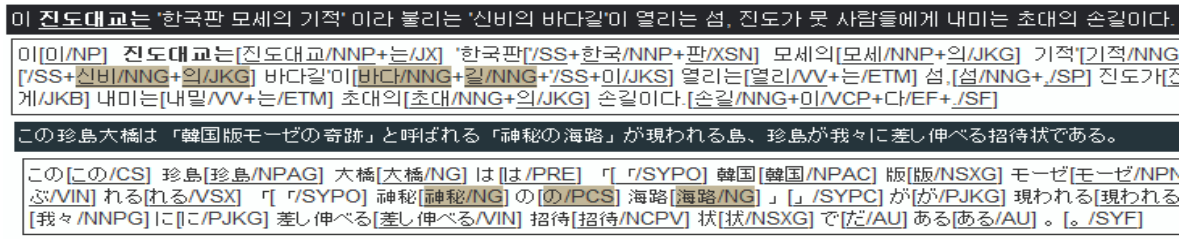

Figure 4. Phrase Alignment for 'a mysterious sea route'

Turning to the output screen, as shown in Figure 4, each underlined word has its corresponding word or phrase. When the pointer is over an underlined word, the system highlights the related words and phrases. If it is necessary to check out more information (e.g. source), one can use 'INFO' buttons. Finally, the interface offers a function to save the current result to a spreadsheet (MS-Excel).

6 Conclusion

Focusing on the *Sejong Korean Japanese Bilingual Corpus* (SKJBC), this paper covers three matters about how to use and show bilingual resources, and provides a user-friendly online interface to search the SKJBC. The NARA interface is applicable to any other bilingual resources in further researches, because it has been designed data-independently. We have already used it for aligned Korean-English text.

Acknowledgments

Part of this work was carried out while the first author was an intern at the NICT Language Infrastructure Group. The fist author was also sponsored by the BK21 Project (Global Internship). We owe special thanks to Prof. Jae-Woong Choe, Prof. Han-Seop Lee, Dr. Dong-Sung Kim, Eric Nichols, Yeolwon Seong, and Inbean Lim.

References

Anne Abeillé. 2003. *Treebanks*. Kluwer Academic Publishers, Hingham, MA, USA.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1): 19-51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation.* Annual Meeting of the ACL.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondˇrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, Alexandra Constantin, Christine Corbett Moran, and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* Annual Meeting of the ACL.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, and Yoshitaka Hirano. 1999. *Japanese Morphological Analysis System ChaSen version 2.0 Manual*. NAIST-ISTR99009.