Proceedings of the ACL 2003 Workshop on Multilingual and
Mixed-language Named Entity Recognition, pp. 17-24.

Construction and Analysis of Japanese-English Broadcast News Corpus with Named Entity Tags

Tadashi Kumano, **Hideki Kashioka** and **Hideki Tanaka**
ATR Spoken Language Translation Research Laboratories
2–2–2, Hikaridai, Keihanna Science City, Kyoto 619–0288, Japan
{tadashi.kumano, hideki.kashioka, hideki.tanaka}@atr.co.jp

Takahiro Fukusima
Otemon Gakuin University
1–15, Nishiai 2-chome, Ibaraki, Osaka 567–8502, Japan
fukusima@res.otemon.ac.jp

Abstract

We are aiming to acquire named entity (NE) translation knowledge from non-parallel, content-aligned corpora, by utilizing NE extraction techniques. For this research, we are constructing a Japanese-English broadcast news corpus with NE tags. The tags represent not only NE class information but also coreference information within the same monolingual document and between corresponding Japanese-English document pairs. Analysis of about 1,100 annotated article pairs has shown that if NE occurrence information, such as classes, number of occurrence and occurrence order, is given for each language, it may provide a good clue for corresponding NEs across languages.

1 Introduction

Studies on named entity (NE) extraction are making progress for various languages, such as English and Japanese. A number of evaluation workshops have been held, including the Message Understanding Conference (MUC)[1] for English and other languages, and the Information Retrieval and Extraction Exercise (IREX)[2] for Japanese. Extraction accuracy for English has reached a nearly practical level (Marsh and Perzanowski, 1998). As for Japanese, it is more difficult to find NE boundaries, however, NE extraction is relatively accurate (Sekine and Isahara, 2000).

Most of the past research on NE extraction used monolingual corpora, but the application of NE extraction techniques to bilingual (or multilingual) corpora is expected to obtain NE translation pairs. We are developing a Japanese-English machine translation system for documents including many NEs, such as news articles or documents about current topics. Translating NE correctly is indispensable for conveying information correctly. NE translations, however, are not listed in conventional dictionaries. It is necessary to retrieve NE translation knowledge from the latest bilingual documents.

When extracting translation knowledge from bilingual corpora, using literally translated parallel corpora, such as official documents written in several languages makes it easier to get the desired information. However, not many of such corpora contain the latest NEs. There are few Japanese-English corpora which are translated literally. Therefore, we decided to extract NE translation pairs from content-aligned corpora, such as multilingual broadcast news articles including new NEs daily, which are not literally translated.

Sentential alignment (Brown et al., 1991; Gale and Church, 1993; Kay and Röscheisen, 1993; Utsuro et al., 1994; Haruno and Yamazaki, 1996) is commonly used as a starting point for finding the translations of words or expressions from bilingual corpora. However, it is not always possible to correspond non-parallel corpora in sentences. Past statistical methods for non-parallel corpora (Fung and Yee, 1998) are not valid for finding translations of

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
[2] http://nlp.cs.nyu.edu/irex/

words or expressions with low frequency. These methods have a problem in covering NEs because there are many NEs that appear only once in a corpus. So we need a specialized method for extracting NE translation pairs. Transliteration is used for finding the translations of NE in the source language from texts in the target language (Stalls and Knight, 1998; Goto et al., 2001; Al-Onazian and Knight, 2002). Transliteration is useful for the names of persons and places; however, it is not applicable to all sorts of NEs.

Content-aligned documents, such as a bilingual news corpus, are made to convey the same topics. Since NEs are the essential element of document contents, content-aligned documents are likely to share NEs pointing to the same objects. Consequently, when extracting all NEs with NE class information from each of a pair of bilingual documents separately by applying monolingual NE extraction techniques, the distribution of the NEs in each document may be similar enough to recognize correspondences between the NE translation pairs.

A technique for finding bilingual NE correspondences will have a wide range of applications other than NE translation-pair extraction. For example,

- Bilingual NE correspondences have clues for identifying corresponding parts in a pair of noisy bilingual documents.
- The similarity of any two documents in different languages can be estimated by NE translation-pair correspondence.

For this research, we obtained a Japanese-English broadcast news corpus (Kumano et al., 2002) by the Japanese broadcast company NHK[3], and we are manually tagging NEs in the corpus to analyze it and to conduct NE translation-pair extraction experiments.

The tag specifications are based on the IREX NE task (Sekine and Isahara, 1999), the evaluation workshop of Japanese NE extraction. We extended the specifications to English NEs. In addition, coreference information between NEs, within the same monolingual document and between the corresponding Japanese-English document pairs (henceforth, we call these *in a language* and *across languages*, respectively), is added to each of the tagged NEs, for NE translation-pair extraction studies.

In Section 2, we will introduce the bilingual corpus used in this study and describe its characteristics. Then, we will discuss tag design for NE extraction studies, and explain the tag specifications and existing problems. The current status of corpus annotation under these specifications will also be introduced. We analyzed an annotated part of the corpus in terms of NE occurrence and translation. This analysis will be shown in Section 3. In Section 4, we will mention future plans for the extraction of NE translation-pairs.

2 Constructing a Japanese-English broadcast news corpus with NE tags

2.1 Characteristics of the NHK Japanese-English broadcast news corpus

We are annotating an NHK broadcast news corpus with NE tags. The corpus is composed of Japanese news articles for domestic programs and English news articles translated for international broadcasting[4] and domestic bilingual programs[5].

Figure 1 shows an example of a Japanese news article and its translation in English. The original Japanese article and the translated English article deal with the same topic, but they differ much in details. The difference arises from the following reasons (Kumano et al., 2002).

Audience Content might be added or deleted, according to the audience, especially for international broadcasting.

Broadcasting date The broadcasting of English news is often delayed compared to the original Japanese news. The time expressions might be changed sometimes or new facts might be added to the articles.

News styles / languages Comparing news articles of two languages reveals that they have different presentation styles, for example, facts are sometimes introduced in a different order. The

[3] *Nippon Hoso Kyokai* (Japan Broadcasting Corporation) (`http://www.nhk.or.jp/englishtop/`)

[4] NHK WORLD (`http://www.nhk.or.jp/nhkworld/`)

[5] `http://www.nhk.or.jp/englishtop/program_list/`

Original article in Japanese (and its literal translation in English by authors):

1: 地震が続いている伊豆諸島できょう午前六時四十二分頃強い地震があり式根島で震度五弱を観測しました。
(There was a strong earthquake at 6:42 this morning in Izu Islands, the site of recent numerous earthquakes. An earthquake of a little less than five in seismic intensity was observed at Shikine Island.)

2: このほか震度四が新島、神津島、震度三が利島、三宅島、また関東各地や静岡県の一部で震度二や一の揺れを観測しました。
(In addition, an event of seismic intensity four was observed for Niijima and Kozu Island, events seismic intensity three for Toshima Island and Miyake Island, and events of seismic intensity two and one for various parts of Kanto Area and Shizuoka Prefecture.)

3: この地震による津波の心配はありません。
(There is no risk of tsunamis resulting from this earthquake.)

4: 気象庁の観測によりますと震源地は新島・神津島の近海で震源の深さは十キロ、地震の規模を示すマグニチュードは五点一と推定されています。
(According to observations by the Meteorological Agency, the earthquake epicenter was located in the sea at a depth of ten kilometers near Niijima and Kozu Island. The magnitude of the earthquakes was estimated to be five point one.)

5: 六月末から地震活動が始まった伊豆諸島では活動が活発な状態とやや落ち着いた状態を繰り返していて、先月三十日も三宅島で震度六弱の強い地震を一回観測した他震度五強の地震が二回起きました。
(In Izu Islands, where seismic activity has been observed from the end of June, repeated cycles of seismic activity and dormancy have been observed. On the 30th of the previous month, a single strong earthquake having seismic intensity of a little less than six was observed at Miyake Island, while two earthquakes having seismic intensity of five were also observed there.)

6: これらの地震を含めて一連の地震活動では神津島や新島、三宅島で震度六弱の強い揺れを四回観測したのを含めてこれまでに震度五弱以上の地震が十七回起きています。
(In a series of seismic events, seventeen earthquakes having seismic intensity over five have been observed up to this point, including strong tremors with a seismic intensity of a little less than six observed four times at Kozu Island, Niijima, and Miyake Island.)

Translated article in English:

1: A strong earthquake jolted Shikine Island, one of the Izu islands south of Tokyo, early on Thursday morning.

2: The Meteorological Agency says the quake measured five-minus on the Japanese scale of seven.

3: The quake affected other islands nearby.

4: Seismic activity began in the area in late July, and 17 quakes of similar or stronger intensity have occurred.

5: Officials are warning of more similar or stronger earthquakes around Niijima and Kozu Islands.

6: Tokyo police say there have been no reports of damage from the latest quake.

Figure 1: An article pair in an NHK broadcast news corpus

difference is due to language and socio-cultural backgrounds.

2.2 NE tag design

We designed NE tags for NE translation-pair extraction research and working efficiency for manual annotation. The specifications are shown below.

- It is desirable that NE recognition guidelines be consistent with NE tags of existing corpora. Past guidelines of MUC and IREX should be respected because they were configured as a result of many discussions. Consistent guidelines enable us to utilize existing annotated corpora and systems designated for the corpora.

- Within each bilingual document pair, coreference between NEs in a language and across languages will be specified. When several NEs exist for the same referent in a document, it is not always possible to determine the actual translation for each instance of the NEs from the counterpart document, because our corpus is not composed of literal translations. Therefore, coreference between NEs in a language should be marked so that the coreference across languages can be assigned between NE groups that have the same referent. Coreference between NE groups is sufficient for our purpose.

- Assignment of coreference in a language is limited between NEs only. Although NEs may have the same referent with pronouns or non-NE expressions, these elements are ignored to avoid complicating the annotation work.

2.3 Tag specifications

1. The tag specifications conform to IREX NE tag specifications (IREX Committee, 1999) (an English description in (Sekine and Isahara, 1999)) as regards the markup form, NE classes, and NE recognition guidelines.

Japanese:

地震が続いている<LOCATION ID="1" COR="2">伊豆諸島(Izu Islands)</LOCATION>で<DATE ID="2" COR="4">きょう(today)</DATE><TIME ID="3" COR="5">午前(a.m.)六時四十二分(6:42)</TIME>頃強い地震があり<LOCATION ID="4" COR="1">式根島(Shikine Island)</LOCATION>で震度五弱を観測しました。 ...

English:

A strong earthquake jolted <LOCATION ID="1" COR="4">Shikine Island</LOCATION>, one of the <LOCATION ID="2" COR="1">Izu islands</LOCATION> south of <LOCATION ID="3">Tokyo</LOCATION>, early on <DATE ID="4" COR="2">Thursday</DATE> <TIME ID="5" COR="3">morning</TIME>. ...

Figure 2: An annotation example

NE Class	Example
Named entities (in the narrow sense):	
ORGANIZATION	The Diet; IREX Committee
PERSON	(Mr.) Obuchi; Wakanohana
LOCATION	Japan; Tokyo; Mt. Fuji
ARTIFACT	Pentium Processor; Nobel Prize
Temporal expressions:	
DATE	September 2, 1999; Yesterday
TIME	11 PM; midnight
Number expressions:	
MONEY	100 yen; $12,345
PERCENT	10%; a half

Table 1: NE Classes

Eight NE classes were defined at the IREX NE task — the same 7 classes as MUC-7 (3 types of named entities in the narrow sense, 2 types of temporal expressions, and 2 types of number expressions), and ARTIFACT (concrete objects like commercial products and abstract objects such as laws or intellectual properties). Table 1 shows a list of these.

2. IREX's NE classes and NE recognition guidelines are applied to English for consistency between Japanese and English NEs. For English-specific annotation, such as prepositions or determiners in NE, the MUC-7 Named Entity Task Definition (Chinchor, 1997) is consulted[6].

3. The SGML markup form of the IREX tag is extended by adding the following two tag attributes, which represent coreference information in a language, and across languages.

   ID="*NE group ID***"** (mandatory)
   Each NE is assigned an attribute ID and an ID number as its value. All coreferent NEs in each language document are given the same ID number[7]. The same ID number is assigned to NEs that have different forms, such as the full name and the first name or the official name and the abbreviated form, in addition to NEs with the same form. Basically, NE are assigned the same ID number when they belong to an NE class and have the identical surface form[8].

   COR="*ID for corresponding NE groups in the other language***"** (optional)
   When there exists a corresponding NE (group) belonging to the same NE class in the other language, an attribute COR is given to each NE (group) in both languages, and the ID number for the counterpart is assigned as a value to each other.

Annotations by the specifications are illustrated in Figure 2.

2.4 Current status of the corpus annotation

Annotators who have experience in translation work and in the production of linguistic data are engaging in the tag annotation. Plans call for a total of 2,000 article pairs to be annotated, and about 1,100 pairs have been finished up to the present.

2.5 Problems

Some problems became obvious in the course of discussions of tag specifications and tag annotation work. They confuse annotators and make the result inaccurate. Typical cases are shown below.

2.5.1 The granularity difference between Japanese and English

In Japanese, a unit smaller than a morpheme may be accepted as an NE according to IREX guidelines.

[6]The tag specifications of IREX NE and those of MUC-7 do not differ radically, because IREX NE tags are designed based on the discussions of MUC.

[7]ID numbers do not maintain uniqueness across the documents.

[8]There are some exceptions. See Section 2.5.3.

(last Sunday and this Sunday)
sensyuu-no nichiyou -to konsyuu-no nichiyou
J: 先週の<DATE ID="1">日曜</DATE>と今週の<DATE ID="2">日曜</DATE>
E: <DATE COR="1">last Sunday</DATE> and <DATE COR="2">this Sunday</DATE>

Figure 3: Assignment of different group IDs with NEs having the same surface form

On the other hand, English does not accept any unit smaller than a word by MUC-7 guidelines. Some Japanese NEs cannot have a counterpart English NE, even if they have a corresponding English expression because of the difference in the segmentation granularity. For example, "アメリカ (*amerika*; America)" in the Japanese morpheme "アメリカ人 (*amerika-jin*; America-people)" is treated as an NE, while no NE can be tagged to "American", the English counterpart of "アメリカ人."

2.5.2 Translation problems

NEs have the same problem that translation in general has: What is the exact translation word(s) for an expression?

- Semantically corresponding expressions may not be assigned corresponding NE relations, because they belong to different NE classes or an expression in a language is not recognized as an NE. For example, a non-NE word "政府 (*seifu*; government)" which means Japanese government in Japanese articles is often translated as the English NE: "Japan."
- A non-literal translation of an NE may cause difficulty in recognizing corresponding relations. Correspondences for some expressions cannot be decided with the information represented in documents: Relative temporal expressions in Japanese are often translated as absolute expressions in English and those correspondences cannot be identified without consulting the calendar; Money expressions are generally converted to dollars and the exchange rate at the relative time is needed to confirm correspondences. For example, we found a translation pair of money expressions "三千億円 (*sanzen-oku-en*; three hundred billion yen)" and "three billion U-S dollars" in our corpus, which constitutes a rough conversion from yen into dollars when the articles were produced.

2.5.3 Assigning NE group IDs

We defined NEs that have the identical surface form and the same NE class to be coreferent and assigned the same NE group ID, in order to make coreference judgment easier. There are some cases where we cannot apply this rule, especially to temporal expressions or number expressions.

The example in Figure 3 shows the translation pair "先週の日曜と今週の日曜 (last Sunday and this Sunday)" and "last Sunday and this Sunday" annotated with NE tags. Japanese temporal expressions "先週の日曜 (last Sunday)" and "今週の日曜 (this Sunday)" are translated into English as "last Sunday" and "this Sunday" respectively. When annotating NE tags for this translation pair, only "日曜 (Sunday)" in those temporal expressions in Japanese is regarded as an NE according to the IREX's NE specifications. This causes a problem in which the two NEs of the same surface form that are assigned the same NE class have different referents. Each of them should assign correspondence to different NEs in the counterpart: the former to "last Sunday" and the latter to "this Sunday."

Tentatively, we allowed a different NE group ID to be assigned to an NE with the identical surface form in an NE class, as shown in Figure 3. It would be better reexamine the consistency of the NE tag specification between Japanese and English, and the necessity of coreference information for temporal expressions and number expressions.

3 Analysis

We conducted an elementary investigation into 1,096 pairs of annotated Japanese and English articles.

3.1 Corpus size

Table 2 shows the content size of our corpus by the number of sentences and the morphemes/words. The content decreases significantly when translating from Japanese to English. This fact points out that

NE class	Japanese		English	
	tokens (avr. per art. / sent.)	types (avr. per art. / sent.)	tokens (avr. per art. / sent.)	types (avr. per art. / sent.)
Total	24,147 (22.03 / 4.13)	12,809 (11.69 / 2.19)	15,844 (14.46 / 2.03)	10,353 (9.45 / 1.32)
ORGANIZATION	5,160 (4.71 / 0.88)	2,558 (2.33 / 0.44)	2,882 (2.63 / 0.37)	1,863 (1.70 / 0.24)
PERSON	3,525 (3.22 / 0.60)	1,628 (1.49 / 0.28)	2,800 (2.55 / 0.36)	1,410 (1.29 / 0.18)
LOCATION	8,737 (7.97 / 1.49)	3,752 (3.42 / 0.64)	5,792 (5.28 / 0.74)	3,302 (3.01 / 0.42)
ARTIFACT	455 (0.42 / 0.08)	282 (0.26 / 0.05)	241 (0.22 / 0.03)	193 (0.18 / 0.02)
DATE	4,342 (3.96 / 0.74)	2,959 (2.70 / 0.51)	2,990 (2.73 / 0.38)	2,620 (2.39 / 0.34)
TIME	854 (0.78 / 0.15)	740 (0.68 / 0.13)	245 (0.22 / 0.03)	232 (0.21 / 0.03)
MONEY	577 (0.53 / 0.10)	462 (0.42 / 0.08)	517 (0.47 / 0.07)	375 (0.34 / 0.05)
PERCENT	497 (0.45 / 0.08)	428 (0.39 / 0.07)	377 (0.34 / 0.05)	358 (0.33 / 0.05)

Table 3: NE frequency

	articles	sentences (avr. per article)	morphemes/words (avr. per sent.)
J	1,096	5,851 (5.34)	321,204 (54.90)
E		7,815 (7.13)	181,180 (23.18)

Table 2: Corpus size

the content tends to be lost through the translation process.

3.2 In-language characteristics of NE occurrences

3.2.1 Frequency

The number of occurrences for each NE class is listed in Table 3. The distribution of NE classes is almost the same as that in the data for MUC-7 or IREX.

By comparing the decrease in content (cf. Table 2), the number of NE tokens also decreases for translations. However, the degree of the NE decrease is less than that of the morphemes/words. It is also remarkable that the number of NE types is fairly well preserved. Notice that only a small number of tokens in the NE class TIME appear in English. The reason may be that detailed time information may become less important for English articles, which are intended for audiences outside of Japan and broadcast later than the original Japanese articles.

3.2.2 NE characteristics within NE groups

To examine the surface form distribution in the same NE groups, we counted the number of members (*freq*) and sorts of surface form (*sort*) for each NE group in each article. The probability that a given member has a unique surface form in a group that has two or more members (*uniq*) has also been calculated as follows:

$$uniq = \frac{freq - 2C sort - 2}{freq - 1C sort - 1} = \frac{sort - 1}{freq - 1} \quad (freq \geq 2).$$

NE class	Japanese			English		
	freq	*sort*	*uniq*	*freq*	*sort*	*uniq*
Average	1.89	1.10	0.131	1.53	1.14	0.332
ORG.	2.02	1.12	0.144	1.55	1.16	0.345
PERSON	2.17	1.12	0.121	1.99	1.49	0.655
LOCATION	2.33	1.14	0.114	1.75	1.07	0.105
ARTIFACT	1.61	1.05	0.072	1.25	1.05	0.216
DATE	1.47	1.08	0.175	1.14	1.03	0.200
TIME	1.15	1.02	0.098	1.06	1.01	0.182
MONEY	1.25	1.03	0.109	1.38	1.35	0.936
PERCENT	1.16	1.00	0.008	1.05	1.06	0.278

Table 4: Surface form distribution in the same NE groups

Table 4 shows the values averaged for all the NE groups that appeared in all articles.

In English, a repetition of the same expression is not conventionally desirable. Therefore, pronouns or paraphrases are used frequently. On the other hand, Japanese does not have such a convention. This difference is considered to be the reason for the result shown in Table 4: *freq* in English is smaller than that in Japanese, and *sort* in English is larger than that in Japanese. As a result, *uniq* in English is higher than that in Japanese. These tendencies differ slightly according to the NE classes.

- The *sort* of English PERSON is notably large. In English, the name of a person is usually first expressed in full, and after that, it tends to be expressed only by the family name. In Japanese, only the family name is generally used from the beginning, especially for well-known persons.

NE class	J → E		J ← E	
	token	type	token	type
Average	0.742	0.639	0.842	0.786
ORGANIZATION	0.684	0.612	0.877	0.837
PERSON	0.881	0.777	0.938	0.898
LOCATION	0.799	0.673	0.833	0.753
ARTIFACT	0.701	0.628	0.925	0.912
DATE	0.717	0.656	0.761	0.742
TIME	0.207	0.184	0.596	0.591
MONEY	0.593	0.595	0.781	0.733
PERCENT	0.712	0.692	0.830	0.827

Table 5: Cross-language corresponding rate

NE class	Japanese			English		
	freq	*sort*	*uniq*	*freq*	*sort*	*uniq*
Average	2.19	1.14	0.134	1.64	1.17	0.342
ORG.	2.25	1.17	0.164	1.62	1.19	0.364
PERSON	2.45	1.14	0.110	2.07	1.53	0.645
LOCATION	2.77	1.19	0.117	1.94	1.10	0.112
ARTIFACT	1.80	1.06	0.075	1.27	1.05	0.222
DATE	1.60	1.10	0.167	1.17	1.04	0.211
TIME	1.30	1.04	0.106	1.07	1.01	0.250
MONEY	1.24	1.04	0.138	1.47	1.43	0.934
PERCENT	1.20	1.00	0.010	1.06	1.01	0.250

Table 6: Surface form distribution in the same NE groups (only for those having cross-language correspondences)

- The *uniq* of English MONEY is quite high. A money expression in Japanese tends to be translated into English as both the original currency (usually yen) and dollars.
- The *freq* of temporal and number expressions are smaller than those of named entities in the narrow sense.

3.3 Cross-language characteristics of NE occurrences

3.3.1 Correspondence across languages

We calculated the rates for a given NE in a document to have a corresponding NE in the counterpart language. The units of NE correspondences we used for these calculations are both NE token and NE group (type). The results, shown in Table 5, show that an NE that appeared in English will have a Japanese NE correspondent with a high rate.

We also conducted the same survey as we did in Table 4 for only NEs having cross-language coreferences, whose results are shown in Table 6. A comparison of both results shows that the *freq* for only NEs having cross-language coreferences is larger, especially in Japanese. An NE occurring more times in an article may have more important information and is more likely to appear in the translation.

NE class	J → E		J ← E	
	All	Corr. only	All	Corr. only
All NEs	0.291	0.774	0.483	0.774
Average	0.304	0.790	0.494	0.790
ORG.	0.269	0.808	0.568	0.809
PERSON	0.403	0.877	0.671	0.875
LOCATION	0.318	0.746	0.461	0.745
ARTIFACT	0.410	0.725	0.662	0.710
DATE	0.307	0.805	0.428	0.805
TIME	0.033	0.815	0.227	0.815
MONEY	0.170	0.829	0.407	0.829
PERCENT	0.509	0.903	0.658	0.903

Table 7: Preservation ratio of NE order

3.3.2 Preservation of NE order

We investigated how well the order of NEs occurring in an article is preserved in the counterpart language as follows:

1. In every article, we eliminated all NEs except the first occurrence of every NE group.
2. We calculated the ratio between all of the possible NE pairs in the source language and those translated into the target language with the same order of occurrence.

Table 7 lists the average preservation ratios of the NE order for all NEs ("All") and for NEs having corresponding NEs in the counterpart ("Corr. only"). The scores labeled "All NEs" express ratios for the order of all NEs. The preservation ratio for each NE class is listed below in the table. The NE orders are preserved so well even for all NEs that they can be used for determining cross-language correspondences.

4 Conclusion

In this paper, in which we aimed to acquire NE translation knowledge, we described our construction of a Japanese-English broadcast news corpus with NE tags for NE translation-pair extraction. The tags represent NE characteristics and coreference information in a language and across languages. Analysis of the annotated 1,097 article pairs has shown that if NE occurrence information, such as classes, number of occurrences and occurrence order, is given for

each language side, it may provide a good clue for determining NE correspondence across languages.

Our future plans are listed below.

- The problems in Section 2.5 need to be reexamined from the point of view of what information bilingual corpora should have for NE translation-pair extraction research.
- The proposed analysis in Section 3 pointed out that identifying coreferences in a language is very important for achieving NE translation-pair extraction. Richer coreference information should be annotated in our corpus for coreference identification studies. We are planning to annotate coreference information for pronouns and some other non-NE expressions, referring to the MUC-7 coreference task definition (Hirschman and Chinchor, 1997).
- Corpora with different characteristics, such as a bilingual newspaper corpus, will be annotated and analyzed.

Acknowledgments This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

Yaser Al-Onazian and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 400–408.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 169–176.

Nancy Chinchor. 1997. MUC-7 named entity task definition. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html`.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, volume I, pages 414–420.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Isao Goto, Noriyoshi Uratani, and Terumasa Ehara. 2001. Cross-language information retrieval of proper nouns using context information. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 571–578.

Masahiko Haruno and Takefumi Yamazaki. 1996. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th International Conference on Computational Linguistics (ACL '96)*, pages 131–138.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 coreference task definition. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`.

IREX Committee. 1999. Named entity extraction task definition (version 990214). `http://nlp.cs.nyu.edu/irex/NE/df990214.txt`. (In Japanese).

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Tadashi Kumano, Isao Goto, Hideki Tanaka, Noriyoshi Uratani, and Terumasa Ehara. 2002. A translation aid system by retrieving bilingual news database. *Systems and Computers in Japan*, 33(8):19–29. (Original written in Japanese is in *Transactions of the Institute of Electronics, Information and Communication Engineers*, J85-D-II(6):1175–1184. 2001).

Elaine Marsh and Dennis Perzanowski. 1998. MUC-7 evaluation of IE technology: Overview and results. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/marsh_slides.pdf`.

Satoshi Sekine and Hitoshi Isahara. 1999. IREX project overview. `http://nlp.cs.nyu.edu/irex/Paper/irex-e.ps`. (Original written in Japanese is in *Proceedings of the IREX Workshop*, pages 1–5).

Satoshi Sekine and Hitoshi Isahara. 2000. IREX: IR and IE evaluation project in Japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pages 1475–1480.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text. In *Proceedings of the Workshop on Computational Approaches of the Semitic Languages*, pages 34–41.

Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. 1994. Bilingual text matching using bilingual dictionary and statistics. In *Proceedings of the 32th International Conference on Computational Linguistics (ACL-94)*, pages 1076–1082.