Using the Web as a Bilingual Dictionary

Masaaki NAGATA
NTT Cyber Space Laboratories
1-1 Hikarinooka, Yokoshuka-shi
Kanagawa, 239-0847 Japan
nagata@nttnly.isl.ntt.co.jp

Teruka SAITO
Chiba University
1-33 Yayoi-cho, Inage-ku
Chiba-shi, Chiba, 263-8522 Japan
t-saito@icsd4.tj.chiba-u.ac.jp

Kenji SUZUKI
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi
Aichi, 441-8580 Japan
ksuzuki@ss.ics.tut.ac.jp

Abstract

We present a system for extracting an English translation of a given Japanese technical term by collecting and scoring translation candidates from the web. We first show that there are a lot of partially bilingual documents in the web that could be useful for term translation, discovered by using a commercial technical term dictionary and an Internet search engine. We then present an algorithm for obtaining translation candidates based on the distance of Japanese and English terms in web documents, and report the results of a preliminary experiment.

1 Introduction

In the field of computational linguistics, the term 'bilingual text' is often used as a synonym for 'parallel text', which is a pair of texts written in two different languages with the same semantic contents. In Asian languages such as Japanese, Chinese and Korean, however, there are a large number of 'partially bilingual texts', in which the monolingual text of an Asian language contains several sporadically interlaced English words as follows:

> "緑内障が早期発見によって管理可能な病気となったため、黄斑変性 (macular degeneration) が先進国の視力障害の主な原因となりつつある。"

The above sentence is taken from a Japanese medical document, which says "Since glaucoma is now manageable if diagnosed early, macular degeneration is becoming a major cause of visual impairment in developed nations". These partially bilingual texts are typically found in technical documents, where the original English technical terms are indicated (usually in parenthesis) just after the first usage of the Japanese technical terms. Even if you don't know Japanese, you can easily guess '黄斑変性' is the translation of 'macular degeneration'.

Partially bilingual texts can be used for machine translation and cross language information retrieval, as well as bilingual lexicon construction, because they not only give a correspondence between Japanese and English terms, but also give the context in which the Japanese term is translated to the English term. For example, the Japanese word '変性' can be translated into many English words, such as 'degeneration', 'denaturation', and 'conversion'. However, the words in the Japanese context such as '病気 (disease)' and '障害 (impairment)' can be used as informants guiding the selection of the most appropriate English word.

In this paper, we investigate the possibility of using web-sourced partially bilingual texts as a continually-updated, wide-coverage bilingual technical term dictionary.

Extracting the English translation of a given Japanese technical term from the web on the fly is different from collecting a set of arbitrary many pairs of English and Japanese technical terms. The former can be thought of example-based

translation, while the latter is a tool for bilingual lexicon construction.

Internet portals are starting to provide on-line bilingual dictionary and translation services. However, technical terms and new words are unlikely to be well covered because they are too specific or too new. The proposed term translation extractor could be an useful Internet tool for human translators to complement the weakness of existing on-line dictionaries and translation services.

In the following sections, we first investigate the coverage provided by partially bilingual texts in the web as discovered by using a commercial technical term dictionary and an Internet search engine. We then present a simple algorithm for extracting English translation candidates of a given Japanese technical term. Finally, we report the results of a preliminary experiment and discuss future work.

2 Partially Bilingual Text in the Web

2.1 Coverage of Fields

It is very difficult to measure precisely in what field of science there are a large number of partially bilingual text in the web. However, it is possible to get a rough estimate on the relative amount in different fields, by asking a search engine for documents containing both Japanese and English technical terms in each field several times.

For this purpose, we used a Japanese-to-English technical term dictionary licensed from NOVA, a maker of commercial machine translation systems. The dictionary is classified into 19 categories, ranging from aeronautics to ecology to trade, as shown in Table 1. There are 1,082,594 pairs of Japanese and English technical terms[1].

We randomly selected 30 pairs of Japanese and English terms from each category and sent queries to an Internet search engine, Google (Google, 2001), to see whether there are any documents that contain both Japanese and English technical terms. The fourth column in Table 1 shows the percentage of queries (J-E pairs) returned by at least one document.

[1]The dictionary can be searched in their web site (NOVA Inc., 2000).

It is very encouraging that, on average, 42% of the queries returned at least one document. The results show that the web is worth mining for bilingual lexicon, in fields such as aeronautics, computer, and law.

2.2 Classification of Format

In order to implement a term translation extractor, we have to analyze the format, or structural pattern of the partially bilingual documents. There are at least three typical formats in the web. Figure 1 shows examples.

- aligned paragraph format
- table format
- plain text format

In ‘aligned paragraph’ format, each paragraph contains one language and the paragraphs with different languages are interlaced. This format is often found in web pages designed for both Japanese and foreigners, such as official documents by governments and academic papers by researchers (usually title and abstract only).

In ‘table’ format, each row contains a pair of equivalent terms. They are not necessarily marked by the TABLE tag of HTML. This format is often found in bilingual glossaries of which there are many in the web. Some portals offer hyper links to such bilingual glossaries, such as kotoba.ne.jp (kotoba.ne.jp, 2000).

In ‘plain text’ format, phrases of different language are interlaced in the monolingual text of the baseline language. The vast majority of partially bilingual documents in the web belongs to this category.

The formats of the web documents are so wildly different that it is impossible to automatically classify them to estimate the relative quantities belonging to each format. Instead, we examined the distance (in bytes) from a Japanese technical term to its corresponding English technical term in the documents retrieved from the web by the experiment described in the Section 2.1

Figure 2 shows the results. Positive distance indicates that the English term appeared after the Japanese term, while negative distance indicates the reverse. It is observed that the English and Japanese terms are likely to appear very close to

Registration for Foreign Residents and Birth Registration
がいこくじんとうろく しゅっせいとど
外国人登録と 出生届け
The official name for registration for foreign residents in Japan, as determined by the Ministry of Justice, is “Alien Registration”.
...
Anyone staying in Japan for more than 90 days, children born in Japan, ...
90日以上日本に滞在するとき、子供が日本で生まれたとき、...
...
(http://www.pref.akita.jp/life/g090.htm)

(a) An example of ‘aligned paragraph format’ taken from a life guide for foreigners.

日本救急医学会・用語集1(和英)

...
【あ】
喘ぎ呼吸 gasping respiration
アカラシア achalasia
亜急性細菌性心内膜炎 subacute bacterial endocarditis
...
【い】
胃 stomach
胃液 gastric juice
異化 catabolism
...
(http://apollo.m.ehime-u.ac.jp/GHDNet/98/waei.html)

(b) An example of ‘table format’ taken from a medical glossary.

●温暖化とは
温暖化とは、人間の活動が活発になるにつれて「温室効果ガス」が大気中に大量に放出され、地球全体の平均気温が急激に上がり始めている現象のことをいいます。大気中に微量に含まれる二酸化炭素（CO2）、メタン（CH4）、亜酸化窒素（N2O）、フロンなどが、温室効果ガス（Green House Gases：GHGs）といわれています。
...
(http://www.eic.or.jp/cop3/ondan/ondan.html)

(c) An example of ‘plain text format’ taken from a document on global worming.

Figure 1: Three typical formats of partially bilingual documents in the web

Table 1: The percentage of documents including both Japanese and English words

fields	words	samples	found	Example of Japanese-English pair
aeronautics and space	17862	30	57%	黄道座標 ecliptic coordinates
architecture	32049	30	30%	耐荷力 load capacity
biotechnology	59766	30	50%	系統発生学 phylogeny
business	50201	30	57%	カラ売り short selling
chemicals	122232	30	43%	ギ酸メチル methyl formate
computers	117456	30	57%	ＯＳローダー OS loader
defense	4787	30	17%	識別特性 signature
ecology	32440	30	40%	永久凍土層 permafrost
electronics	87942	30	47%	内接歯車ポンプ internal gear pump
energy	15804	30	50%	サイクロトロン加熱 cyclotron heating
finance	57097	30	37%	営業経費 operating expenses
law	36033	30	60%	保証人 sponsor
math and physics	76304	30	40%	変形エネルギー deformation energy
mechanical engineering	86371	30	30%	正方晶系 tetragonal system
medical	135158	30	27%	整形外科学 orthopedics
metals	25595	30	37%	電解加工 electrochemical machining
ocean	13215	30	43%	係留運転 mooring trial
(industrial) plant	95756	30	53%	自動製図装置 plotter
trade	16526	30	20%	採算価格 remunerative price
total	1082594	570	42%	

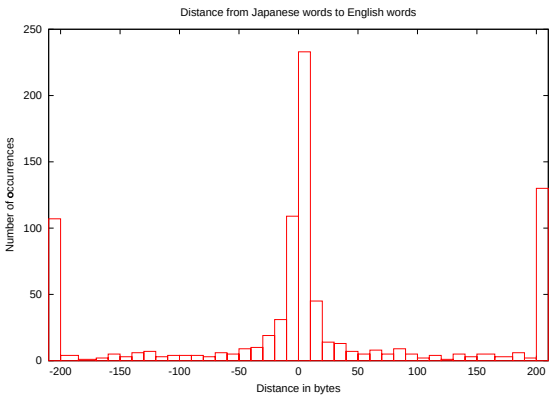

Figure 2: Distance from Japanese terms to English terms

each other. 28% (=233/847) of English terms appeared just after (within 10 bytes) the corresponding Japanese terms. 58% (=490/847) of English terms appeared within ± 50 bytes. They probably reflect either table or plain text format.

Although there are 28% (=237/847) English terms appeared outside the window of ± 200 bytes, we find this 'distance heuristics' very powerful, so it was used in the term translation algorithm described in the next section.

3 Term Translation Extraction Algorithm

Let j and e be Japanese and English technical terms which are translations of each other. Let d be a document, and let $D(j)$ be a set of documents which includes the Japanese term j. Let $P_t(j, e)$ be a statistical translation model which gives the likelihood (or score) that j and e are translations of each other.

Figure 3 shows the basic (conceptual) algorithm for extracting the English translation of a given Japanese technical term from the web. First, we retrieve all documents $D(j)$ that contain the

```
1  foreach d in D(j)
2    if d is a bilingual document then
3      foreach e in d
4        compute P_t(j, e)
5      end
6    endif
7  end
8  output ê = arg max P_t(j, e)
```

Figure 3: Conceptual algorithm for extracting English translation of Japanese term

given Japanese technical term j using a search engine. We then eliminate the Japanese only documents. For each English term e contained in the (partially) bilingual documents, we compute the translation probability $P_t(j, e)$, and select the English term $\hat{e}$ which has the highest translation probability.

In practise, it is often prohibitive to down load all documents that include the Japanese term. Moreover, a reliable Japanese-English statistical translation model is not available at the moment because of the scarcity of parallel corpora. Rather, one of the aim of this research is to collect the resources for building such translation models. We therefore employed a very simplistic approach.

Instead of using all documents including the Japanese term, we used only the predetermined number of documents (top 100 documents based on the rank given by the search engine). This entails the risk of missing the documents including the English terms we are looking for.

Instead of using a statistical translation model, we used a scoring function in the form of a geometric distribution as shown in Equation (1).

$$w(j, e) = p(1 - p)^{floor(d(j,e)/10)} \qquad (1)$$

Here, $d(j, e)$ is the byte distance between Japanese term j and English term e. It is divided by 10 and the integer part of the quotient is used as the variable in the geometric distribution ($floor$ indicates flooring operation). The parameter (the average) of the geometric distribution p is set to 0.6 in our experiment.

There is no theoretical background to the scoring function Equation (1). It was designed, after a trial and error, so that the likelihood of candidates pairs being translations of each other decreases exponentially as the distance between the two terms increases. Starting from the score of 0.6, it decreases 40% for every 10 bytes.

If we observed the same pair of Japanese and English terms more than once, it is more likely that they are valid translations. Therefore, we sum the score of Equation (1) for each occurrence of pair (j, e) and select the highest scoring English term $\hat{e}$ as the translation of the Japanese term j.

Table 3: Term translation extraction accuracy tested by 34 Japanese terms

rank	exact		partial-1		partial-2	
1	15%	(5)	15%	(5)	18%	(6)
5	29%	(10)	29%	(19)	41%	(14)
10	47%	(16)	53%	(18)	62%	(21)
50	56%	(19)	71%	(24)	79%	(27)
all	62%	(21)	76%	(26)	91%	(31)

4 Experiments

4.1 Test Terms

In order to factor out the characteristics of the search engine and the proposed term extraction algorithm, we used, as a test set, those words that are guaranteed to have at lease one retrieved document that includes both Japanese and English terms.

First, we randomly selected 50 pairs of such Japanese and English terms, from the pairs used in the experiment described in Section 2.1. They are shown in Figure 2. We then sent each Japanese term as a query to an Internet search engine, Google, and down loaded the top 100 web documents. “o” indicates that at least one of the down loaded documents included both terms. “x” indicates that no document included both terms. This resulted in a test set of 34 pairs of Japanese and English terms.

For example, although there are a lot of documents which include both “西” and “west”, the top 100 documents retrieved by “西” as the query did not contain “west” since “西” is a highly frequent Japanese word.

Table 2: A list of Japanese and English technical terms used in the experiment.

o 国家情報基盤	National Information Infrastructure	x 比強度	specific strength
o 地球型惑星	terrestrial planet	o アースケーブル	earth cable
o 耐荷力	load capacity	o テヌアゾン酸	tenuazonic acid
o 多因子	multiple factor	o 動物行動学	ethology
o 放射性核種	radionuclide	o ジョブショップスケジューリング	job shop scheduling
o 政府印刷局	Government Printing Office	o 発射装置	launcher
x 支出報告	expense reporting	o ギ酸メチル	methyl formate
o ネットワークゲーム	network game	o ウォーゲーム	war game
o フェニックス	Phoenix	x 西	west
x 立冬	first day of winter	o サイクルタイム	cycle time
o 半二重回線	half duplex circuit	o 市場研究	market research
o 内接歯車ポンプ	internal gear pump	o 閉じたループ	closed loop
o サイクロトロン加熱	cyclotron heating	x 営業経費	operating expenses
x 福祉	well-being	o 世界市場	world market
x 信義	faith	o 法廷	courtroom
x 法律専門書	treatise	x 保証人	sponsor
o アドレス	address	x 気候研究	climate study
o 地磁気逆転	geomagnetic reversal	x かど	edge
o 密度	density	o 終動脈	end artery
o 整形外科学	orthopedics	x 製鋼プロセス	steelmaking process
x 握り	knob	o 係留運転	mooring trial
o 低圧タービン	low pressure turbine	o 豆コック	petcock
x 控え	stay	o 航行システム	navigation system
x 全圧	total pressure	o 借方	debit
x 外国為替相場	foreign exchange rate	o 光ファイバー	optical fiber

4.2 Extraction Accuracy

Table 3 shows the extraction accuracy of the English translation of Japanese term. Since both Japanese and English terms could occur as a subpart of more longer terms, we need to consider local alignment to extract the English subpart corresponding to the Japanese query. Instead of doing this alignment, we introduced two partial match measures as well as exact matching.

In Table 3, ‘exact’ indicates that the output is exactly matched to the correct answer, while ‘partial-1’ indicates that the correct answer was a subpart of the output; ‘partial-2’ indicates that at least one word of the output is a subpart of the correct answer.

For example, the eye disease ‘黄斑変性’, whose translation is ‘macular degeneration’, is sometimes more formally refereed to as ‘加齢性黄斑変性’, whose translation is ‘age-related macular degeneration’. ‘Partial-1’ holds if ‘age-related macular degeneration’ is extracted when the query is ‘黄斑変性’. ‘Partial-2’ holds if ‘degeneration’ is included in the output when the query is ‘黄斑変性’.

It is encouraging that useful outputs (either exact or partial matches) are included in the top 10 candidates with the probability of around 60%. Since we used simple string matching to measure the accuracy automatically, the evaluation reported in Table 3 is very conservative. Because the output contains acronyms, synonyms, and related words, the overall performance of the system is fairly credible.

For example, the extracted translations for the query ‘国家情報基盤’ (National Information Infrastructure) were as follows, where the second candidate is the correct answer.

```
18.721123: nii
13.912146: national informa-
tion infrastructure
2.137008: gii
1.398144: unii
```

NII (nii) is the acronym for National Information Infrastructure, while GII (gii) and UNII (unii) stand for Global Information Infrastructure and Unlicensed National Information Infrastructure, respectively.

If the query is a chemical substance, its molecular formula, instead of acronym, is often extracted, such as ‘HCOOCH3’ for ‘ギ酸メチル’ (methyl formate).

```
1.801008: methyl formate
0.840786: hcooch3
0.84: hcooh
```

As for synonyms, although we took ‘operating expenses’ to be the correct translation for ‘営業経費’, the following third candidate ‘operating cost’ is also a legitimate translation. This is counted as ‘partial-2’ because ‘operating’ is a subpart of the correct answer.

```
1.8: fa
0.606144: ohr
0.6: operating cost
```

For your information, OHR (Over Head Ratio) is a management index and equals to the operating cost divided by the gross operating profit. ‘Fa’ happened to be used three times in a tutorial document on accounting to stand for ‘operating expenses’, such as “営業経費 (Fa)=原価 (E)*23%”, where ‘原価’ means ‘cost’.

The following example is a combination of the acronyms, synonyms and related words, which is, in a sense, a typical output of the proposed system. The query is ‘気候研究’, and ‘climate study’ is the translation we assumed to be correct.

```
10.736611: wcrp
2.282483: wmo
1.220275: no
1.2: wc rp
0.72: igbp
0.6: sparc
0.6: wcp
0.6: applied climatology
0.2784: world climate research programme
```

A subpart of the 9th candidate ‘climate research’ is also a legitimate translation. ‘WCRP’ is the acronym for ‘World Climate Research Programme’, which is the 9th candidate and is translated to ‘世界気候研究計画’ which includes the original Japanese query. ‘WMO’ stands for World Meteorological Organization, which hosts this international program.

In short, if you look at the extracted translations together with the context from which they are extracted, you can learn a lot about the relevant information of the query term and its translation candidates. We think this is a useful tool for human translators, and it could provide a useful resource for statistical machine translation and cross language information retrieval.

5 Discussion and Related Works

Previous studies on bilingual text mainly focused on either parallel texts, non-parallel texts, or comparable texts, in which a pair of texts are written in two different languages (Veronis, 2000). However, except for governmental documents from Canada (English/French) and Hong Kong (Chinese/English), bilingual texts are usually subject to such limitations as licensing conditions, usage fees, domains, language pairs, etc. One approach that partially overcomes these limitations is to collect parallel texts from the web (Nie et al., 1999; Resnik, 1999).

To provide better coverage with fewer restrictions, we focused on partially bilingual text. Considering the enormous volume of such texts and the variety of fields covered, we believe they are the best resource to mine for MT-related applications that involve English and Asian languages.

The current system for extracting the translation of a given term is more similar to the information extraction system for term descriptions (Fujii and Ishikawa, 2000) than any other machine translation systems. In order to collect descriptions for technical term X, such as ‘data mining’, (Fujii and Ishikawa, 2000) collected phrases like “X is Y” and “X is defined as Y”, from the web. As our system used a scoring function based solely on byte distance, introducing this kind of pattern matching might improve its accuracy.

Practically speaking, the factor that most influences the accuracy of the term translation extractor is the set of documents returned from the search engine. In order to evaluate the system, we used a test set that guarantees to contain at least one document with both the Japanese term and its English translation; this is a rather optimistic assumption.

Since the search engine is an uncontrollable factor, one possible solution is to make your own search engine. We are very interested in combining such ideas as focused crawling (Chakrabarti et al., 1999) and domain-specific Internet portals (McCallum et al., 2000) with the proposed term translation extractor to develop a domain-specific on-line dictionary service.

6 Conclusion

We investigated the possibility of using the web as a bilingual dictionary, and reported the preliminary results of an experiment on extracting the English translations of given Japanese technical terms from the web.

One interesting approach to extending the current system is to introduce a statistical translation model (Brown et al., 1993) to filter out irrelevant translation candidates and to extract the most appropriate subpart from a long English sequence as the translation by locally aligning the Japanese and English sequences.

Unlike ordinary machine translation which generates English sentences from Japanese sentences, this is a recognition-type application which identifies whether or not a Japanese term and an English term are translations of each other. Considering the fact that what the statistical translation model provides is the joint probability of Japanese and English phrases, this could be a more natural and prospective application of statistical translation model than sentence-to-sentence translation.

References

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

Soumen Chakrabarti, Martin van den Berg, and Byron Dom. 1999. Focused crawling: a new approach to topic-specific web resource. In *Proceedings of the Eighth International World Wide Web Conference*, pages 545–562.

Atsushi Fujii and Tetsuya Ishikawa. 2000. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeging of the Association for Computational Linguistics*, pages 488–495.

Google. 2001. Google. `http://www.google.com`.

kotoba.ne.jp. 2000. Translators' internet resources (in Japanese). `http://www.kotoba.ne.jp`.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81.

NOVA Inc. 2000. Technical term dictionary lookup service (in Japanese). `http://wwwd.nova.co.jp/webdic/webdic.html`.

Rhilip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534.

Jean Veronis, editor. 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*, volume 13 of *Text, Speech, and Language Technology*. Kluwer Academic Publishers.