Compiling French-Japanese Terminologies from the Web

Xavier Robitaille†, Yasuhiro Sasaki†, Masatsugu Tonoike†,
Satoshi Sato‡ and Takehito Utsuro†

†Graduate School of Informatics,
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto 606-8501 Japan

‡Graduate School of Engineering,
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya 464-8603 Japan

{xavier, sasaki, tonoike, utsuro}@pine.kuee.kyoto-u.ac.jp,
ssato@nuee.nagoya-u.ac.jp

Abstract

We propose a method for compiling bilingual terminologies of multi-word terms (MWTs) for given translation pairs of seed terms. Traditional methods for bilingual terminology compilation exploit parallel texts, while the more recent ones have focused on comparable corpora. We use bilingual corpora collected from the web and tailor made for the seed terms. For each language, we extract from the corpus a set of MWTs pertaining to the seed's semantic domain, and use a compositional method to align MWTs from both sets. We increase the coverage of our system by using thesauri and by applying a bootstrap method. Experimental results show high precision and indicate promising prospects for future developments.

1 Introduction

Bilingual terminologies have been the center of much interest in computational linguistics. Their applications in machine translation have proven quite effective, and this has fuelled research aiming at automating terminology compilation. Early developments focused on their extraction from parallel corpora (Daille et al. (1994), Fung (1995)), which works well but is limited by the scarcity of such resources. Recently, the focus has changed to utilizing comparable corpora, which are easier to obtain in many domains. Most of the proposed methods use the fact that words have comparable contexts across languages. Fung (1998) and Rapp (1999) use so called context vector methods to extract translations of general words. Chiao and Zweigenbaum (2002) and Déjean and Gaussier (2002) apply similar methods to technical domains. Daille and Morin (2005) use specialized comparable corpora to extract translations of multi-word terms (MWTs).

These methods output a few thousand terms and yield a precision of more or less 80% on the first 10-20 candidates. We argue for the need for systems that output fewer terms, but with a higher precision. Moreover, all the above were conducted on language pairs including English. It would be possible, albeit more difficult, to obtain comparable corpora for pairs such as French-Japanese. We will try to remove the need to gather corpora beforehand altogether. To achieve this, we use the web as our only source of data. This idea is not new, and has already been tried by Cao and Li (2002) for base noun phrase translation. They use a compositional method to generate a set of translation candidates from which they select the most likely translation by using empirical evidence from the web.

The method we propose takes a translation pair of seed terms in input. First, we collect MWTs semantically similar to the seed in each language. Then, we work out the alignments between the MWTs in both sets. Our intuition is that both seeds have the same related terms across languages, and we believe that this will simplify the alignment process. The alignment is done by generating a set of translation candidates using a compositional method, and by selecting the most probable translation from that set. It is very similar to Cao and Li's, except in two respects. First, the generation makes use of thesauri to account for lexical divergence between MWTs in the source and target language. Second, we validate candidate translations using a set of terms collected from the web, rather than

using empirical evidence from the web as a whole. Our research further differs from Cao and Li's in that they focus only on finding valid translations for given base noun phrases. We attempt to both collect appropriate sets of related MWTs and to find their respective translations.

The initial output of the system contains 9.6 pairs on average, and has a precision of 92%. We use this high precision as a bootstrap to augment the set of Japanese related terms, and obtain a final output of 19.6 pairs on average, with a precision of 81%.

2 Related Term Collection

Given a translation pair of seed terms (s_f, s_j), we use a search engine to gather a set F of French terms related to s_f, and a set J of Japanese terms related to s_j. The methods applied for both languages use the framework proposed by Sato and Sasaki (2003), outlined in Figure 1. We proceed in three steps: corpus collection, automatic term recognition (ATR), and filtering.

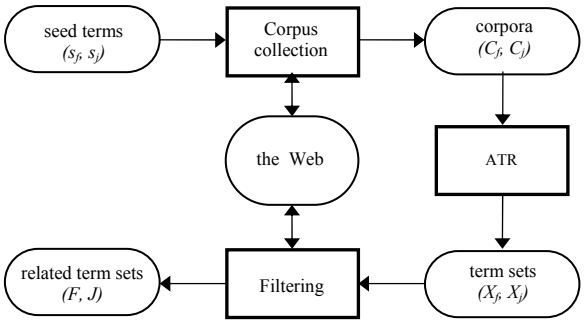

Figure 1: Related term collection

2.1 Corpus Collection

For each language, we collect a corpus C from web pages by selecting passages that contain the seed.

Web page collection

In French, we use Google to find relevant web pages by entering the following three queries: "s_f", "s_f *est*" (s_f is), and "s_f *sont*" (s_f are). In Japanese, we do the same with queries "s_j", "s_jとは", "s_jは", "s_jという", and "s_jの", where とは *toha*, は *ha,* という *toiu*, and の *no* are Japanese functional words that are often used for defining or explaining a term. We retrieve the top pages for each query, and parse those pages looking for hyperlinks whose anchor text contain the seed. If such links exist, we retrieve the linked pages as well.

Sentence extraction

From the retrieved web pages, we remove html tags and other noise. Then, we keep only properly structured sentences containing the seed, as well as the preceding and following sentences – that is, we use a window of three sentences around the seed.

2.2 Automatic Term Recognition

The next step is to extract candidate related terms from the corpus. Because the sentences composing the corpus are related to the seed, the same should be true for the terms they contain. The process of extracting terms is highly language dependent.

French ATR

We use the C-value method (Frantzi and Ananiadou (2003)), which extracts compound terms and ranks them according to their termhood. It consists of a linguistic part, followed by a statistical part.

The **linguistic part** consists in applying a linguistic filter to constrain the structure of terms extracted. We base our filter on a morphosyntactic pattern for the French language proposed by Daille et al. It defines the structure of multi-word units (MWUs) that are likely to be terms. Although their work focused on MWUs limited to two content words (nouns, adjectives, verbs or adverbs), we extend our filter to MWUs of greater length. The pattern is defined as follows:

$$(Noun|Num)(Adj|Prep\,Det^{?}(Noun|Num))^{+}$$

The **statistical part** measures the termhood of each compound that matches the linguistic pattern. It is given by the C-value:

$$\text{C-value}(a) = \begin{cases} \log_2|a|\,\mathrm{f}(a) & \text{if } a \text{ is not nested}, \\ \log_2|a|\,\mathrm{f}(a)\left(\mathrm{f}(a) - \dfrac{\sum_{b\in T_a}\mathrm{f}(b)}{\mathrm{P}(T_a)}\right) & \text{otherwise} \end{cases}$$

where a is the candidate string, $f(a)$ is its frequency of occurrence in all the web pages retrieved, T_a is the set of extracted candidate terms that contain a, and $P(T_a)$ is the number of these candidate terms.

The nature of our variable length pattern is such that if a long compound matches the pattern, all the shorter compounds it includes also match. For example, consider the N-Prep-N-

Prep-N structure in *système à base de connaissances* (knowledge based system). The shorter candidate *système à base* (based system) also matches, although we would prefer not to extract it.

Fortunately, the strength of the C-value is the way it effectively handles nested MWTs. When we calculate the termhood of a string, we subtract from its total frequency its frequency as a substring of longer candidate terms. In other words, a shorter compound that almost always appears nested in a longer compound will have a comparatively smaller C-value, even if its total frequency is higher than that of the longer compound. Hence, we discard MWTs whose C-value is smaller than that of a longer candidate term in which it is nested.

Japanese ATR

Because compound nouns represent the bulk of Japanese technical MWTs, we extract them as candidate related terms. As opposed to Sato and Sasaki, we ignore single nouns. Also, we do not limit the number of candidates output by ATR as they did.

2.3 Filtering

Finally, from the output set of ATR, we select only the technical terms that are part of the seed's semantic domain. Numerous measures have been proposed to gauge the semantic similarity between two words (van Rijsbergen (1979)). We choose the Jaccard coefficient, which we calculate based on search engine hit counts. The similarity between a seed term s and a candidate term x is given by:

$$Jac = \frac{H(s \wedge x)}{H(s \vee x)}$$

where $H(s \wedge x)$ is the hit count of pages containing both s and x, and $H(s \vee x)$ is the hit count of pages containing s or x. The latter can be calculated as follows:

$$H(s \vee x) = H(s) + H(x) - H(s \wedge x)$$

Candidates that have a high enough coefficient are considered related terms of the seed.

3 Term Alignment

Once we have collected related terms in both French and Japanese, we must link the terms in the source language to the terms in the target language. Our alignment procedure is twofold. First, we first generate Japanese translation candidates for each collected French term. Second, we select the most likely translation(s) from the set of candidates. This is similar to the generation and selection procedures used in the literature (Baldwin and Tanaka (2004), Cao and Li, Langkilde and Knight (1998)).

3.1 Translation Candidates Generation

Translation candidates are generated using a compositional method, which can be divided in three steps. First, we decompose the French MWTs into combinations of shorter MWU elements. Second, we look up the elements in bilingual dictionaries. Third, we recompose translation candidates by generating different combinations of translated elements.

Decomposition

In accordance with Daille et al., we define the length of a MWU as the number of content words it contains. Let n be the length of the MWT to decompose. We produce all the combinations of MWU elements of length less or equal to n. For example, consider the French translation of "knowledge based system":

système	à	base	de	connaissances
Noun	Prep	Noun	Prep	Noun

It has a length of three and yields the following four combinations[1]:

[système à base de connaissances**]**
[système**]** **[**base de connaissances**]**
[système à base**]** **[**connaissances**]**
[système**]** **[**base**]** **[**connaissances**]**

Note the treatment given to the prepositions and determiners: we leave them in place when they are interposed between content words within elements, otherwise we remove them.

Dictionary Lookup

We look up each element in bilingual dictionaries. Because some words appear in their inflected forms, we use their lemmata. In the example given above, we look up *connaissance* (lemma) rather than *connaissances* (inflected). Note that we do not lemmatize MWUs such as *base de connaissances*. This is due to the complexity of gender and number agreements of French compounds. However, only a small part of the MWTs are collected in their inflected forms, and French-Japanese bilingual dictionaries do not contain that many MWTs to begin with. The performance hit should therefore be minor.

Already at this stage, we can anticipate problems arising from the insufficient coverage of

[1] A MWT of length n produces 2^{n-1} combinations, including itself.

French-Japanese lexicon resources. Bilingual dictionaries may not have enough entries, and existing entries may not include a great variety of translations for every sense. The former problem has no easy solution, and is one of the reasons we are conducting this research. The latter can be partially remedied by using thesauri – we augment each element's translation set by looking up in thesauri all the translations obtained with bilingual dictionaries.

Recomposition

To recompose the translation candidates, we simply generate all suitable combinations of translated elements for each decomposition. The word order is inverted to take into account the different constraints in French and Japanese. In the example above, if the lookup phase gave {知識 *chishiki*}, {土台 *dodai*, ベース *besu*} and {体系 *taikei*, システム *shisutemu*} as respective translation sets for *système, base* and *connaissance*, the fourth decomposition given above would yield the following candidates:

connaissance	base	système
知識	土台	体系
知識	土台	システム
知識	ベース	体系
知識	ベース	システム

If we do not find any translation for one of the elements, the generation fails.

3.2 Translation Selection

Selection consists of picking the most likely translation from the translation candidates we have generated. To discern the likely from the unlikely, we use the empirical evidence provided by the set of Japanese terms related to the seed. We believe that if a candidate is present in that set, it could well be a valid translation, as the French MWT in consideration is also related to the seed. Accordingly, our selection process consists of picking those candidates for which we find a complete match among the related terms.

3.3 Relevance of Compositional Methods

The automatic translation of MWTs is no simple task, and it is worthwhile asking if it is best tackled with a compositional method. Intricate problems have been reported with the translations of compounds (Daille and Morin, Baldwin and Tanaka), notably:

- **fertility:** source and target MWTs can be of different lengths. For example, *table de vérité* (truth table) contains two content words and translates into 真理･値･表 *shinri・chi・hyo* (lit. truth-value-table), which contains three.
- **variability of forms in the translations:** MWTs can appear in many forms. For example, *champ electromagnétique* (electromagnetic field) translates both into 電磁･場 *denji・ba* (lit. electromagnetic field)電磁･界 *denji・kai* (lit. electromagnetic "region").
- **constructional variability in the translations:** source and target MWTs have different morphological structures. For example, in the pair *apprentissage automatique*↔ 機 械･学 習 *kikai・gakushu* (machine learning) we have (N-Adj)↔(N-N). In the pair *programmation par contraintes*↔パターン･認識 *patan・ninshiki* (pattern recognition) we have (N-par-N)↔(N-N).
- **non-compositional compounds:** some compounds' meaning cannot be derived from the meaning of their components. For example, the Japanese term 赤･点 *aka・ten* (failing grade, lit. "red point") translates into French as *note d'échec* (lit. failing grade) or simply *échec* (lit. failure).
- **lexical divergence:** source and target MWTs can use different lexica to express a concept. For example, *traduction automatique* (machine translation, lit. "automatic translation") translates as 機 械･翻 訳 *kikai・honyaku* (lit. machine translation).

It is hard to imagine any method that could address all these problems accurately.

Tanaka and Baldwin (2003) found that 48.7% of English-Japanese Noun-Noun compounds translate compositionality. In a preliminary experiment, we found this to be the case for as much as 75.1% of the collected MWTs. If we are to maximize the coverage of our system, it is sensible to start with a compositional approach. We will not deal with the problem of fertility and non-compositional compounds in this paper. Nonetheless, lexical divergence and variability issues will be partly tackled by broader translations and related words given by thesauri.

Id	French	Japanese	(English)
1	analyse vectorielle	ベクトル·解析 *bekutoru ·kaiseki*	(vector analysis)
2	circuit logique	論理·回路 *ronri ·kairo*	(logic circuit)
3	intelligence artificielle	人工·知能 *jinko ·chinou*	(artificial intelligence)
4	linguistique informatique	計算·言語学 *keisan ·gengogaku*	(computational linguistics)
5	reconnaissance des formes	パターン·認識 *patan ·ninshiki*	(pattern recognition)
6	reconnaissance vocale	音声·認識 *onsei ·ninshiki*	(speech recognition)
7	science cognitive	認知·科学 *ninchi ·kagaku*	(cognitive science)
8	traduction automatique	機械·翻訳 *kikai ·honyaku*	(machine translation)

Table 1: Seed pairs

4 Evaluation

4.1 Linguistic Resources

The bilingual dictionaries used in the experiments are the Crown French-Japanese Dictionary (Ohtsuki et al. (1989)), and the French-Japanese Scientific Dictionary (French-Japanese Scientific Association (1989)). The former contains about 50,000 entries of general usage single words. The latter contains about 50,000 entries of both single and multi-word scientific terms. These two complement each other, and by combining both entries we form our base dictionary to which we refer as Dic_{FJ}.

The main thesaurus used is *Bunrui Goi Hyo* (National Institute for Japanese Language (2004)). It contains about 96,000 words, and each entry is organized in two levels: a list of synonyms and a list of more loosely related words. We augment the initial translation set by looking up the Japanese words given by Dic_{FJ}. The expanded bilingual dictionary comprised of the words from Dic_{FJ} combined with their synonyms is denoted Dic_{FJJ}. The dictionary resulting of Dic_{FJJ} combined with the more loosely related words is denoted Dic_{FJJ2}.

Finally, we build another thesaurus from a Japanese-English dictionary. We use Eijiro (Electronic Dictionary Project (2004)), which contains 1,290,000 entries. For a given Japanese entry, we look up its English translations. The Japanese translations of the English intermediaries are used as synonyms/related words of the entry. The resulting thesaurus is expected to provide even more loosely related translations (and also many irrelevant ones). We denote it Dic_{FJEJ}.

4.2 Notation

Let F and J be the two sets of related terms collected in French and Japanese. F' is the subset of F for which $Jac \geq 0.01$:

$$F' = \left\{ f \in F \middle| Jac(f) \geq 0.01 \right\}$$

F'^* is the subset of valid related terms in F', as determined by human evaluation. P is the set of all potential translation pairs among the collected terms ($P=F\times J$). P' is the set of pairs containing either a French term or a Japanese term with $Jac \geq 0.01$:

$$P' = \left\{ (f \in F, j \in J) \middle| Jac(f) \geq 0.01 \vee Jac(j) \geq 0.01 \right\}$$

P'^* is the subset of valid translation pairs in P', determined by human evaluation. These pairs need to respect three criteria: 1) contain valid terms, 2) be related to the seed, and 3) constitute a valid translation. M is the set of all translations selected by our system. M' is the subset of pairs in M with $Jac \geq 0.01$ for either the French or the Japanese term. It is also the output of our system:

$$M' = \left\{ (f, j) \in M \middle| Jac(f) \geq 0.01 \vee Jac(j) \geq 0.01 \right\}$$

M'^* is the intersection of M' and P'^*, or in other words, the subset of valid translation pairs output by our system.

Set	|M'|	|M'*|	Prec.	Recall
FJ	10.5	9.6	92%	40%
FJJ	15.3	12.6	83%	53%
FJJ2	20.5	13.4	65%	56%
FJEJ	30.9	14.1	46%	59%

Table 2: Results for the baseline

4.3 Baseline Method

Our starting point is the simplest possible alignment, which we refer to as our baseline. It is worked out by using each of the aforementioned dictionaries independently. The output set obtained using Dic_{FJ} is denoted *FJ*, the one using Dic_{FJJ} is denoted *FJJ*, and so on. The experiment is made using the eight seed pairs given in Table 1. On average, we have $|F'| = 74.3$, $|F'^*| = 51.0$ and $|P'^*| = 24.0$. Table 2 gives a summary of the key results. The precision and the recall are given by:

$$precision = \frac{|M'^*|}{|M'|}, \quad recall = \frac{|M'^*|}{|P'^*|}$$

Dic_{FJ} contains only Japanese translations corresponding to the strict sense of French elements. Such a dictionary generates only a few translation candidates which tend to be correct when present in the target set. On the other hand, the lookup in Dic_{FJJ2} and Dic_{FJEJ} interprets French

Set	|M'|	|M'*|	Prec.	Recall
FJJ′	14.0	12.3	88%	51%
FJJ2′	16.1	12.8	79%	53%
FJEJ′	29.1	15.5	53%	65%

Table 3: Results for the incremental selection

MWT elements with more laxity, generating more translations and thus more alignments, at the cost of some precision.

4.4 Incremental Selection

The progressive increase in recall given by the increasingly looser translations is in inverse proportion to the decrease in precision, which hints that we should give precedence to the alignments obtained with the more accurate methods. Consequently, we start by adding the alignments in *FJ* to the output set. Then, we augment it with the alignments from *FJJ* whose terms are not already in *FJ*. The resulting set is denoted *FJJ′*. We then augment *FJJ′* with the pairs from *FJJ2* whose terms are not in *FJJ′*, and so on, until we exhaust the alignments in *FJEJ*.

For instance, let *FJ* contain (*synthèse de la parole*↔ 音 声 · 合 成 *onsei · gousei* (speech synthesis)) and *FJJ* contain this pair plus (*synthèse de la parole*↔音声·解析 *onsei·kaiseki* (speech analysis)). In the first iteration, the pair in *FJ* is added to the output set. In the second iteration, no pair is added because the output set already contains an alignment with *synthèse de la parole*.

Table 3 gives the results for each incremental step. We can see an increase in precision for *FJJ′*, *FJJ2′* and *FJEJ′* of respectively 5%, 9% and 8%, compared to *FJJ*, *FJJ2* and *FJEJ*. We are effectively filtering output pairs and, as expected, the increase in precision is accompanied by a slight decrease in recall. Note that, because *FJEJ* is not a superset of *FJJ2*, we see an increase in both precision and recall in *FJEJ′* over *FJEJ*. Nonetheless, the precision yielded by *FJEJ′* is not sufficient, which is why Dic_{FJEJ} is left out in the next experiment.

4.5 Bootstrapping

The coverage of the system is still shy of the 20 pairs/seed objective we gave ourselves. One cause for this is the small number of valid translation pairs available in the corpora. From an average of 51 valid related terms in the source set, only 24 have their translation in the target set. To counter that problem, we increase the coverage of Japanese related terms and hope that by doing so, we will also increase the coverage of the system as a whole.

Set	|M'|	|M'*|	Prec.
FJ^+	20.9	16.8	80%
FJJ^+	30.9	21.3	69%
$FJJ2^+$	45.8	22.6	49%

Table 4: Results for the baseline method with bootstrap expansion

Set	|M'|	|M'*|	Prec.
$FJ^+{}'$	19.5	16.1	83%
$FJJ^+{}'$	22.5	18.6	83%
$FJJ^+{}''$	24.3	19.6	81%
$FJJ2^+{}'$	25.6	20.1	79%
$FJJ2^+{}''$	28.6	20.6	72%

Table 5: Results for the incremental selection with bootstrap expansion

Once again, we utilize the high precision of the baseline method. The average 10.5 pairs in *FJ* include 92% of Japanese terms semantically similar to the seed. By inputting these terms in the term collection system, we collect many more terms, some of which are probably the translations of our French MWTs.

The results for the baseline method with bootstrapping are given in Table 4. The ones using incremental selection and bootstrapping are given in Table 5. FJ^+ consists of the alignments given by a generation process using Dic_{FJ} and a selection performed on the augmented set of related terms. FJJ^+ and $FJJ2^+$ are obtained in the same way using Dic_{FJJ} and Dic_{FJJ2}. $FJ^+{}'$ contains the alignments from FJ, augmented with those from FJ^+ whose terms are not in *FJ*. $FJJ^+{}'$ contains $FJ^+{}'$, incremented with terms from *FJJ*. $FJJ^+{}''$ contains $FJJ^+{}'$, incremented with terms from FJJ^+, and so on.

The bootstrap mechanism grows the target term set tenfold, making it very laborious to identify all the valid translation pairs manually. Consequently, we only evaluate the pairs output by the system, making it impossible to calculate recall. Instead, we use the number of valid translation pairs as a makeshift measure.

Bootstrapping successfully allows for many more translation pairs to be found. FJ^+, FJJ^+, and $FJJ2^+$ respectively contain 7.6, 8.7 and 8.5 more valid alignments on average than *FJ*, *FJJ* and *FJJ2*. The augmented target term set is noisier than the initial set, and it produces many more invalid alignments as well. Fortunately, the incremental selection effectively filters out most of the unwanted, restoring the precision to acceptable levels.

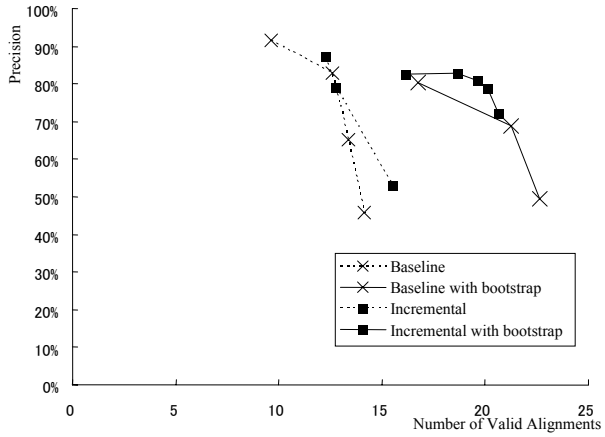

Figure 2: Precision - Valid Alignments curves

4.6 Analysis

A comparison of all the methods is illustrated in the precision – valid alignments curves of Figure 2. The points on the four curves are taken from Tables 2 to 5. The gap between the dotted and filled curves clearly shows that bootstrapping increases coverage. The respective positions of the squares and crosses show that incremental selection effectively filters out erroneous alignments. *FJJ*$^{+}$*″*, with 19.6 valid alignments and a precision of 81%, is at the rightmost and uppermost position in the graph. The detailed results for each seed are presented in Table 6, and the complete output for the seed "logic circuit" is given in Table 7.

From the average 4.7 erroneous pairs/seed, 3.2 (68%) were correct translations but were judged unrelated to the seed. This is not surprising, considering that our set of French related terms contained only 69% (51/74.3) of valid related terms. Also note that, of the 24.3 pairs/seed output, 5.25 are listed in the French-Japanese Scientific Dictionary. However, only 3.9 of those pairs are included in *M′**. The others were deemed unrelated to the seed.

In the output set of "machine translation", 自然·言語·処理 *shizen·gengo·shori* (natural language processing) is aligned to both *traitement du language naturel* and *traitement des langues naturelles*. The system captures the term's variability around *langue/language*. Lexical divergence is also taken into account to some extent. The seed computational linguistics yields the alignment of *langue maternelle* (mother tongue) with 母国·語 *bokoku·go* (literally [[mother-country]-language]). The usage of thesauri enabled the system to include the concept of country in the translated MWT, even though it is not present in any of the French elements.

seed	|F'|	|F'*|	|P'*|	|M'|	|M'*|	Prec.
1	89	40	14	26	13	50%
2	64	55	24	14	14	100%
3	72	59	38	40	33	83%
4	67	49	22	23	18	78%
5	85	70	22	21	17	81%
6	67	50	27	22	21	95%
7	36	27	16	20	17	85%
8	114	58	29	28	24	86%
avg	74.3	51.0	24.0	24.3	19.6	81%

Table 6: Detailed results for *FJJ*$^{+}$*″*

5 Conclusion and future work

We have proposed a method for compiling bilingual terminologies of compositionally translated MWTs. As opposed to previous work, we use the web rather than comparable corpora as a source of bilingual data. Our main insight is to constrain source and target candidate MWTs to only those strongly related to the seed. This allows us to achieve term alignment with high precision. We showed that coverage reaches satisfactory levels by using thesauri and bootstrapping.

Due to the difference in objectives and in corpora, it is very hard to compare results: our method produces a rather small set of highly accurate alignments, whereas extraction from comparable corpora generates much more candidates, but with an inferior precision. These two approaches have very different applications. Our method does however eliminate the requirement of comparable corpora, which means that we can use seeds from any domain, provided we have reasonably rich dictionaries and thesauri.

Let us not forget that this article describes only a first attempt at compiling French-Japanese terminology, and that various sources of improvement have been left untapped. In particular, our alignment suffers from the fact that we do not discriminate between different candidate translations. This could be achieved by using any of the more sophisticated selection methods proposed in the literature. Currently, corpus features are used solely for the collection of related terms. These could also be utilized in the translation selection, which Baldwin and Tanaka have shown to be quite effective. We could also make use of bilingual dictionary features as they did. Lexical context is another resource we have not exploited. Context vectors have successfully been applied in translation selection by Fung as well as Daille and Morin.

On a different level, we could also apply the bootstrapping to expand the French set of related terms. Finally, we are investigating the possibil-

Jac (Fr.)	French term	Japanese term	(English)	eval†
0.100	portes logiques	論理･ゲート *ronri･geeto*	(logic gate)	2/2/2
0.064	fonctions logiques	論理･関数 *ronri･kansuu*	(logic function)	2/2/2
0.064	fonctions logiques	論理･機能 *ronri･kinou*	(logic function)	2/2/2
0.048	registre à décalage	シフト･レジスタ *shifuto･rejisuta*	(shift register)	2/2/2
0.044	simulateur de circuit	回路･シミュレータ *kairo･shimureeta*	(circuit simulator)	2/2/2
0.040	circuit combinatoire	組合せ･回路 *kumiawase･kairo*	(combinatorial circuit)	2/2/2
0.031	nombre binaire	2･進数 *ni･shinsuu*	(binary number)	2/2/2
0.024	niveaux logiques	論理･レベル *ronri･reberu*	(logical level)	2/2/2
0.020	circuit logique combinatoire	組合せ･論理･回路 *kumiawase･ronri･kairo*	(combinatorial logic circuit)	2/2/2
0.017	valeur logique	論理･値 *ronri･chi*	(logical value)	2/2/2
0.013	tension d' alimentation	電源･電圧 *dengen･denatsu*	(supply voltage)	2/2/2
0.011	conception de circuits	回路･設計 *kairo･sekkei*	(circuit design)	2/2/2
0.007	conception d' un circuit logique	論理･回路･設計 *ronri･kairo･sekkei*	(logic circuit design)	2/1/2
0.005	nombre de portes	ゲート･数 *geeto･suu*	(number of gates)	2/1/2

† relatedness / termhood / quality of the translation, on a scale of 0 to 2

Table 7: System output for seed pair *circuit logique* ↔論理回路 (logic circuit)

ity of resolving the alignments in the opposite direction: from Japanese to French. Surely the constructional variability of French MWTs would present some difficulties, but we are confident that this could be tackled using translation templates, as proposed by Baldwin and Tanaka.

References

T. Baldwin and T. Tanaka. 2004. Translation by Machine of Complex Nominals: Getting it Right. In *Proc. of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pp. 24–31, Barcelona, Spain.

Y. Cao and H. Li. 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proc. of COLING -02*, Taipei, Taiwan.

Y.C. Chiao and P. Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proc. of COLING-02*, pp. 1208–1212. Taipei, Taiwan.

B. Daille, E. Gaussier, and J.M. Lange. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proc. of COLING-94*, pp. 515–521, Kyoto, Japan.

B. Daille and E. Morin. 2005. French-English Terminology Extraction from Comparable Corpora, In *IJCNLP-05*, pp. 707–718, Jeju Island, Korea.

H. Déjean., E. Gaussier and F. Sadat. An Approach Based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In *Proc. of COLING-02*, pp. 218–224. Taipei, Taiwan.

Electronic Dictionary Project. 2004. Eijiro Japanese-English Dictionary: version 79. EDP.

K.T. Frantzi, and S. Ananiadou. 2003. The C-Value/NC-Value Domain Independent Method for Multi-Word Term Extraction. *Journal of Natural Language Processing*, 6(3), pp. 145–179.

French Japanese Scientific Association. 1989. French-Japanese Scientific Dictionary: 4th edition. Hakusuisha.

P. Fung. 1995. A Pattern Matching Method for Finding Noun and Proper Noun from Noisy Parallel Corpora. In *Proc of the ACL-95*, pp. 236–243, Cambridge, USA.

P. Fung. 1998. A Statiscal View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In D. Farwell, L. Gerber and L. Hovy eds.: Proceedings of the AMTA-98, Springer, pp. 1–16.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In COLLING/ACL-98, pp. 704–710, Montreal, Canada.

National Institute for Japanese Language. 2004. Bunrui Goi Hyo: revised and enlarged edition Dainippon Tosho.

T. Ohtsuki et al. 1989. Crown French-Japanese Dictionary: 4th edition. Sanseido.

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In Proc. of the ACL-99. pp. 1–17. College Park, USA.

S. Sato and Y. Sasaki. 2003. Automatic Collection of Related Terms from the Web. In *ACL-03 Companion Volume to the Proc. of the Conference*, pp. 121–124, Sapporo, Japan.

T. Tanaka and T. Baldwin. 2003. Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 17–24. Sapporo, Japan.

van Rijsbergen, C.J. 1979. *Information Retrieval*. London: Butterworths. Second Edition.