Proceedings of the ACL 2003 Workshop on Multiword Expressions:
Analysis, Acquisition and Treatment, pp. 17-24.

Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing

Takaaki Tanaka
Communication Science Laboratories
Nippon Telephone and Telegraph Corporation
Kyoto, Japan
takaaki@cslab.kecl.ntt.co.jp

Timothy Baldwin
CSLI
Stanford University
Stanford, CA 94305 USA
tbaldwin@csli.stanford.edu

Abstract

The translation of compound nouns is a major issue in machine translation due to their frequency of occurrence and high productivity. Various shallow methods have been proposed to translate compound nouns, notable amongst which are memory-based machine translation and word-to-word compositional machine translation. This paper describes the results of a feasibility study on the ability of these methods to translate Japanese and English noun-noun compounds.

1 Introduction

Multiword expressions are problematic in machine translation (MT) due to the idiomaticity and overgeneration problems (Sag et al., 2002). **Idiomaticity** is the problem of compositional semantic unpredictability and/or syntactic markedness, as seen in expressions such as *kick the bucket* (= die′) and *by and large*, respectively. **Overgeneration** occurs as a result of a system failing to capture idiosyncratic lexical affinities between words, such as the blocking of seemingly equivalent word combinations (e.g. *many thanks* vs. **several thanks*). In this paper, we target the particular task of the Japanese↔English machine translation of noun-noun compounds to outline the various techniques that have been proposed to tackle idiomaticity and overgeneration, and carry out detailed analysis of their viability over naturally-occurring data.

Noun-noun (NN) compounds (e.g. *web server*, *car park*) characteristically occur with high frequency and high lexical and semantic variability. A summary examination of the 90m-word written component of the British National Corpus (BNC, Burnard (2000)) unearthed over 400,000 NN compound types, with a combined token frequency of 1.3m;[1] that is, over 1% of words in the BNC are NN compounds. Moreover, if we plot the relative token coverage of the most frequently-occurring NN compound types, we find that the low-frequency types account for a significant proportion of the type count (see Figure 1[2]). To achieve 50% token coverage, e.g., we require coverage of the top 5% most-frequent NN compounds, amounting to roughly 70,000 types with a minimum token frequency of 10. NN compounds are especially prevalent in technical domains, often with idiosyncratic semantics: Tanaka and Matsuo (1999) found that NN compounds accounted for almost 20% of entries in a Japanese-English financial terminological dictionary.

Various claims have been made about the level of processing complexity required to translate NN compounds, and proposed translation methods range over a broad spectrum of processing complexity. There is a clear division between the proposed methods based on whether they attempt to interpret the semantics of the NN compound (i.e. use **deep** processing), or simply use the source language word forms to carry out the translation task (i.e. use **shallow** processing). It is not hard to find examples of semantic mismatch in NN compounds to motivate deep translation methods: the Japanese 井戸端 · 会議 *idobata·kaigi* "(lit.) well-side meeting",[3] e.g., translates most naturally into English as "idle gossip", which a shallow method would be hard put to predict. Our interest is in the relative occurrence of such NN compounds and their impact on the performance of shallow translation methods. In particular, we seek to determine what proportion of NN compounds shallow translation translation methods can reasonably translate and answer the question: do shallow methods perform well enough to preclude the need for deep processing? The answer to this question takes the form of an estimation of the upper bound on translation performance for shallow translation methods.

In order to answer this question, we have selected the language pair of English and Japanese, due to the high linguistic disparity between the two languages. We consider the tasks of both English-to-Japanese (EJ) and Japanese-to-English (JE) NN compound translation over fixed datasets of NN compounds, and apply representative shallow MT methods to the data.

[1] Results based on the method described in § 3.1.

[2] The graph for Japanese NN compounds based on the Mainichi Corpus is almost identical.

[3] With all Japanese NN compound examples, we explicitly segment the compound into its component nouns through the use of the "·" symbol.

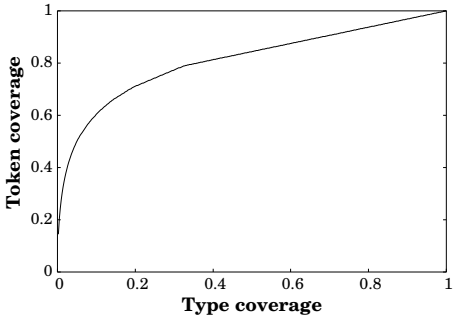

Figure 1: Type vs. token coverage (English)

While stating that English and Japanese are highly linguistically differentiated, we recognise that there are strong syntactic parallels between the two languages with respect to the compound noun construction. At the same time, there are large volumes of subtle lexical and expressional divergences between the two languages, as evidenced between 自転車 · 選手 *jiteNsha·seNshu* "(lit.) bicycle athelete" and its translation *competitive cyclist*. In this sense, we claim that English and Japanese are representative of the inherent difficulty of NN compound translation.

The remainder of this paper is structured as follows. In § 2, we outline the basic MT strategies that exist for translating NN compounds, and in § 3 we describe the method by which we evaluate each method. We then present the results in § 4, and analyse the results and suggest an extension to the basic method in § 5. Finally, we conclude in § 6

2 Methods for translating NN compounds

Two basic paradigms exist for translating NN compounds: memory-based machine translation and dynamic machine translation. Below, we discuss these two paradigms in turn and representative instantiations of each.

2.1 Memory-based machine translation

Memory-based machine translation (MBMT) is a simple and commonly-used method for translating NN compounds, whereby translation pairs are stored in a static **translation database** indexed by their source language strings. MBMT has the ability to produce consistent, high-quality translations (conditioned on the quality of the original bilingual dictionary) and is therefore suited to translating compounds in closed domains. Its most obvious drawback is that the method can translate only those source language strings contained in the translation database.

There are a number of ways to populate the translation database used in MBMT, the easiest of which is to take translation pairs directly from a bilingual dictionary (**dictionary-driven MBMT** or $\mathbf{MBMT_{DICT}}$). $\text{MBMT}_{\text{DICT}}$ offers an extremist solution to the idiomaticity problem, in treating all NN compounds as being fully lexicalised. Overgeneration is not an issue, as all translations are manually determined.

As an alternative to a precompiled bilingual dictionary, translation pairs can be extracted from a **parallel corpus** (Fung, 1995; Smadja et al., 1996; Ohmori and Higashida, 1999), that is a bilingual document set that is translation-equivalent at the sentence or paragraph level; we term this MT configuration **alignment-driven MBMT** (or $\mathbf{MBMT_{ALIGN}}$). While this method alleviates the problem of limited scalability, it relies on the existence of a parallel corpus in the desired domain, which is often an unreasonable requirement.

Whereas a parallel corpus assumes translation equivalence, a **comparable corpus** is simply a crosslingual pairing of corpora from the same domain (Fung and McKeown, 1997; Rapp, 1999; Tanaka and Matsuo, 1999; Tanaka, 2002). It is possible to extract translation pairs from a comparable corpus by way of the following process (Cao and Li, 2002):

1. extract NN compounds from the source language corpus by searching for NN bigrams (e.g. 機械 · 翻訳 *kikai·hoNyaku* "machine translation")

2. compositionally generate translation candidates for each NN compound by accessing translations for each component word and slotting these into translation templates; example JE translation templates for source Japanese string $[N_1\ N_2]_{\mathbf{J}}$ are $[N_1\ N_2]_{\mathbf{E}}$ and $[N_2\ \textit{of}\ N_1]_{\mathbf{E}}$, where the numeric subscripts indicate word coindexation between Japanese and English (resulting in, e.g., *machine translation* and *translation of machine*)

3. use empirical evidence from the target language corpus to select the most plausible translation candidate

We term this process **word-to-word compositional MBMT** (or $\mathbf{MBMT_{COMP}}$). While the coverage of $\text{MBMT}_{\text{COMP}}$ is potentially higher than $\text{MBMT}_{\text{ALIGN}}$ due to the greater accessibility of corpus data, it is limited to some degree by the coverage of the simplex translation dictionary used in Step 2 of the translation process. That is, only those NN compounds whose component nouns occur in the bilingual dictionary can be translated.

Note that both $\text{MBMT}_{\text{ALIGN}}$ and $\text{MBMT}_{\text{COMP}}$ lead to a static translation database. $\text{MBMT}_{\text{COMP}}$ is also subject to overgeneration as a result of dynamically generating translation candidates.

2.2 Dynamic machine translation

Dynamic machine translation (DMT) is geared towards translating arbitrary NN compounds. In this paper, we consider two methods of dynamic translation: word-to-word compositional DMT and interpretation-driven DMT.

Word-to-word compositional DMT (or **$\mathbf{DMT_{COMP}}$**) differs from $MBMT_{COMP}$ only in that the source NN compounds are fed directly into the system rather than extracted out of a source language corpus. That is, it applies Steps 2 and 3 of the method for $MBMT_{COMP}$ to an arbitrary source language string.

Interpretation-driven DMT (or **$\mathbf{DMT_{INTERP}}$**) offers the means to deal with NN compounds where strict word-to-word alignment does not hold. It generally does this in two stages:

1. use semantics and/or pragmatics to carry out deep analysis of the source NN compound, and map it into some intermediate (i.e. interlingual) semantic representation (Copestake and Lascarides, 1997; Barker and Szpakowicz, 1998; Rosario and Hearst, 2001)

2. generate the translation directly from the semantic representation

DMT_{INTERP} removes any direct source/target language interdependence, and hence solves the problem of overgeneration due to crosslingual bias. At the same time, it is forced into tackling idiomaticity head-on, by way of interpreting each individual NN compound. As for DMT_{COMP}, DMT_{INTERP} suffers from undergeneration.

With DMT_{INTERP}, context must often be called upon in interpreting NN compounds (e.g. *apple juice seat* (Levi, 1978; Bauer, 1979)), and minimal pairs with sharply-differentiated semantics such as *colour/group photograph* illustrate the fine-grained distinctions that must be made. It is interesting to note that, while these examples are difficult to interpret, in an MT context, they can all be translated word-to-word compositionally into Japanese. That is, *apple juice seat* translates most naturally as アップルジュース・の・席 *appurujuusu·no·seki* "apple-juice seat",[4] which retains the same scope for interpretation as its English counterpart; similarly, *colour photograph* translates trivially as カラー・写真 *karaa·shashiN* "colour photograph" and *group photograph* as 団体・写真 *daNtai·shashiN* "group photograph". In these cases, therefore, DMT_{INTERP} offers no advantage over DMT_{COMP}, while incurring a sizeable cost in producing a full semantic interpretation.

3 Methodology

We selected the tasks of Japanese-to-English and English-to-Japanese NN compound MT for evaluation, and tested $MBMT_{DICT}$ and DMT_{COMP} on each task. Note that we do not evaluate $MBMT_{ALIGN}$ as results would have been too heavily conditioned on the makeup of the parallel corpus and the particular alignment method adopted. Below, we describe the data and method used in evaluation.

[4]Here, *no* is the genitive marker.

3.1 Testdata

In order to generate English and Japanese NN compound testdata, we first extracted out all NN bigrams from the BNC (90m word tokens, Burnard (2000)) and 1996 Mainichi Shimbun Corpus (32m word tokens, Mainichi Newspaper Co. (1996)), respectively. The BNC had been tagged and chunked using fnTBL (Ngai and Florian, 2001), and lemmatised using morph (Minnen et al., 2001), while the Mainichi Shimbun had been segmented and tagged using ALT-JAWS.[5] For both English and Japanese, we took only those NN bigrams adjoined by non-nouns to ensure that they were not part of a larger compound nominal. In the case of English, we additionally measured the entropy of the left and right contexts for each NN type, and filtered out all compounds where either entropy value was < 1.[6] This was done in an attempt to, once again, exclude NNs which were embedded in larger MWEs, such as *service department* in *social service department*.

We next extracted out the 250 most common NN compounds from the English and Japanese data, and from the remaining data, randomly selected a further 250 NN compounds of frequency 10 or greater (out of 20,748 English and 169,899 Japanese NN compounds). In this way, we generated a total of 500 NN compounds for each of English and Japanese. For the Japanese NN compounds, any errors in segmentation were post-corrected. Note that the top-250 NN compounds accounted for about 7.0% and 3.3% of the total token occurrences of English and Japanese NN compounds, respectively; for the random sample of 250 NN compounds, the relative occurrence of the English and Japanese compounds out of the total token sample was 0.5% and 0.1%, respectively.

We next generated a unique gold-standard translation for each of the English and Japanese NN compounds. In order to reduce the manual translation overhead and maintain consistency with the output of $MBMT_{DICT}$ in evaluation, we first tried to translate each English and Japanese NN compound automatically by $MBMT_{DICT}$. In this, we used the union of two Japanese-English dictionaries: the ALTDIC dictionary and the on-line EDICT dictionary (Breen, 1995). The **ALTDIC** dictionary was compiled from the ALT-J/E MT system (Ikehara et al., 1991), and has approximately 400,000 entries including more than 200,000 proper nouns; EDICT has approximately 150,000 entries. In the case that multiple translation candidates were found for a given NN compound, the most appropriate of these was selected manually, or in the case that the dictionary translations were considered

[5]`http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html`

[6]For the left token entropy, if the most-probable left context was *the*, *a* or a sentence boundary, the threshold was switched off. Similarly for the right token entropy, if the most-probable right context was a punctuation mark or sentence boundary, the threshold was switched off.

Templates (JE)	*Examples*	*#*
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_1\ N_2]_{\mathbf{E}}$	市場 · 経済 *shijou·keizai* "market economy"	83
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [Adj_1\ N_2]_{\mathbf{E}}$	医療 · 機関 *iryou·kikaN* "medical institution"	71
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_1\ Np_2]_{\mathbf{E}}$	調査 · 結果 *chousa·kekka* "survey results"	14
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_2$ *of (the)* $N_1]_{\mathbf{E}}$	政権 · 交替 *seikeN·koutai* "change of government"	11
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_2$ *of (the)* $Np_1]_{\mathbf{E}}$	意見 · 交換 *ikeN·koukaN* "exchange of ideas"	8
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [Adj_1\ Np_2]_{\mathbf{E}}$	経済 · 制裁 *keizai·seisai* "economic sanctions"	8
Templates (EJ)	***Examples***	***#***
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_1\ N_2]_{\mathbf{J}}$	*exchange rate* 為替 · レート "*kawase·reeto*"	192
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_1$ *no* $N_2]_{\mathbf{J}}$	*hotel room* ホテル · の · 部屋 "*hoteru·no·heya*"	20
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_2\ N_1]_{\mathbf{J}}$	*carbon dioxide* 二酸化 · 炭素 "*nisaNka·taNso*"	1

Table 1: Example translation templates (*N* = noun (base), *Np* = noun (plural), and *Adj* = adjective)

to be sub-optimal or inappropriate, the NN compound was put aside for manual translation. Finally, all dictionary-based translations were manually checked for accuracy.

The residue of NN compounds for which a translation was not found were translated manually. Note that as we manually check all translations, the accuracy of $\text{MBMT}_{\text{DICT}}$ is less than 100%. At the same time, we give $\text{MBMT}_{\text{DICT}}$ full credit in evaluation for containing an optimal translation, by virtue of using the dictionaries as our primary source of translations.

3.2 Upper bound accuracy-based evaluation

We use the testdata to evaluate $\text{MBMT}_{\text{DICT}}$ and DMT_{COMP}. Both methods potentially produce multiple translations candidates for a given input, from which a unique translation output must be selected in some way. So as to establish an upper bound on the feasibility of each method, we focus on the translation candidate generation step in this paper and leave the second step of translation selection as an item for further research.

With $\text{MBMT}_{\text{DICT}}$, we calculate the upper bound by simply checking for the gold-standard translation within the translation candidates. In the case of DMT_{COMP}, rather than generating all translation candidates and checking among them, we take a predetermined set of translation templates and a simplex translation dictionary to test for word alignment. Word alignment is considered to have been achieved if there exists a translation template and set of word translations which lead to an isomorphic mapping onto the gold-standard translation. For 領土 · 問題 *ryoudo·moNdai* "territorial dispute", for example, alignment is achieved through the word-level translations 領土 *ryoudo* "territory" and 問題 *moNdai* "dispute", and the mapping conforms to the $[N_1\ N_2]_{\mathbf{J}} \Rightarrow [Adj_1\ N_2]_{\mathbf{E}}$ translation template. It is thus possible to translate 領土 · 問題 by way of DMT_{COMP}. Note here that derivational morphology is used to convert the nominal translation of *territory* into the adjective *territorial*.

On the first word-alignment pass for DMT_{COMP}, the translation pairs in each dataset were automatically aligned using only ALTDIC. We then manual inspected the unaligned translation pairs for translation pairs which were not aligned simply because of patchy coverage in ALTDIC. In such cases, we manually supplemented ALTDIC with simplex translation pairs taken from the Genius Japanese-English dictionary (Konishi, 1997),[7] resulting in an additional 178 simplex entries. We then performed a second pass of alignment using the supplemented ALTDIC (**ALTDIC**+). Below, we present the results for both the original ALTDIC and ALTDIC+.

3.3 Learning translation templates

DMT_{COMP} relies on translation templates to map the source language NN compound onto different constructions in the target language and generate translation candidates. For the JE task, the question of what templates are used becomes particularly salient due to the syntactic diversity of the gold standard English translations (see below). Rather than assuming a manually-specified template set for the EJ and JE NN compound translation tasks, we learn the templates from NN compound translation data. Given that the EJ and JE testdata is partitioned equally into the top-250 and random-250 NN compounds, we cross-validate the translation templates. That is, we perform two iterations over each of the JE and EJ datasets, taking one dataset of 250 NN compounds as the test set and the remaining dataset as the training set in each case. We first perform word-alignment on the training dataset, and in the case that both source language nouns align leaving only closed-class function words in the target language, extract out the mapping schema as a translation template (with word coindices). We then use this extracted set of translation templates as a filter in analysing word alignment in the test set.

A total of 23 JE and 3 EJ translation templates were learned from the training data in each case, a sample of which are shown in Table 1.[8] Here, the count for each template is the combined number of activations over each combined dataset of 500 compounds.

[7] The reason that we used Genius here is that, as an edited dictionary, Genius has a more complete coverage of translations for simplex words.

[8] For the 3 EJ templates learned on each iteration, there was an intersection of 2, and for the 23 JE templates, the intersection was only 10.

	TOP 250			*RAND 250*			*TOTAL*		
	Cov	***Acc***	***F***	***Cov***	***Acc***	***F***	***Cov***	***Acc***	***F***
JE	83.6	93.8	88.4	27.2	82.4	40.9	55.4	91.0	68.9
EJ	94.4	94.5	94.5	60.0	91.3	72.4	77.2	93.3	84.5

Table 2: Results for $\text{MBMT}_{\text{DICT}}$ (***F*** = F-score)

3.4 Evaluation measures

The principal evaluatory axes we consider in comparing the different methods are coverage and accuracy: **coverage** is the relative proportion of a given set of NN compounds that the method can generate some translation for, and **accuracy** describes the proportion of translated NN compounds for which the gold-standard translation is reproduced (irrespective of how many other translations are generated). These two tend to be in direct competition, in that more accurate methods tend to have lower coverage, and conversely higher coverage methods tend to have lower accuracy. So as to make cross-system comparison simple, we additionally combine these two measures into an **F-score**, that is their harmonic mean.

4 Results

We first present the individual results for $\text{MBMT}_{\text{DICT}}$ and DMT_{COMP}, and then discuss a cascaded system combining the two.

4.1 Dictionary-driven MBMT

The source of NN compound translations for $\text{MBMT}_{\text{DICT}}$ was the combined ALTDIC and EDICT dictionaries. Recall that this is the same dictionary as was used in the first pass of generation of gold standard translations (see § 3.1), but that the gold-standard translations were manually selected in the case of multiple dictionary entries, and an alternate translation manually generated in the case that a more appropriate translation was considered to exist.

The results for $\text{MBMT}_{\text{DICT}}$ are given in Table 2, for both translation directions. In each case, we carry out evaluation over the 250 most-commonly occurring NN compounds (***TOP 250***), the random sample of 250 NN compounds (***RAND 250***) and the combined 500-element dataset (***ALL***).

The accuracies (***Acc***) are predictably high, although slightly lower for the random-250 than the top-250. The fact that they are below 100% indicates that the translation dictionary is not infallible and contains a number of sub-optimal or misleading translations. One such example is 救済 · 基金 *kyuusai·kikiN* "relief fund" for which the dictionary provides the unique, highly-specialised translation *lifeboat*.

Coverage (***Cov***) is significantly lower than accuracy, but still respectable, particularly for the random-250 datasets. This is a reflection of the inevitable emphasis by lexicographers on more frequent expressions, and underlines the brittleness of $\text{MBMT}_{\text{DICT}}$. An additional reason for coverage being generally lower than accuracy is that dictionaries tend not to contain transparently compositional compounds, an observation which applies particularly to ALTDIC as it was developed for use with a full MT system. Coverage is markedly lower for the JE task, largely because ALTJAWS—which uses ALTDIC as its system dictionary—tends to treat the compound nouns in ALTDIC as single words. As we used ALTJAWS to pre-process the corpus we extracted the Japanese NN compounds from, a large component of the compounds in the translation dictionary was excluded from the JE data. One cause of a higher coverage for the EJ task is that many English compounds are translated into single Japanese words (e.g. *interest rate* vs. 利率 *riritsu*) and thus reliably recorded in bilingual dictionaries. There are 127 single word translations in the EJ dataset, but only 31 in the JE dataset.

In summary, $\text{MBMT}_{\text{DICT}}$ offers high accuracy but mid-range coverage in translating NN compounds, with coverage dropping off appreciably for less-frequent compounds.

4.2 Word-to-word composional DMT

In order to establish an upper bound on the performance of DMT_{COMP}, we word-aligned the source language NN compounds with their translations, using the extracted translation templates as described in § 3.3. The results of alignment are classified into four mutually-exclusive classes, as detailed below:

(A) Completely aligned All component words align according to one of the extracted translation templates.

(B) No template The translation does not correspond to a known translation template (irrespective of whether component words align in the source compound).

(C) Partially aligned Some but not all component words align. We subclassify instances of this class into: **C1** compounds, where there are unaligned words in both the source and target languages; **C2** compounds, where there is an unaligned word in the source language only; and **C3** compounds where there are unaligned words in the target language only.

(D) No alignment No component words align between the source NN compound and translation. We subclassify D instances into: **D1** compounds, where the translation is a single word; and **D2** compounds, where no word pair aligns.

The results of alignment are shown in Table 3, for each of the top-250, random-250 and combined 500-element datasets. The alignment was carried out using both the basic ALTDIC and ALTDIC+ (ALTDIC with 178 manually-added simplex entries). Around 40% of the data align completely using ALTDIC+ in both translation directions. Importantly, DMT_{COMP} is slightly more robust over the random-250 dataset

		JAPANESE-TO-ENGLISH						ENGLISH-TO-JAPANESE					
		ALTDIC			*ALTDIC+*			*ALTDIC*			*ALTDIC+*		
		Top	***Rand***	***All***	***Top***	***Rand***	***All***	***Top***	***Rand***	***All***	***Top***	***Rand***	***All***
Completely aligned (A)	Total	**26.4**	**26.0**	**26.2**	**39.6**	**43.6**	**41.6**	**29.6**	**34.4**	**32.0**	**39.2**	**45.6**	**42.4**
No template (B)	Total	**5.2**	**5.2**	**5.2**	**5.2**	**6.0**	**5.6**	**0.4**	**0.4**	**0.4**	**0.4**	**0.8**	**0.6**
Partially aligned (C)	Total	**44.0**	**48.8**	**46.4**	**38.4**	**36.4**	**37.4**	**29.2**	**39.2**	**34.2**	**24.8**	**30.8**	**27.8**
	C1	40.8	46.4	43.6	35.6	33.6	34.6	25.2	36.8	31.0	20.8	28.4	24.6
	C2	3.2	2.4	2.8	2.8	2.4	2.6	4.0	2.4	3.2	4.0	2.4	3.2
	C3	0.0	0.0	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
No alignment (D)	Total	**24.4**	**20.0**	**22.2**	**16.8**	**14.0**	**15.4**	**40.8**	**26.0**	**33.4**	**35.6**	**22.8**	**29.2**
	D1	5.2	2.4	3.8	5.2	2.4	3.8	31.2	13.2	22.2	31.2	13.2	22.2
	D2	19.2	17.6	18.4	11.6	11.6	11.6	9.6	12.8	11.2	4.4	9.6	7.0

Table 3: Alignment-based results for DMT_{COMP}

	JE			EJ		
	Cov	***Acc***	***F-score***	***Cov***	***Acc***	***F-score***
$\text{MBMT}_{\text{DICT}}$	55.4	91.0	68.9	77.2	93.3	84.5
DMT_{COMP}	96.4	43.1	59.6	87.0	48.7	62.5
Cascaded	96.4	71.6	82.2	95.6	87.0	91.1

Table 4: Cascaded translation results

than top-250, in terms of both completely aligned and partially aligned instances. This contrasts with $\text{MBMT}_{\text{DICT}}$ which was found to be brittle over the less-frequent random-250 dataset.

4.3 Combination of $\text{MBMT}_{\text{DICT}}$ and DMT_{COMP}

We have demonstrated $\text{MBMT}_{\text{DICT}}$ to have high accuracy but relatively low coverage (particularly over lower-frequency NN compounds), and DMT_{COMP} to have medium accuracy but high coverage. To combine the relative strengths of the two methods, we test a cascaded architecture, whereby we first attempt to translate each NN compound using $\text{MBMT}_{\text{DICT}}$, and failing this, resort to DMT_{COMP}.

Table 4 shows the results for $\text{MBMT}_{\text{DICT}}$ and DMT_{COMP} in isolation, and when cascaded (***Cascade***). For both translation directions, cascading results in a sharp increase in F-score, with coverage constantly above 95% and accuracy dropping only marginally to just under 90% for the EJ task. The cascaded method represents the best-achieved shallow translation upper bound achieved in this research.

5 Analysis and extensions

In this section, we offer qualitative analysis of the unaligned translation pairs (i.e. members of classes B, C and D in Table 3) with an eye to improving the coverage of DMT_{COMP}. We make a tentative step in this direction by suggesting one extension to the basic DMT_{COMP} paradigm based on synonym substition.

5.1 Analysis of unaligned translation pairs

We consider there to be 6 basic types of misalignment in the translation pairs, each of which we illustrate with examples (in which underlined words are aligned and **boldface words** are the focus of discussion). In listing each misalignment type, we indicate the corresponding alignment classes in § 4.2.

(a) Missing template (B) An example of misalignment due to a missing template (but where all component words align) is:

(a1) 決勝·進出 *kesshou·shiNshutsu* "advancement to finals"

Simply extending the coverage of translation templates would allow DMT_{COMP} to capture examples such as this.

(b) Single-word translation (C2,D1) DMT_{COMP} fails when the gold-standard translation is a single word:

(b1) **情報**·開示 *jouhou·kaiji* "(lit.) information disclosure" ⇒ *disclosure*

(b2) 首脳 · 会談 *shunou·kaidaN* "(lit.) leader meeting" ⇒ *summit*

(b3) *interest rate* ⇒ 利率 *riritsu*

In (b1), the misalignment is caused by the English *disclosure* default-encoding *information*; a similar case can be made for (b2), although here *summit* does not align with 会談 *kaidaN*. DMT_{COMP} could potentially cope with these given a lexical inference module interfacing with a semantically-rich lexicon (particularly in the case of (b1) where translation selection at least partially succeeds), but $\text{DMT}_{\text{INTERP}}$ seems the more natural model for coping with this type of translation. (b3) is slightly different again, in that 利率 *riritsu* can be analysed as a two-character abbreviation derived from 利息 *risoku* "interest" and 率 *ritsu* "rate", which aligns fully with *interest rate*. Explicit abbreviation expansion could unearth the full wordform and facilitate alignment.

(c) Synonym and association pairs (C1) This class contains translation pairs where one or more pairs of component nouns does not align under exact translation, but are conceptually similar:

(c1) ***budget*** *deficit* ⇒ **財政**·赤字 *zaisei·akaji* "finance deficit"

(c2) **加盟**·国 *kamei·koku* "affiliation state" ⇒ ***member state***

In (c1), although 財政 *zaisei* "finance" is not an exact translation of *budget*, they are both general financial terms. It may be possible to align such words using word similarity, which would enable DMT_{COMP} to translate some component of the C1 data. In (c2), on the other hand, 加盟 *kamei* "affiliation" is lexically-associated with the English *membership*, although here the link becomes more tenuous.

(d) Mismatch in semantic explicitness (C1) This translation class is essentially the same as class (b) above, in that semantic content explicitly described in the source NN compound is made implicit in the translation. The only difference is that the translation is not a single word so there is at least the potential for word-to-word compositionality to hold:

(d1) **州知事**·選挙 *shuuchiji·seNkyo* "(lit.) state-governor election" ⇒***state election***

(e) Concept focus mismatch (C1-2,D2) The source NN compound and translation express the same concept differently due to a shift in semantic focus:

(e1) 就職 · **活動** *shuushoku·katsudou* "(lit.) activity for getting new employment" ⇒ *job* ***hunting***.

Here, the mismatch is between the level of directed participation in the process of finding a job. In Japanese, 活動 *katsudou* "activity" describes simple involvement, whereas *hunting* signifies a more goal-oriented process.

(f) Lexical gaps (C3,D2) Members of this class cannot be translated compositionally as they are either non-compositional expressions or, more commonly, there is no conventionalised way of expressing the denoted concept in the target language:

(f1) 族 · 議員 *zoku·giiN* "legistors championing the causes of selected industries"

These translation pairs pose an insurmountable obstacle for DMT_{COMP}.

Of these types, (a), (b) and (c) are the most realistically achievable for DMT_{COMP}, which combined account for about 20% of coverage, suggesting that it would be worthwhile investing effort into resolving them.

5.2 Performance vs. translation fan-out

As mentioned in § 5.1, there are a number of avenues for enhancing the performance of DMT_{COMP}. Here, we propose synonym-based substitution as a means of dealing with synonym pairs from class (c).

The basic model of word substitution can be extending simply by inserting synonym translations as well as direct word translations into the translation templates. We test-run this extended method for the JE translation task, using the *Nihongo Goi-taikei* thesaurus (Ikehara et al., 1997) as the source of source language synonyms, and ALTDIC+ as our translation dictionary. The *Nihongo Goi-taikei* thesaurus classifies the contents of ALTDIC into 2,700 semantic classes. We consider words occurring in the same class to be synonyms, and add in the translations for each. Note that we test this configuration over only C1-type compounds due to the huge fan-out in translation candidates generated by the extended method (although performance is evaluated over the full dataset, with results for non-C1 compounds remaining constant throughout).

Configuration	***Cov***	***Acc***	***F-score***	***Fan-out***
$\text{MBMT}_{\text{DICT}}$ (orig)	55.4	91.0	68.9	2
DMT_{COMP} (orig)	96.4	43.1	59.6	74
DMT_{COMP} (6 TTs−sim)	95.6	41.4	57.8	20
DMT_{COMP} (6 TTs+sim)	95.6	47.1	63.1	6,577
DMT_{COMP} (13 TTs−sim)	96.6	43.2	59.7	43
DMT_{COMP} (13 TTs+sim)	96.6	48.1	64.1	13,911

Table 5: Performance vs. translation fan-out (JE)

One significant disadvantage of synonym-based substitution is that it leads to an exponential increase in the number of translation candidates. If we analyse the complexity of simple word-based substitution to be $O(n^2)$ where n is the average number of translations per word, the complexity of synonym based substitution becomes $O((m^2+1)n^2)$ where m is the average number of synonyms per class.

Table 5 shows the translation performance and also translation fan-out (average number of translation candidates) for DMT_{COMP} with and without synonym-based substitution (±***sim***) over the top 6 and 13 translation templates (***TTs***). As baselines, we also present the results for $\text{MBMT}_{\text{DICT}}$ (***MBMT*$_{\text{DICT}}$ *(orig)***) and DMT_{COMP} (***DMT*$_{\text{COMP}}$ *(orig)***) in their original configurations (over the full 23 templates and without synonym-substitution for DMT_{COMP}). From this, the exponential translation fan-out for synonym-based substitution is immediately evident, but accuracy can also be seen to increase by over 4 percentage points through the advent of synonym substitution. Indeed, the accuracy when using synonym-substitution over only the top 6 translation templates is greater than that for the basic DMT_{COMP} method, although the number of translation candidates is clearly greater. Note the marked difference in fan-out for $\text{MBMT}_{\text{DICT}}$ vs. the various incarnations of DMT_{COMP}, and that considerable faith is placed in the ability of translation selection with DMT_{COMP}.

While the large number of translation candidates produced by synonym-substitution make translation selection appear intractable, most candidates are meaningless word sequences, which can easily be filtered out based on target language corpus evidence. Indeed, Tanaka (2002) successfully combines synonym-substitution with translation selection and

achieves appreciable gains in accuracy.

6 Conclusion and future work

This paper has used the NN compound translation task to establish performance upper bounds on shallow translation methods and in the process empirically determine the relative need for deep translation methods. We focused particularly on dictionary-driven MBMT and word-to-word compositional DMT, and demonstrated the relative strengths of each. When cascaded these two methods were shown to achieve 95%+ coverage and potentially high translation accuracy. As such, shallow translation methods are able to translate the bulk of NN compound inputs successfully.

One question which we have tactfully avoided answering is how deep translation methods perform over the same data, and how successfully they can handle the data that shallow translation fails to produce a translation for. We leave these as items for future research. Also, we have deferred the issue of translation selection for the methods described here, and in future work hope to compare a range of translation selection methods using the data developed in this research.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Emily Bender, Francis Bond, Dan Flickinger, Stephan Oepen, Ivan Sag and the three anonymous reviewers for their valuable input on this research.

References

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 96–102, Montreal, Canada.

Laurie Bauer. 1979. On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics*, 3:45–50.

Jim Breen. 1995. *Building an electronic Japanese-English dictionary*. Japanese Studies Association of Australia Conference

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Yunbo Cao and Hang Li. 2002. Base noun phrase translation using Web data and the EM algorithm. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Ann Copestake and Alex Lascarides. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, pages 136–43, Madrid, Spain.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proc. of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. of the 33rd Annual Meeting of the ACL*, pages 236–43, Cambridge, USA.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Proc. of the Third Machine Translation Summit (MT Summit III)*, pages 101–106, Washington DC, USA.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Nihongo Goi-Taikei – A Japanese Lexicon*. Iwanami Shoten.

Tomoshichi Konishi, editor. 1997. *Genius English-Japanese and Japanese-English Dictionary CD-ROM edition*. Taishukan Publishing Co., Ltd.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, USA.

Mainichi Newspaper Co. 1996. Mainichi Shimbun CD-ROM 1996.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

Kumiko Ohmori and Masanobu Higashida. 1999. Extracting bilingual collocations from non-aligned parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI99)*, pages 88–97, Chester, UK.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the 37th Annual Meeting of the ACL*, pages 1–17, College Park, USA.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, USA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Takaaki Tanaka and Yoshihiro Matsuo. 1999. Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 109–19, Chester, UK.

Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 981–7, Taipei, Taiwan.