Translation by Machine of Complex Nominals: Getting it Right

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

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

Abstract

We present a method for compositionally translating noun-noun (NN) compounds, using a word-level bilingual dictionary and syntactic templates for candidate generation, and corpus and dictionary statistics for selection. We propose a support vector learning-based method employing target language corpus and bilingual dictionary data, and evaluate it over a English↔Japanese machine translation task. We show the proposed method to be superior to previous methods and also robust over low-frequency NN compounds.

1 Introduction

Noun-noun (NN) compounds (e.g. *web server*, 機械·翻訳 *kikai·hoNyaku* "machine translation",[1] the elements of which we will refer to as N_1 and N_2 in linear order of occurrence) are a very real problem for both machine translation (MT) systems and human translators due to:

constructional variability in the translations:
機械·翻訳 *kikai·hoNyaku* "machine translation" (N-N) vs. 民間·企業 *miNkaN·kigyou* "private company" (Adj-N) vs. 関係·改善 *kaNkei·kaizeN* "improvement in relations" (N *in* N);

lexical divergences in Japanese and English:
配布·計画 *haifu·keikaku* "distribution schedule" vs. 経済·計画 *keizai·keikaku* "economic plan/programme" vs. 主要·計画 *shuyou·keikaku* "major project";

semantic underspecification: compounds generally have multiple interpretations, and can only be reliably interpreted in context (Levi, 1978);

the existence of non-compositional NN compounds:
井戸端·会議 *idobata·kaigi* "(lit.) well-side meeting", which translates most naturally into English as "idle gossip";

high productivity and frequency

In order to quantify the high productivity and frequency of NN compounds, we carried out a

---

[1]With all Japanese NN compound examples, we segment the compound into its component nouns through the use of the "·" symbol. Note that no such segmentation boundary is indicated in the original Japanese.

	BNC	*Reuters*	*Mainichi*
Token coverage	2.6%	3.9%	2.9%
Total no. types	265K	166K	889K
Ave. token freq.	4.2	12.7	11.1
Singletons	60.3%	44.9%	45.9%

Table 1: Corpus occurrence of NN compounds

basic study of corpus occurrence in English and Japanese. For English, we based our analysis over: (1) the written portion of the British National Corpus (BNC, 84M words: Burnard (2000)), and (2) the Reuters corpus (108M words: Rose et al. (2002)). For Japanese, we focused exclusively on the Mainichi Shimbun Corpus (340M words: Mainichi Newspaper Co. (2001)). We identified NN compounds in each corpus using the method described in § 2.2 below, and from this, derived the statistics of occurrence presented in Table 1. The token coverage of NN compounds in each corpus refers to the percentage of words which are contained in NN compounds; based on our corpora, we estimate this figure to be as high as 3-5%. If we then look at the average token frequency of each distinct NN compound type, we see that it is a relatively modest figure given the size of each of the corpora, the reason for which is seen in the huge number of distinct NN compound types. Combining these observations, we see that a translator or MT system attempting to translate one of these corpora will run across NN compounds with high frequency, but that each individual NN compound will occur only a few times (with around 45-60% occuring only once). The upshot of this for MT systems and translators is that NN compounds are too varied to be able to pre-compile an exhaustive list of translated NN compounds, and must instead be able to deal with novel NN compounds on the fly. This claim is supported by Tanaka and Baldwin (2003a), who found that static bilingual dictionaries had a type coverage of around 84% and 94% over the top-250 most frequent English and Japanese NN compounds, respectively, but only 27% and 60%, respectively, over a random sample of NN compounds occurring more than 10 times in the corpus.

We develop and test a method for translating NN compounds based on Japanese↔English MT. The method can act as a standalone module in an MT

system, translating NN compounds according to the best-scoring translation candidate produced by the method, and it is primarily in this context that we present and evaluate the method. This is congruent with the findings of Koehn and Knight (2003) that, in the context of statistical MT, overall translation performance improves when source language noun phrases are prescriptively translated as noun phrases in the target language. Alternatively, the proposed method can be used to generate a list of plausible translation candidates for each NN compound, for a human translator or MT system to select between based on the full translation context.

In the remainder of the paper, we describe the translation procedure and resources used in this research (§ 2), and outline the translation candidate selection method, a benchmark selection method and pre-processors our method relies on (§ 3). We then evaluate the method using a variety of data sources (§ 4), and finally compare our method to related research (§ 5).

2 Preliminaries

2.1 Translation procedure

We translate NN compounds by way of a two-phase procedure, incorporating generation and selection (similarly to Cao and Li (2002) and Langkilde and Knight (1998)).

Generation consists of looking up word-level translations for each word in the NN compound to be translated, and running them through a set of constructional translation templates to generate translation candidates. In order to translate 関係·改善 *kaNkei·kaizeN* "improvement in relations", for example, possible word-level translations for 関係 are *relation*, *connection* and *relationship*, and translations for 改善 are *improvement* and *betterment*. Constructional templates are of the form $[N_2^E\ in\ N_1^E]$ (where N_i^E indicates that the word is a noun (N) in English (E) and corresponds to the ith-occurring noun in the original Japanese; see Table 3 for further example templates and Kageura et al. (2004) for discussion of templates of this type). Each slot in the translation template is indexed for part of speech (POS), and derivational morphology is optionally used to convert a given word-level translation into a form appropriate for a given template. Example translation candidates for 関係·改善, therefore, are *relation improvement*, *betterment of relationship*, *improvement connection* and *relational betterment*. Note that generation fails in the instance that we are unable to find a word-level translation for N_1 and/or N_2.

Selection consists of selecting the most likely translation for the original NN compound from the generated translation candidates. Selection is performed based on a combination of monolingual target language and crosslingual evidence, obtained from corpus or web data.

Ignoring the effects of POS constraints for the moment, the number of generated translations is $O(mnt)$ where m and n are the fertility of Japanese nouns N_1^J and N_2^J, respectively, and t is the number of translation templates. As a result, there is often a large number of translation candidates to select between, and the selection method crucially determines the efficacy of the method.

This translation procedure has the obvious advantage that it can generate a translation for any NN compound input assuming that there are word-level translations for each of the component nouns; that is it has high coverage. It is based on the assumption that NN compounds translate compositionality between Japanese and English, which Tanaka and Baldwin (2003a) found to be the case 43.1% of the time for Japanese–English (JE) MT and 48.7% of the time for English–Japanese (EJ) MT. In this paper, we focus primarily on selecting the correct translation for those NN compounds which can be translated compositionally, but we also investigate what happens when non-compositional NN compounds are translated using a compositional method.

2.2 Translation data

In order to generate English and Japanese NN compound testdata, we first extracted out all NN bigrams from the Reuters Corpus and Mainichi Shimbun Corpus. The Reuters Copus was first tagged and chunked using fnTBL (Ngai and Florian, 2001), and lemmatised using morph (Minnen et al., 2001), while the Mainichi Shimbun was segmented and tagged using ChaSen (Matsumoto et al., 1999). 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. 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.[2] 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 calculated the frequency of occurrence of each NN compound type identified in the English and Japanese corpora, and ranked the NN compound types in order of corpus frequency. Based on this ranking, we split the NN compound types into three partitions of equal token frequency, and from each partition, randomly selected 250 NN com-

---

[2] 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.

Band		English	Japanese
HIGH	Freq. range	346–24,025	336–64,835
	Types	791	4,009
MED	Freq. range	44–345	37–336
	Types	6,576	32,283
LOW	Freq. range	1–44	1–37
	Types	158,215	852,328

Table 2: Frequency bands

pounds. In doing so, we produced NN compound data representative of three disjoint frequency bands of equal token size, as detailed in Table 2. This allows us to analyse the robustness of our method over data of different frequencies.

Our motivation in testing the proposed method over NN compounds according to the three frequency bands is to empirically determine: (a) whether there is any difference in translation-compositionality for NN compounds of different frequency, and (b) whether our method is robust over NN compounds of different frequency. We return to these questions in § 4.1.

In order to evaluate basic translation accuracy over the test data, we generated a unique **gold-standard translation** for each NN compound to represent its optimally-general default translation. This was done with reference to two bilingual Japanese-English dictionaries: the ALTDIC dictionary and the on-line EDICT dictionary. 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** (Breen, 1995) has approximately 150,000 entries. Note that the existence of a translation for a given NN compound in one of the dictionaries does not guarantee that we used it as our gold-standard, and 35% of JE translations and 25% of EJ translations were rejected in favour of a manually-generated translation. In generating the gold-standard translation data, we checked the validity of each of the randomly-extracted NN compounds, and rejected a total of 0.5% of the initial random sample of Japanese strings, and 6.6% of the English strings, on the grounds of: (1) not being NN compounds, (2) being proper nouns, or (3) being part of a larger MWE. In each case, the rejected string was replaced with an alternate randomly-selected NN compound.

2.3 Translation templates

The generation phase of translation relies on translation templates to recast the source language NN compound into the target language. The translation templates were obtained by way of word alignment over the JE and EJ gold-standard translation datasets, generating a total of 28 templates for the JE task and 4 templates for the EJ task. The reason for the large number of templates in the JE task is that they are used to introduce prepositions and possessive markers, as well as indicating word class conversions (see Table 3).

3 Selection methodology

In this section, we describe a benchmark selection method based on monoligual corpus data, and a novel selection method combining monolingual corpus data and crosslingual data derived from bilingual dictionaries. Each method takes the list of generated translation candidates and scores each, returning the highest-scoring translation candidate as our final translation.

3.1 Benchmark monolingual method

The monolingual selection method we benchmark ourselves against is the **corpus-based translation quality** (CTQ) method of Tanaka and Baldwin (2003b). It rates a given translation candidate according to corpus evidence for both the fully-specified translation and its parts in the context of the translation template in question. This is calculated as:[3]

$$ctq(w_1^{L2}, w_2^{L2}, t) = \alpha \; p(w_1^{L2}, w_2^{L2}, t) \; + \beta \; p(w_1^{L2}, t)p(w_2^{L2}, t)$$

where w_1^{L2} and w_2^{L2} are the word-level translations of the source language N_1^{L1} and N_2^{L1}, respectively, and t is the translation template.[4] Each probability is calculated according to a maximum likelihood estimate based on relative corpus occurrence. The formulation of CTQ is based on linear interpolation over α and β, where $0 \leq \alpha, \beta \leq 1$ and $\alpha + \beta = 1$. We set α to 0.9 and β to 0.1 throughout evaluation.

The basic intuition behind decomposing the translation candidate into its two parts within the context of the translation template ($p(w_1^{L2}, t)$ and $p(w_2^{L2}, t)$) is to capture the subcategorisation properties of w_1^{L2} and w_2^{L2} relative to t. For example, if w_1^{L2} and w_2^{L2} were *Bandersnatch* and *relation*, respectively, and $p(w_1^{L2}, w_2^{L2}, t) = 0$ for all t, we would hope to score *relation to (the) Bandersnatch* as being more likely than *relation on (the) Bandersnatch*. We could hope to achieve this by virtue of the fact that *relation* occurs in the form *relation to ...* much more frequently than *relation on ...*, making the value of $p(w_2^{L2}, t)$ greater for the template $[N_2^E \text{ to } N_1^E]$ than $[N_2^E \text{ on } N_1^E]$.

In evaluation, Tanaka and Baldwin (2003b) found the principal failing of this method to be its treatment of all translations contained in the transfer

---
[3]In the original formulation, the product $p(w_1^{L2})p(w_2^{L2})p(t)$ was included as a third term, but Tanaka and Baldwin (2003b) found it to have negligible impact on translation accuracy, so we omit it here.

[4]w_1^{L2} and w_2^{L2} are assumed to be POS-compatible with t.

Template (JE)	Example
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_1\ N_2]_{\mathbf{E}}$	市場·経済 *shijou·keizai* "market economy"
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_2\ N_1]_{\mathbf{E}}$	賛成·多数 *saNsei·tasuu* "majority agreement"
$[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_2\ of\ (the)\ N_1]_{\mathbf{E}}$	政権·交替 *seikeN·koutai* "change of government"

Template (EJ)	Example
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_1\ N_2]_{\mathbf{J}}$	*exchange rate* 為替·レート "*kawase·reeto*"
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_1\ teki\ N_2]_{\mathbf{J}}$	*world leader* 世界·的·リーダー "*sekai·teki·leader*"
$[N_1\ N_2]_{\mathbf{E}} \Rightarrow [N_2\ no\ N_1]_{\mathbf{J}}$	*baby girl* 女·の·赤ちゃん "*oNna·no·akachaN*"

Table 3: Example translation templates (*N* = noun and *Adj* = adjective)

dictionary as being equally likely, where in fact there is considerable variability in their applicatility. One example of this is the simplex 記事 *kiji* which is translated as either *article* or *item* (in the sense of a newspaper) in ALTDIC, of which the former is clearly the more general translation. Lacking knowledge of this conditional probability, the method considers the two translations to be equally probable, giving rise to the preferred translation of *related item* for 関連·記事 *kaNreN·kiji* "related article" due to the markedly greater corpus occurrence of *related item* over *related article*. It is this aspect of selection that we focus on in our proposed method.

3.2 Proposed selection method

The proposed method uses the corpus-based monolingual probability terms of CTQ above, but also mono- and crosslingual terms derived from bilingual dictionary data. In doing so, it attempts to preserve the ability of CTQ to model target language expressional preferences, while incorporating more direct translation preferences at various levels of lexical specification. For ease of feature expandability, and to avoid interpolation over excessively many terms, the backbone of the method is the TinySVM support vector machine (SVM) learner.[5]

The way we use TinySVM is to take all source language inputs where the gold-standard translation is included among the generated translation candidates, and construct a single feature vector for each translation candidate. We treat those feature vectors which correspond to the (unique) gold-standard translation as positive exemplars, and all other feature vectors as negative exemplars. We then run TinySVM over the training exemplars using the ANOVA kernel (the only kernel which was found to converge). Strictly speaking, SVMs produce a binary classification, by returning a continuous value and determining whether it is closest to $+1$ (the positive class) or -1 (the negative class). We treat this value as a translation quality rating, and rank the translation candidates accordingly. To select the best translation candidate, we simply take the best-scoring exemplar, breaking ties through random se-

lection.

The selection method makes use of three basic feature types in generating a feature vector for each source language–translation candidate pair: corpus-based features, bilingual dictionary-based features and template-based features.

Corpus-based features

Each source language–translation pair is mapped onto a total of 8 corpus-based feature types, in line with the CTQ formulation above:

- $ctq(w_1^{L2}, w_2^{L2}, t)$
- $freq(w_1^{L2}, w_2^{L2}, t)$
- $freq(w_1^{L2}, t)$ and $freq(w_2^{L2}, t)$
- $freq(w_1^{L2})$, $freq(w_2^{L2})$ and $freq(t)$
- $mwe(w_1^{L2}, w_2^{L2}, t)$

$mwe(w_1^{L2}, w_2^{L2}, t)$ is a normalisation parameter used to estimate the frequency of occurrence of multiword expression (MWE) translations from that of the head. E.g., in generating translations for 不動産·会社 *fudousaN·gaisha* "real estate company", we get two word-level translations for 不動産: *real estate* and *real property*. In each case, we identify the final word as the head, and calculate the number of times the MWEs (i.e. *real estate* and *real property*) occur in the overall corpus as compared to the head (i.e. *estate* and *property*, respectively). In calculating the values of each of the frequency-based features involving these translations, we determine the frequency of the head in the given context, and multiply this by the normalisation parameter. The reason for doing this is for ease of calculation and, wherever possible, to avoid zero values for frequencies involving MWEs. The feature $mwe(w_1^{L2}, w_2^{L2}, t)$ is generated by multiplying the MWE parameters for each of w_1^{L2} and w_2^{L2} (which are set to 1.0 in the case that the translation is simplex) and intended to model the tendency to prefer simplex translations over MWEs when given a choice.

We construct an additional feature from each of these values, by normalising (by simple division to generate a value in the range $[0, 1]$) relative to the maximum value for that feature among the translation candidates generated for a given source language input. For each corpus, therefore, the total number of corpus-based features is $8 \times 2 = 16$.

In EJ translation, the corpus-based feature values were derived from the Mainichi Shimbun Corpus, whereas in JE translation, we used the BNC and Reuters Corpus, and concatenated the feature values from each.

Bilingual dictionary-based features

Bilingual dictionary data is used to generate 6 features:

- $freq_{\mathrm{DIC\_COMB}}(w_1^{L2}, w_2^{L2}, t|w_1^{L1}, w_2^{L1})$
- $freq_{\mathrm{DIC\_COMB}}(w_1^{L2}, w_2^{L2}, t)$
- $freq_{\mathrm{DIC}}(w_1^{L2}, t|w_1^{L1})$ and $freq_{\mathrm{DIC}}(w_2^{L2}, t|w_2^{L1})$
- $freq_{\mathrm{DIC}}(w_1^{L2}|w_1^{L1})$ and $freq_{\mathrm{DIC}}(w_2^{L2}|w_2^{L1})$

$freq_{\mathrm{DIC\_COMB}}(w_1^{L2}, w_2^{L2}, t|w_1^{L1}, w_2^{L1})$ is the total number of times the given translation candidate occurs as a translation for the source language NN compound *across all dictionaries*. While this feature may seem to give our method an unfair advantage over CTQ, it is important to realise that only limited numbers of NN compounds are listed in the dictionaries (12% for English and 28% for Japanese), and that the gold-standard accuracy when the dictionary translation is selected is not as high as one would expect (65% for English and 75% for Japanese). $freq_{\mathrm{DIC\_COMB}}(w_1^{L2}, w_2^{L2}, t)$ describes the total occurrences of the translation candidate across all dictionaries (irrespective of the source language expression it translates), and is considered to be an indication of conventionalisation of the candidate.

The remaining features are intended to capture word-level translation probabilities, optionally in the context of the template used in the translation candidate. Returning to our 関連·記事 *kaNreN·kiji* "related article" example from above, of the translations *article* and *item* for 記事, *article* occurs as the translation of 記事 for 42% of NN entries with 記事 as the N_2, and within 18% of translations for complex entries involving 記事 (irrespective of the form or alignment between *article* and 記事). For *item*, the respective statistics are 9% and 4%. From this, we can conclude that *article* is the more appropriate translation, particularly for the given translation template.

As with the corpus-based features, we additionally construct a normalised variant of each feature value, such that the total number of bilingual dictionary-based features is $4 \times 2 = 8$.

In both JE and EJ translation, we derived bilingual dictionary-based features from the EDICT and ALTDIC dictionaries independently, and concatenated the features derived from each.

Template-based features

We use a total of two template-based features: the template type and the target language head (N1 or N2). For template $[N_1\ N_2]_{\mathbf{J}} \Rightarrow [N_2\ N_1]_{\mathbf{E}}$ (see § 2.3), e.g., the template type is N-N and the target language head is N1.

3.3 Corpus data

The corpus frequencies were extracted from the same three corpora as were described in § 1: the BNC and Reuters Corpus for English, and Mainichi Shimbun Corpus for Japanese. We chose to use the BNC and Reuters Corpus because of their complementary nature: the BNC is a balanced corpus and hence has a rounded coverage of NN compounds (see Table 1), whereas the Reuters Corpus contains newswire data which aligns relatively well in content with the newspaper articles in the Mainichi Shimbun Corpus.

We calculated the corpus frequencies based on the tag and dependency output of RASP (Briscoe and Carroll, 2002) for English, and CaboCha (Kudo and Matsumoto, 2002) for Japanese. RASP is a tag sequence grammar-based stochastic parser which attempts to exhaustively resolve inter-word dependencies in the input. CaboCha, on the other hand, chunks the input into head-annotated "bunsetsu" or base phrases, and resolves only inter-phrase dependencies. We thus independently determined the intra-phrasal structure from the CaboCha output based on POS-conditioned templates.

4 Evaluation

We evaluate the method over both JE and EJ translation selection, using the two sets of 750 NN compounds described in § 2.2. In each case, we first evaluate system performance according to gold-standard accuracy, i.e. the proportion of inputs for which the (unique) gold-standard translation is ranked top amongst the translation candidates. For the method to have a chance at selecting the gold-standard translation, we clearly must be able to generate it. The first step is thus to identify inputs which have **translation-compositional** gold-standard translations, and generate the translation candidates for each. The translation-compositional data has the distribution given in Table 4. The overall proportion of translation-compositional inputs is somewhat lower than suggested by Tanaka and Baldwin (2003a), although this is conditional on the coverage of the particular dictionaries we use. The degree of translation-compositionality appears to be relatively constant across the three frequency bands, a somewhat surprising finding as we had expected the lower frequency NN compounds to be less conventionalised and therefore have more straightforwardly compositional translations.

We use the translation-compositional test data to evaluate the proposed method ($\mathrm{SVM}_{\mathrm{L2+X}}$) against CTQ and a simple baseline derived from CTQ, which takes the most probable fully-specified translation

	JE	EJ
ALL	297/750	272/750
HIGH	99/250	108/250
MED	98/250	81/250
LOW	100/250	83/250

Table 4: Analysis of translation compositionality

	Baseline	CTQ	SVM_{L2}	SVM_X	SVM_{L2+X}
JE	.317	.367	.390	.382	**.434**
EJ	.400	.416	.441	.296	**.514**

Table 5: Gold-standard translation accuracies

	Baseline	CTQ	SVM_{L2}	SVM_X	SVM_{L2+X}
JE	.616	.721	.764	.693	**.839**
EJ	.621	.654	**.721**	.419	**.783**

Table 6: Silver-standard translation accuracies

Band	Training data	Baseline G	S	CTQ G	S	SVM_{L2+X} G	S
HIGH	All	.425	.789	.445	.806	.464	**.879**
	Local					.462	.857
MED	All	.315	.665	.368	.797	**.474**	**.889**
	Local					**.480**	**.878**
LOW	All	.210	.393	.280	.569	.332	**.742**
	Local					.320	**.720**

Table 7: JE translation accuracies across different frequency bands

candidate (i.e. is equivalent to setting $\alpha = 1$ and $\beta = 0$). We additionally tested the proposed method using just corpus-based features (SVM_{L2}) and bilingual dictionary-based features (SVM_X) to get a better sense for the relative impact of each on overall performance. In the case of the proposed method and its derivants, evaluation is according to 10-fold stratified cross-validation, with stratification taking place across the three frequency bands. The average number of translations generated for the JE dataset was 205.6, and that for the EJ dataset was 847.5. We were unable to generate any translations for 17 (2.3%) and 57 (7.6%) of the NN compounds in the JE and EJ datasets, respectively, due to there being no word-level translations for N_1 and/or N_2 in the combined ALTDIC/EDICT dictionaries.

The gold-standard accuracies are presented in Table 5, with figures in **boldface** indicating a statistically significant improvement over both CTQ and the baseline.[6] Except for SVM_X in the EJ task, all evaluated methods surpass the baseline, and all variants of SVM surpassed CTQ. SVM_{L2+X} appears to successfully consolidate on SVM_{L2} and SVM_X, indicating that our modelling of target language corpus and crosslingual data is complementary. Overall, the results for the EJ task are higher than those for the JE task. Part of the reason for this is that Japanese has less translation variability for a given pair of word translations, as discussed below.

In looking through the examples where a gold-standard translation was not returned by the different methods, we often find that the uniqueness of gold-standard translation has meant that equally good translations (e.g. *dollar note* vs. the gold-standard translation *dollar bill* for ドル・紙幣 *doru·shihei*) or marginally lower-quality but perfectly acceptable translations (e.g. *territorial issue* vs. the gold-standard translation of *territorial dispute* for 領土・問題 *ryoudo·moNdai*) are adjudged incorrect. To rate the utility of these near-miss translations, we rated each non-gold-standard first-ranking translation according to **source language-recoverability** (L1-recoverability). L1-recoverable

---

[6]Based on the paired t test, $p < 0.05$

translations are defined to be syntactically unmarked, capture the basic semantics of the source language expression and allow the source language expression to be recovered with reasonable confidence. While evaluation of L1-recoverability is inevitably subjective, we minimise bias towards any given system by performing the L1-recoverability annotation for all methods in a single batch, without giving the annotator any indication of which method selected which translation. The average number of English and Japanese L1-recoverable translations were 1.9 and 0.85, respectively. The principle reason for the English data being more forgiving is the existence of possessive- and PP-based paraphrases of NN gold-standard translations (e.g. *ammendment of rule(s)* as an L1-recoverable paraphrase of *rule ammendment*).

We combine the gold-standard data and L1-recoverable translation data together into a single **silver standard** translation dataset, based upon which we calculate silver-standard translation accuracy. The results for the translation-compositional data are given in Table 6. Once again, we find that the proposed method is superior to the baseline and CTQ, and that the combination of crosslingual and target language corpus data is superior to the individual data sources. SVM_X fares particularly badly under silver-standard evaluation as it is unable to capture the target language lexical and constructional preferences as are needed to generate syntactically-unmarked, natural-sounding translations. Unsurprisingly, the increment between gold-standard accuracy and silver-standard accuracy is greater for English than Japanese.

4.1 Accuracy over each frequency band

We next analyse the breakdown in gold- and silver-standard accuracies across the three frequency bands. In doing this, we test the hypothesis that training over only translation data from the same frequency band will produce better results than

Band	Training data	Baseline		CTQ		$\text{SVM}_{\text{L2+X}}$	
		G	S	G	S	G	S
HIGH	All	.451	.641	.463	.657	**.630**	**.842**
	Local					**.657**	**.850**
MED	All	.420	.655	.452	.674	.532	.762
	Local					**.546**	.776
LOW	All	.314	.561	.341	.633	.396	**.755**
	Local					.374	.708

Table 8: EJ translation accuracies across different frequency bands

	Baseline	CTQ	SVM_{L2}	SVM_{X}	$\text{SVM}_{\text{L2+X}}$
JE	.358	.515	.490	.308	**.549**
EJ	.208	.285	.350	.162	.277

Table 9: Silver-standard translation accuracies over non-translation-compositional data

training over all the translation data. The results for the JE and EJ translation tasks are presented in Tables 7 and 8, respectively. The results based on training over data from all frequency bands are labelled **All** and those based on training over data from only the same frequency band are labelled **Local**; **G** is the gold-standard accuracy and **S** is the silver-standard accuracy.

For each of the methods tested, we find that the gold- and silver-standard accuracies drop as we go down through the frequency bands, although the drop off is markedly greater for gold-standard accuracy. Indeed, silver-standard accuracy is constant between the high and medium bands for the JE task, and the medium and low frequency bands for the EJ task. $\text{SVM}_{\text{L2+X}}$ appears to be robust over low-frequency data for both tasks, with the absolute difference in silver-standard accuracy between the high and low frequency bands around only 0.10, and never dropping below 0.70 for either the EJ or JE task. There was very little difference between training over data from all frequency bands as compared to only the local frequency band, suggesting that there is little to be gained from conditioning training data on the relative frequency of the NN compound we are seeking to translate.

4.2 Accuracy over non-translation-compositional data

Finally, we evaluate the performance of the methods over the non-translation compositional data. We are unable to give gold-standard accuracies here as, by definition, the gold-standard translation is not amongst the translation candidates generated for any of the inputs. We are, however, able to evaluate according to silver-standard accuracy, constructing L1-recoverable translation data as for the translation-compositional case described above. The classifier is learned from all the translation-compositional data, treating the gold-standard trans-

lations as positive exemplars as before.

The results are presented in Table 9. A large disparity is observable here between the JE and EJ accuracies, which is, once again, a direct result of Japanese being less forgiving when it comes to L1-recoverable translations. For the translation-compositional data, the EJ task displayed a similarly diminished accuracy increment when the L1-recoverable translation data was incorporated, but this was masked by the higher gold-standard accuracy for the task. The relative results for the JE task largely mirror those for the translation-compositonal data. In contrast, $\text{SVM}_{\text{L2+X}}$ actually performs marginally worse than CTQ over the EJ task, despite SVM_{L2} performing above CTQ. That is, the addition of dictionary data diminishes overall accuracy, a slightly surprising result given the complementary of corpus and dictionary data in all other aspects of evaluation. It is possible that we could get better results by treating both L1-recoverable and gold-standard translations in the training data as positive exemplars, which we leave as an item for future research.

Combining the results from Table 9 with those from Table 6, the overall silver-standard accuracy over the JE data is 0.671 for $\text{SVM}_{\text{L2+X}}$ (compared to 0.602 for CTQ), and that over the EJ data is 0.461 (compared to 0.419 for CTQ).

In summary, we have shown our method to be superior to both the baseline and CTQ over EJ and JE translation tasks in terms of both gold- and silver-standard accuracy. We also demonstrated that the method successfully combines crosslingual and target language corpus data, and is relatively robust over low frequency inputs.

5   Related work

One piece of research relatively closely related to our method is that of Cao and Li (2002), who use bilingual bootstrapping over Chinese and English web data in various forms to translate Chinese NN compounds into English. While we rely on bilingual dictionaries to determine crosslingual similarity, their method is based on contextual similarity in the two languages, without assuming parallelism or comparability in the corpus data. They report an impressive F-score of 0.73 over a dataset of 1000 instances, although they also cite a prior-based F-score (equivalent to our Baseline) of 0.70 for the task, such that the particular data set they are dealing with would appear to be less complex than that which we have targeted. Having said this, contextual similarity is an orthogonal data source to those used in this research, and has the potential to further improve the accuracy of our method.

Nagata et al. (2001) use "partially bilingual" web

pages, that is web pages which are predominantly Japanese, say, but interspersed with English words, to extract translation pairs. They do this by accessing web pages containing a given Japanese expression, and looking for the English expression which occurs most reliably in its immediate vicinity. The method achieves an impressive gold-standard accuracy of 0.62, at a recall of 0.68, over a combination of simplex nouns and compound nominals.

Grefenstette (1999) uses web data to select English translations for compositional German and Spanish noun compounds, and achieves an impressive accuracy of 0.86–0.87. The translation task Grefenstette targets is intrinsically simpler than that described in this paper, however, in that he considers only those compounds which translate into NN compounds in English. It is also possible that the historical relatedness of languages has an effect on the difficulty of the translation task, although further research would be required to confirm this prediction. Having said this, the successful use of web data by a variety of researchers suggests an avenue for future research in comparing our results with those obtained using web data.

6   Conclusion and future work

We have proposed a method for translating NN compounds which compositionally generates translation candidates and selects among them using a target language model based on corpus statistics and a translation model based on bilingual dictionaries. Our SVM-based implementation was shown to outperform previous methods and be robust over low-frequency NN compounds for JE and EJ translation tasks.

Acknowledgements

References

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

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands.

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.

Gregory Grefenstette. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Translating and the Computer 21: ASLIB'99*, London, UK.

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.

Kyo Kageura, Fuyuki Yoshikane, and Takayuki Nozawa. 2004. Parallel bilingual paraphrase rules for noun compounds: Concepts and rules for exploring Web language resources. In *Proc. of the Fourth Workshop on Asian Language Resources*, pages 54–61, Sanya, China.

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 63–9, Taipei, Taiwan.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 704–710, Montreal, Canada.

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

Mainichi Newspaper Co. 2001. Mainichi Shimbun CD-ROM 2001.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, and Yoshitaka Hirano. 1999. *Japanese Morphological Analysis System ChaSen Version 2.0 Manual*. Technical Report NAIST-IS-TR99009, NAIST.

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

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the Web as a bilingual dictionary. In *Proc. of the ACL/EACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 95–102, Toulouse, France.

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.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus volume 1 – from yesterday's news to tomorrow's language resources. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 827–33, Las Palmas, Canary Islands.

Takaaki Tanaka and Timothy Baldwin. 2003a. 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*, pages 17–24, Sapporo, Japan.

Takaaki Tanaka and Timothy Baldwin. 2003b. Translation selection for Japanese-English noun-noun compounds. In *Proc. of the Ninth Machine Translation Summit (MT Summit IX)*, pages 89–96, New Orleans, USA.