# Translation Selection for Japanese-English Noun-Noun Compounds

**Takaaki Tanaka**

NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
2-4 Hikari-dai Seika-cho, Kyoto 619-0237, Japan
takaaki@cslab.kecl.ntt.co.jp

**Timothy Baldwin**

CSLI
Stanford University
Stanford, CA 94305-4115, USA
tbaldwin@csli.stanford.edu

## Abstract

We present a method for compositionally translating Japanese NN compounds into English, using a word-level transfer dictionary and target language monolingual corpus. The method interpolates over fully-specified and partial translation data, based on corpus evidence. In evaluation, we demonstrate that interpolation over the two data types is superior to using either one, and show that our method performs at an F-score of 0.68 over translation-aligned inputs and 0.66 over a random sample of 500 NN compounds.

## 1 Introduction

This paper addresses the task of the machine translation (MT) of Japanese noun-noun (NN) compounds into English. NN compounds are defined to take the form of a concatenated noun pair, the elements of which we will refer to as $N_1^J$ and $N_2^J$ (in linear order of occurrence). Examples of Japanese NN compounds are 機械 · 翻訳 *kikai·hoNyaku* "machine translation", 民間 · 企業 *miNkaN·kigyou* "private company" and 関係 · 改善 *kaNkei·kaizeN* "improvement in relations".[1]

Our interest in NN compounds stems from the realisation that they are highly frequent and highly productive in Japanese, and translate into a variety of English constructions. We estimate that the token occurrence of Japanese NN compounds in the 1996 Mainichi Shimbun Corpus (32m word tokens, Mainichi Newspaper Co. (1996)) is roughly 10%, underlining their high frequency. The average token frequency per NN compound type is around 7, and slightly more than half of the NN compounds occur only once in the corpus (mirroring the results of Lapata and Lascarides (2003) for English). Additionally, new NN compounds are constantly evolving (Nagata et al., 2001), all of which motivate a robust translation method which is able to handle novel NN compounds. In terms of MT, Japanese-English NN compound translation is made difficult because of: (a) constructional variability in the English translations (evidenced in the English translations for the example Japanese NN compounds above); (b) lexical idisoyncracies in Japanese and English (e.g. 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"); and (c) non-compositional NN compounds (e.g. 井戸端 · 会議 *idobata·kaigi* "(lit.) well-side meeting", which translates most naturally into English as "idle gossip").

We use the Japanese-English NN compound MT task as a test case for a compositional translation method which makes use of a word-level translation lexicon and monolingual corpus data. Much work on the similar task of terminology translation has relied on parallel or comparable corpora (Fung and McKeown, 1997; Rapp, 1999; Tanaka and Matsuo, 1999; Lee and Kim, 2002; Tanaka, 2002). We attempt to use only target language corpus evidence in the translation process, to reduce data sparseness and make the method as portable to novel domains as possible.

We suggest that the method proposed in this research can be used as a specialist NN compound translation module in a full-scale MT system. This is supported by the finding of Koehn and Knight (2003) that, in the context of statistical MT, over-

[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.

all translation performance improves when noun phrases are translated independently of the sentential context.

The remainder of this paper is structured as follows. Section 2 describes the proposed method, and Section 3 outlines the resources used in this research. Section 4 provides detailed evaluation of the proposed method. Section 5 contextualises this research with respect to previous work, and Section 6 concludes the paper.

## 2 Proposed method

We translate NN compounds through the composition of word-level translations and a constructional translation template. In order to translate 関係·改善 *kaNkei·kaizeN* "improvement in relations", for example, we make use of the three independent pieces of evidence: the translations of *relations* and *improvement* for 関係 and 改善, respectively, and the (English) translation template [$N_2^E$ *in* $N_1^E$] (where $N_i^E$ indicates that the word[2] is a noun (*N*) in English ($E$) and corresponds to the $i$th-occurring noun in the original Japanese). This method has the obvious advantage that it can generally generate a translation for a given 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 Japanese NN compounds translate compositionality into English, which Tanaka and Baldwin (2003) found to be the case 43.1% of the time in a Japanese-English NN compound translation task. In this paper, we focus primarily on selecting the correct translation for those NN compounds which can be translated compositionally, but also investigate what happens when non-compositional NN compounds are translated using a compositional method.

Our method can be broken down into two basic stages: generation and selection (similarly to Cao and Li (2002) and Langkilde and Knight (1998)). **Generation** consists of taking the cross-product of all word-level translations for each of the two Japanese nouns, and slotting each such pairing into the various translation templates. Each word translation and template slot is annotated for part of speech (POS), and we constrain the generation process to output only those candidates where each slot and filler in the translation template agree in POS; e.g., no translation would be generated from the combination of the adjectival *relational* and template [$N_2^E$ *in* $N_1^E$] (with *relational* corresponding to either $N_1^J$ or $N_2^J$ in the original Japanese).

In **selection**, we take the generated translation candidates and score each, returning the highest-scoring translation candidate as our final translation. 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.

Our scoring method rates the **corpus-based translation quality** ($ctq$) of a given translation candidate according to both corpus evidence for the fully-specified translation and its parts in the context of the translation template in question. This is calculated as:

$$ctq(w_1^E, w_2^E, t) = \alpha\, p(w_1^E, w_2^E, t) + \beta\, p(w_1^E, t)p(w_2^E, t) + \gamma\, p(w_1^E)p(w_2^E)p(t) \quad (1)$$

where $w_1^E$ and $w_2^E$ are the word-level translations of the Japanese $N_1^J$ and $N_2^J$, respectively, and $t$ is the translation template.[3] Each probability is calculated according to a maximum likelihood estimate based on relative corpus occurrence. The formulation of *ctq* is based on linear interpolation, by which we use weights to combine the probabilities of the fully-specified translation candidates with those of the translation parts first conditioned on translation templates and second independently. The weights take the form of $\alpha$, $\beta$ and $\gamma$, where $0 \leq \alpha, \beta, \gamma \leq 1$ and $\alpha + \beta + \gamma = 1$. Our use of linear interpolation constitutes a basic form of smoothing, in that we would like to have some means of selecting the most likely translation in the absence of evidence for the fully-specified translation candidate.

The basic intuition behind decomposing the translation candidate into its two parts within the context of the translation template ($p(w_1^E, t)$ and $p(w_2^E, t)$ in the second term of equation 1) is to capture the subcategorisation properties of $w_1^E$ and $w_2^E$ relative to $t$. For example, if $w_1^E$ and $w_2^E$ were *Bandersnatch*

[2] Strictly speaking, word-level translations are "listemes" not words, as they can be made up of multiple English words. We always treat the word-level translation in the manner of a word, however, i.e. as a single unit with unique part of speech.

[3] $w_1^E$ and $w_2^E$ are assumed to be POS-compatible with $t$.

and *relation*, respectively, and $p(w_1^E, w_2^E, t) = 0$ 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 $p(w_2^E, t)$ greater for the template [$N_2^E$ *to* $N_1^E$] than [$N_2^E$ *on* $N_1^E$]. The third term in equation 1 ($p(w_1^E)p(w_2^E)p(t)$) represents a second level of backing-off and treats $w_1^E$, $w_2^E$ and $t$ as independent items.

## 3 Resources

The proposed method makes use of a number of resources, namely: (a) pre-processed target language corpus data; (b) a word-level translation dictionary; (c) test data over which to evaluate the method; and (d) an inventory of translation templates.

### 3.1 Corpus data

The corpus data was taken from the 80m word written component of the British National Corpus (BNC, Burnard (2000)). The British National Corpus was chosen because of its size and domain-inspecificity. Corpus size will affect the relative coverage of fully-specified translations, and the fact that the BNC covers a broad range of domains helps us to accurately capture the subcategorisation properties of a given word (in the form of $p(w_i^E, t)$, as described in Section 2).

We dependency-parsed the BNC using RASP (Briscoe and Carroll, 2002), a tag sequence grammar-based stochastic parser. RASP captures noun-noun dependencies using the `ncmod` relation between a head noun and its noun dependent, optionally linked via a preposition or genitive construction. The noun-noun dependency structure of the NP *the Jubjub bird's relation to the frumious Bandersnatch*, for example, would be captured by three `ncmod` relations:

```
ncmod(_,bird,Jubjub)
ncmod(POSS,relation,bird)
ncmod(to,relation,Bandersnatch)
```

which represent the NN, genitive and *to*-PP construction, respectively. Note that the head of the dependency relation occupies the second position in the tuple, and the dependent the third position. All `ncmod` tuples are normalised for number and case using morph (Minnen et al., 2001), meaning that it is possible to sum up the token count of each `ncmod` triple type, and convert it directly into a maximum likelihood-based probability.

### 3.2 Translation dictionary

The word-level translation dictionary used in this research is the **ALTDIC** dictionary. ALTDIC 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.

In order to make ALTDIC directly compatible with the BNC-derived RASP tuples, we: (a) converted any American spellings to British spellings, and (b) lemmatised the spelling-normalised words using morph and the POS tags supplied in ALTDIC. Spelling normalisation was based on simple lookup in the VARCON table of American-British spelling variants.[4]

### 3.3 Test data

The primary test data used in this research comprises 500 Japanese NN compounds extracted from the 1996 Mainichi Shimbun Corpus (Mainichi Newspaper Co., 1996), as described in Tanaka and Baldwin (2003). We first segmented and tagged the corpus using ALTJAWS[5] and then extracted out all NN bigrams adjoined by non-nouns. We next filtered off all NN compounds with a token occurrence of less than 10. As our test data, we took the 250 most frequent NN compounds, and a random selection of 250 NN compounds from the remainder of the extracted data.[6]

In order to evaluate translation accuracy over the test data, we generated a unique **gold-standard translation** for each Japanese NN compound to represent its optimally-general English translation. This was done with reference to the ALTDIC dictionary and the public domain EDICT dictionary (Breen, 1995), although a significant number of novel translations were generated due to: (a) NN compounds not occurring in either dictionary, and (b) the dictionary translations being inappropriate.

We next normalised all English translations by: (a) converting any American spellings to British

[4] `http://wordlist.sourceforge.net/`

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

[6] In fact, some post-filtering of NN compounds took place, in that a small number of NN compounds extracted out in the first sample were not genuine compounds, in which case an alternate compound was sought.

| *Template* | *Example* |
|---|---|
| [$N_1^E$ $N_2^E$] | 市場 · 経済 *shijou·keizai* "market economy" |
| [$J_1^E$ $N_2^E$] | 医療 · 機関 *iryou·kikaN* "medical institution" |
| [$N_2^E$ of $N_1^E$] | 威信 · 低下 *ishiN·teika* "loss of dignity" |
| [$N_2^E$ $N_1^E$] | 賛成 · 多数 *saNsei·tasuu* "majority agreement" |
| [$N_1^E$ $VG_2^E$] | 情報 · 収集 *jouhou·shuushuu* "information gathering" |
| [$J_1^E$ $VG_2^E$] | 調査 · 報道 *chousa·houdou* "investigative reporting" |

Table 1: Example translation templates (J = adjective, VG = gerund)

| *Dataset* | $N$ | *Mean gold-standard translations* | *Mean generated translations* | *Baseline-1 ($\alpha = 1$)* | *Baseline-2 ($\beta = 1$)* | *Best F-score* |
|---|---|---|---|---|---|---|
| $\text{ALIGN}_{\text{GOLD}}$ | 224 | 1.0 | 60.3 | 0.46 | 0.37 | 0.49 |
| $\text{ALIGN}_{\text{RECOVER}}$ | 224 | 2.7 | 60.3 | 0.64 | 0.59 | 0.68 |
| NIKKEI | 401 | 2.2 | 54.8 | 0.45 | 0.39 | 0.46 |
| TERMDICT | 2245 | 1.3 | 45.7 | 0.47 | 0.35 | 0.48 |

Table 2: An outline of the translation datasets

spelling; (b) tokenising words (largely to separate possessive suffices from words); (c) lemmatising all component words using morph (hence normalising number and case); and (d) deleting all determiners. E.g., *citizens' group* would be normalised to *citizen 's group* and *all countries of the world* to *all country of world*.

We then examined each Japanese-English translation pair to determine if both $N_1^J$ and $N_2^J$ aligned in the English translation via word-level translations contained in ALTDIC. If such alignment was found to hold, we further checked that there were no unaligned open class words in the translation, based on the output of the RASP POS tagger for that string. If this condition was found to hold, we classified that translation pair as being an **aligned NN compound**, and any residual translation pairs were classified as **unaligned NN compounds**. With aligned NN compounds, we made note of the pattern of alignment for use as a translation template (see below). A total of 224 NN compounds were classified as belonging to the aligned NN compound class, leaving 276 unaligned compounds. Below, we refer to the set of aligned NN compounds with unique gold-standard translations as $\text{ALIGN}_{\text{GOLD}}$, and the set of unaligned compounds (also with unique gold-standard translations) as $\text{UNALIGN}_{\text{GOLD}}$.

For the aligned NN compound subset, we constructed a second set of **source language-recoverable translations** (incorporating the original gold-standard translation). This was done in order to give the method credit for "near-miss" translations, that is translations which are syntactically unmarked, capture the basic semantics of the source language expression and from which the source language expression is recoverable with reasonable confidence. Examples include *issue of blame* and *responsibility issue* as alternative translations for 責任問題 *sekiniN·moNdai* "liability issue". In this, we set ourselves apart from the highly subjective method of manual translation evaluation where bilingual annotators are presented with the source language expression and system output, and asked to rate the output for "plausibility" or "usability" (Tanaka and Matsuo, 1999; Cao and Li, 2002). By pre-generating our source language-recoverable translations (or L1-coverable translations) we remove the annotator from direct contact with our method and hence hope to make evaluation as objective as possible. Given that we are only ever required to evaluate translations generated by our method, we only consider those translation candidates our method can generate for a given source language expression using ALTDIC. To reduce the annotation overhead, we break the task down into two steps: (1) identify appropriate word-level translations for each member of the NN compound as conditioned by the context of that compound, and (2) generate all translation candidates using only those word-level translations from step 1, and select from among them. In

this way, we reduce the number of translation candidates per input the annotator must look at by around two-thirds as compared to a one-pass method of sifting through all translation candidates we are able to generate. Below, we refer to this set of aligned NN compounds with source language-recoverable translations as $\text{ALIGN}_{\text{RECOV}}$.

We use another two datasets in secondary evaluation. The first consists of frequently-occurring compounds found in the Nikkei 1996 corpus (Nikkei Publishing Co., 1996). Each NN compound was translated by a professional translator, guided by 2 sentences from the Nikkei 1996 corpus containing that compound. The translator was asked to provide suitable translations for each NN compound, with no stipulation on the number or linguistic form of the translations. Out of the translated NN compound data, we selected those which were: (a) not found in the 500 NN compounds from above, and (b) compositionally translatable (i.e. both component Japanese nouns were contained in the word-level translation dictionary). This resulted in 401 items, which we refer to as NIKKEI below.

The second dataset used in secondary evaluation originates from a Japanese-English terminological dictionary. We extracted out all Japanese NN compounds which satisfied the two conditions outlined above for the NIKKEI dataset. This resulted in 2,245 items, which we refer to as TERMDICT below.

### 3.4 Translation templates

A total of 28 translation templates were used in this research, a sample of which are shown in Table 1. They were determined by combining all POS-conditioned alignment mappings between Japanese NN compounds and their English translations found in $\text{ALIGN}_{\text{GOLD}}$. One noteworthy translation template is $[\text{N}_2^E\ \text{N}_1^E]$, where the English translation takes the form of an NN compound but the order of the nouns is reversed. That is, in the example 賛成・多数 *saNsei·tasuu* "majority agreement" from Table 1, 多数 *tasuu* translates as *majority* and 賛成 *saNsei* as *agreement*, the order of which is then reversed in the English translation.[7]

## 4 Evaluation

In this section, we evaluate the translation candidate scoring method over the $\text{ALIGN}_{\text{GOLD}}$, $\text{ALIGN}_{\text{RECOV}}$, NIKKEI and TERMDICT datasets according to translation F-score relative to the model translations (see below). We test the effect of varying the values of $\alpha$ and $\gamma$ in equation 1 (with the value of $\beta$ being determined by $1 - \alpha - \gamma$), and also run the method over $\text{UNALIGN}_{\text{GOLD}}$ to test its robustness over data for which translation compositionality does not hold.

We evaluate translation performance according to the standard measures of precision, recall and F-score. Precision is the relative proportion of inputs for which we generate a correct translation (as determined by the translation data set we are evaluating against), recall is the relative proportion of inputs for which we are able to generate a translation, and F-score is the harmonic mean of the two.

We evaluate our method against two baselines derived from $ctq$. The first baseline (Baseline-1) takes the most probable fully-specified translation candidate (i.e. is equivalent to setting $\alpha = 1$, $\beta = 0$ and $\gamma = 0$ in equation 1). The second (Baseline-2) scores translation candidates according to template-specified partial translation probabilities (i.e. is equivalent to setting $\alpha = 0$, $\beta = 1$ and $\gamma = 0$ in equation 1). Baseline-1 is prone to low recall as a result of there being no fully-specified translation candidate attested in the corpus. Baseline-2 is prone to low precision as it does not consider the lexical affinity between the word-level translations. An outline of baseline F-scores for each of the datasets, along with the total number of items, the average number of gold-standard transaltions per item and the average number of translation candidates generated per item, is given in Table 2. We also indicate the best F-score obtained for each dataset, obtained from Figures 1–3.

The results for $\text{ALIGN}_{\text{GOLD}}$ and $\text{ALIGN}_{\text{RECOV}}$ are given in Figure 1, and those for NIKKEI and TERMDICT are given in Figure 2 and Figure 3, respectively. In each case, $\alpha$ is plotted on a logarithmic scale on the $x$ axis, and separate curves are given for $\gamma$ values of 0.00, 0.01, 0.10 and 0.20. Note that a given combination of $\alpha$ and $\gamma$ values will uniquely determine the $\beta$ value. For each curve, a peak is seen for an $\alpha$ value of around 0.8 and $\gamma$ value of around 0.1, above both baselines. Baseline-1 is superior to Baseline-2 in all cases. This suggests that fully-specified translation data (i.e. $p(w_1^E, w_2^E, t)$) is vital in translation selection, but that some degree of smoothing is required via partially-specified translation data (i.e. $p(w_1^E, t)p(w_2^E, t)$) and fully-

[7] One explanation for this word order reversal is that the Japanese compound is left-headed.

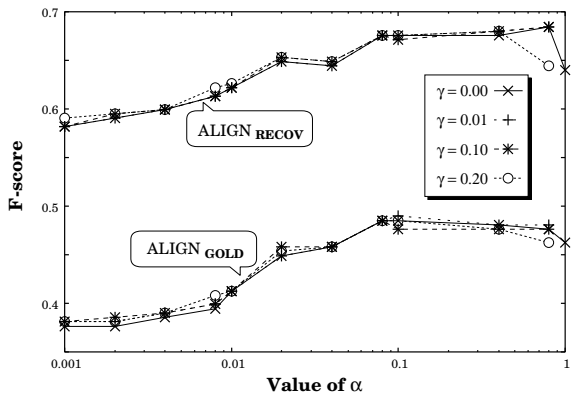


Figure 1: Translation F-scores for $\text{ALIGN}_{\text{GOLD}}$ and $\text{ALIGN}_{\text{RECOV}}$

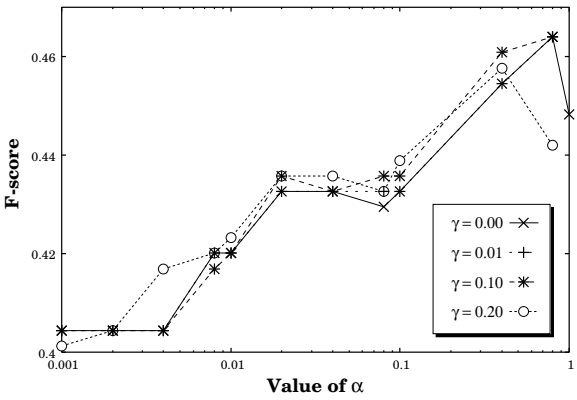


Figure 2: Translation F-scores for NIKKEI

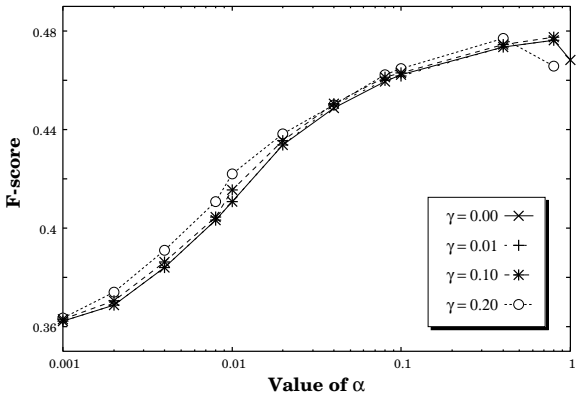


Figure 3: Translation F-scores for TERMDICT

| ***Class*** | ***Basic*** | ***+MBMT*** |
|---|---|---|
| Gold-standard | 0.00 | 0.14 |
| L1-recoverable | 0.23 | 0.20 |
| Basic sense-recoverable | 0.19 | 0.14 |
| Nonsensical | 0.02 | 0.02 |
| Misleading | 0.19 | 0.18 |
| No translations | 0.35 | 0.31 |

Table 3: Breakdown of results over $\text{UNALIGN}_{\text{GOLD}}$

indepedent data (i.e. $p(w_1^E)p(w_2^E)p(t)$).

The L1-recoverable translation F-score of around 0.68 for $\text{ALIGN}_{\text{RECOV}}$ can be interpreted as reflecting the performance of the method in a real-world MT application, where understandability often takes precedence over prescriptive correctness.

Next, we look to the unaligned NN compound data. While we are unable to compositionally generate the gold-standard translation for these inputs, we wish to know how far off the mark the translations we output are. To evaluate the quality of the translation output, we employ the 6-way classification of translation output quality detailed in Table 3. The classification incorporates the gold-standard and L1-recoverable translation classes as used above, but also classes for: no translation output (due principally to either $\text{N}_1^J$ or $\text{N}_2^J$ not being contained in ALTDIC, ***No translations***); basic sense recoverability but where the translation is either syntactically marked or stilted (***Basic sense-recoverable***); nonsensical translations, where the translation is not immediately interpretable (***Nonsensical***); and lastly, misleading translations where there is a contrast in semantics in the source and target languages (***Misleading***).

The results over the $\text{UNALIGN}_{\text{GOLD}}$ dataset are given in Table 3. The ***Basic*** column presents the proposed method applied as described herein, for which over 40% of outputs are either L1-recoverable or basic sense-recoverable. Over one-third of inputs cannot be translated, and only about 20% receive misleading or nonsensical translations. ***+MBMT*** couples the proposed method with memory-based MT (MBMT) in the form of a listing of translation data for Japanese NN compounds as found in the combined ALTDIC and EDICT dictionaries. MBMT is applied as a pre-processor in outputting the dictionary-derived translation directly if the NN compound is found to exist in the dictionaries, falling back to the proposed method only if this fails. This "cascaded" method was shown by Tanaka and Baldwin (2003) to be an effective

way of combining the high-precision of MBMT with the high-recall of compositional translation. For the UNALIGN$_{GOLD}$ dataset, the cascaded translation methodology results in nearly 50% of outputs being assigned classes 1–3, and enhances translation coverage slightly. The combined L1-recoverable F-score over the full 500-element set (in combination with MBMT, based on classes 1–2 in the case of UNALIGN$_{GOLD}$) is 0.66.

One significant area in which our method falls down is that it treats all translations contained in the transfer dictionary as being equally likely, where in fact there is considerable variability in general applicability. 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, but 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*. There are two immediate sources of such conditional probabilities: dictionary alignment data and parallel/comparable corpora. Given that we are committed to using a monolingual corpus, the former method is preferred.

## 5 Related work

One piece of research relatively closely related to our method is that of Cao and Li (2002), who use bilingual web data and various combinations of the EM algorithm, a naive Bayes classifier and TF-IDF to translate Chinese NN compounds into English. 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-1) 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.

Lee and Kim (2002) use definitions from a bilingual dictionary and target language corpora, and treat translation selection as disambiguation of a source word sense and selection of a target word. They report the accuracy of their method for nouns to be 0.55, although in the case of compound nouns it seems to be overkill to rely on such rich semantic resources to achieve relatively modest precision.

Fung and McKeown (1997) extract terminology translations based on the assumption of crosslingual distributional similarity, as defined by seed and previously-extracted translation pairs across comparable corpora. Their results show the usefulness of comparable corpora, but also underline the dependence of the method on closely-correlated crosslingual corpora, which can sometimes be an unreasonable expectation.

Rackow et al. (1992) look at a German–English compound noun MT task over a range of translation templates. They propose the use of "default constructions" for single words in a manner similar to that described in this research, but base determination of such defaults on English translations within a parallel corpus. Specifically, they take the Hansards corpus and observe which of *ecology* and *ecological*, e.g., occurs more often as a noun premodifier and use this information in generating *ecological movement* rather than *ecology movement* as a translation for the German *Umweltbewegung*. Unfortunately, no attempt is made to evaluate the method.

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 filters out the effects of translation template selection by considering 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

We have proposed a method for translating NN compounds and applied it to a Japanese-English MT task. The method interpolates over translation probabilities of different levels of specification, and returns the highest-scoring translation from amongst them. In evaluation, we showed our method to perform at an F-score of 0.68 over aligned data and 0.66 over a random sample of 500 NN compounds.

**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 two anonymous reviewers for their valuable input on this research.

## References

Jim Breen. *Building an electronic Japanese-English dictionary*. Japanese Studies Association of Australia Conference (`http://www.csse.monash.edu.au/~jwb/jsaa_paper/hpaper.html`), 1995.

Ted Briscoe and John Carroll. 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, 2002.

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

Yunbo Cao and Hang Li. 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, 2002.

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

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

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 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, 1991.

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

Irene Langkilde and Kevin Knight. 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, 1998.

Mirella Lapata and Alex Lascarides. Detecting novel compounds: The role of distributional evidence. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, 2003.

Hyun Ah Lee and Gil Chang Kim. Translation selection through source word sense disambiguation and target word selection. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 530–536, Taipei, Taiwan, 2002.

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

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

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 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, 2001.

Nikkei Publishing Co. Nikkei CD-ROM 1996, 1996.

Ulrike Rackow, Ido Dagan, and Ulrike Schwall. Automatic translation of noun compounds. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, pages 1249–53, Nantes, Frances, 1992.

Reinhard Rapp. 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, 1999.

Takaaki Tanaka. 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, 2002.

Takaaki Tanaka and Timothy Baldwin. 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, 2003.

Takaaki Tanaka and Yoshihiro Matsuo. 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, 1999.