Analogical translation of unknown words in a statistical machine translation framework

Etienne Denoual
National Institute of Information and Communications Technology
and ATR - Spoken Language Communication Group
Keihanna, 619-0288 Kyoto, Japan
etienne.denoual@nict.go.jp

Abstract

In this paper we address the problem of translating unknown words in a statistical machine translation framework. In data-driven machine translation, words that are not seen in the data may not be translated and are either discarded or left as is in the output. They are refered to as unknown words. The unknown word problem increases when the available bilingual data is scarce. In order to address this problem, we propose to use proportional analogy at the character-level to translate unknown words. We study and report results of the integration of this approach into a statistical machine translation system translating from Japanese to English with relatively scarce resources. Objective evaluation measures suggest that the translated sentences have a higher adequacy than that produced by a baseline system , while their fluency is similar.

Introduction

In data-driven machine translation, computation is often performed at the level of what intuition hints at being a word in segmenting languages. In the framework of statistical machine translation (Brown et al., 1990), language resources are typically segmented into such tokens before they may be used as training data. These words or tokens may take into account other linguistic information and features made explicit by, say, morphological analysis, the presence of punctuation, etc. Therefore, there is not one segmentation scheme, but many acceptable ones. For instance in the case of machine translation, optimal segmentation highly depends on the language pair that is considered and is, paradoxically, often linguistically unintuitive.

One of the most important problems of data-driven machine translation is that posed by unknown words: in the process of translating, a system is bound to encounter words that were unseen in the available training data. While this is in part due to the aforementioned segmentation issues, it is also often simply due to the lack of training data. This issue becomes exponentially plaguing as available linguistic resources get scarce, and as low-occurence content words often remain untranslated. A number of works have tackled the unknown word problem: (Sinha, 2005) uses a heuristic-based identification and translation method, (Uchimoto et al., 2001) propose a model based on morphological analysis and large amounts of lexical information contained in a dictionary, while (Nagata, 1999) proposes to estimate Part-Of-Speech information of unknown words using a statistical model of morphology and context.

In this work, we propose to address the problem of translating unknown words by going at a lower level than that of words, and to translate unknown words as strings of characters. We propose to use proportional analogy on character strings to capture commutations that occur inside words, in order to grasp more linguistic information from the available data and to compensate insufficient segmentation schemes.

Following this introduction, we first describe the mechanism of proportional analogy on character strings. We then show how proportional analogy allows to translate strings of characters given a set of example data. This translation method is then set in a statistical machine translation (SMT) framework: the system provides additional aligned data that feed the unknown word analogical translation engine, and in a second pass translates the test set with the translated unknown words.

We evaluate the proposed method by performing an experiment in Japanese to English translation using data originating from the IWSLT evaluation campaign. So as to put ourselves in a position where only scarce resources are available, we reproduce the conditions of the IWSLT OPEN track, in which only $40,000$ sentence pairs where available. Translations are then evaluated using automatic evaluation measures BLEU and NIST. Finally, we discuss results in terms of translation quality and show that the proposed method yields a significantly higher adequacy of sentences while maintaining their fluency when compared to a baseline translation system.

Proposed method

Proportional analogy at the character-level

Analogical computation exploits equations of the form:

$$A : B :: C : x$$

where A, B, and C are character strings. The equation may then yield a solution x = D in the form of another character string. This operation allows to capture lexical and syntactical variations along paradigmatic and syntagmatic axes, without explicitly decomposing the strings into fragments. The operation does not require any preprocessing of the aligned examples, such as a step of segmentation into words, as it operates strictly at the character-level. While translating, proportional analogy distributes the information at a lower level than that of words, all over the target string of characters. An algorithm has been developed (Lepage, 1998) , that allows to determine if four terms are in an analogy, and to generate a fourth term if three terms are placed in an analogy. The algorithm may then yield zero, one, or many solutions to the equation.

Analogical translation of unknown words

In this work, we propose to use analogy on character strings in order to translate unknown words. As argued in (Denoual, 2006), Natural Language Processing approaches that rely on word tokens as the single unit of processing neglect the fact that corresponding pieces of information in different languages are distributed over the entire strings of data, and therefore do not necessarily correspond to complete words.

The algorithm that we use here is similar to that exposed in (Lepage and Denoual, 2006). Suppose that we want to translate the character string D, and that we have a bilingual corpus at our disposal. We first put D in a analogical equation[1] in the source part. We then form all analogical equations with the input sentence D and with all relevant[2] pairs of sentences (A_i, B_i) from the source part of the bilingual corpus:

$$A_i : B_i :: x : D$$

The application of the above-mentioned algorithm may allow to solve the equation and yield a solution x.

If x does not belong to the source part of the bilingual corpus, we try to translate x recursively in the same manner until one solution is part of the corpus.

If $x = C_{i,j}$ belongs to the source part of the bilingual corpus, we may then use its translation $\widehat{x} = \widehat{C_{i,j}}$ to form all possible analogical equations in the target side of the bilingual corpus:

$$\widehat{A_i}^k : \widehat{B_i}^k :: \widehat{C_{i,j}}^k : y$$

Such equation may yield a solution $y = \widehat{D_{i,j}}^k$, which is a translation of the character string D. As different equations may yield different or identical solutions, translations are sorted by their frequencies and the top one is selected.

Example

Suppose that we want to translate the following Japanese string into English:

ニューヨーク港 ‘[3]
/nyūyōkukō/

Among all possible pairs of strings in the Japanese part of the corpus, we may find the following two Japanese strings:

上海
/shanhai/ ↔ *Shanghai*

上海港
/shanhaikō/ ↔ *Shanghai Harbor*

This allows us to form the following equation:

$$\text{上海} : \text{上海港} :: x : \text{ニューヨーク港}$$

This equation yields a solution $x = \text{ニューヨーク}$. If this string belongs to the corpus, we therefore know its translation:

ニューヨーク
/nyūyōku/ ↔ *New York*

We may then form the following equation on the English side:

$$\text{Shanghai} : \text{Shanghai Harbor} :: \text{New York} : y$$

Solving this equation yields $y =$ *New York Harbor*, which is by construction a translation of the Japanese string ニューヨーク港 .

Those correspondences are best shown in the view of a parallelopiped, as shown in Figure 1.

Including the method in a statistical machine translation (SMT) system

In data-driven machine translation, words that are not seen in the training data may not be translated and are either discarded or left untranslated in the output. This unknown word problem increases when the available bilingual data is scarce. For instance, in the context of statistical machine translation a bilingual corpus of training data may be used to produce word alignments. Those word alignments may then be used in the form of a lexical translation table, in order to extract consistent phrase pairs. However, if at translation time a word or word sequence is not found in the phrase table, no translation may be retrieved. The tokens remain untranslated, which considerably degrades translation quality.

We propose to apply our method in order to alleviate this problem. Suppose that we have a bilingual corpus available as training data for a statistical machine translation system, and a test set consisting of

[1] Let us stress that the equation is entirely monolingual.

[2] Theoretically, all possible pairs of sentences in the corpus should be placed in the analogical equation. Because this is not feasible practically for a large corpus, heuristics may be used to select the most relevant pairs of sentences.

[3] Lit.: *New York Harbor*

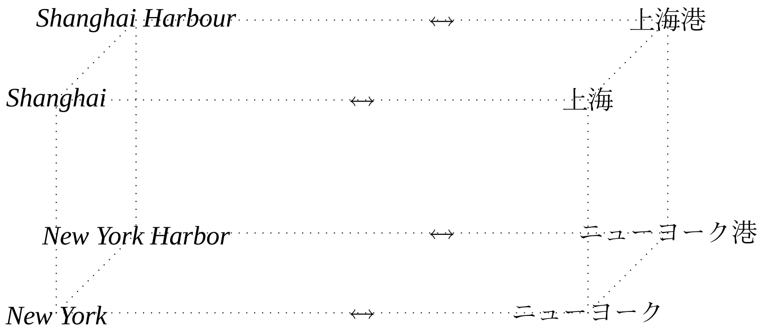

Figure 1: View of the parallelopiped: four terms in each language form a monolingual proportional analogy.

sentences in the source language, that we wish to translate in the target language. Given the available training data, we first build a statistical machine translation system, and translate the test set. The machine translation output includes unknown words, that are automatically extracted and gathered in what we refer to as the *unknown words set*. As argued above, in order to translate this unknown words test set, proportional analogy requires bilingual data. While the given training corpus may be used as is to directly translate, we wish to use additional aligned data that is consistent in terms of size with what we wish to translate: tokens consisting in short character strings. This additional data may be extracted from the lexical translation table that was estimated from statistical word alignments when building the statistical machine translation system. While low-probability alignments may be discarded, extracting the N-first alignments for each source target word allows to conveniently build a basic dictionary. For instance, a word in the source language may be aligned to many words in the target language as in the following example:

卵 /tamago/	↔	*eggs*	0.2989691
		egg	0.2222222
		flakes	0.100000
		boiled	0.0175439
		done	0.0103093
		sir	0.0052910
		your	0.0018643
		want	0.0004474
		like	0.0002205

In this case, it seems reasonable to retain the two top-ranked alignments. Of course, the dictionary may be unaccurate due to word alignment errors. In order to retain an acceptable accuracy, selection may be performed by using a threshold on the lexical translation probability. After the dictionary is extracted, it is concatenated to the original training data.

The unknown words test set is then translated using the analogical translation method described above. The result of this translation is then added to the word alignments that were estimated when building the original statistical machine translation system, and a second training step is performed in order to reestimate a new lexical translation table and eventually a new phrase table. By construction, the new phrase table includes analogical translations of the unknown words contained in the test set.

In the next section, we perform a translation experiment in order to assess the performance of the proposed method.

Experiments

Data

In order to assess the performance of the proposed method, we perform a translation experiment by reproducing some of the conditions of the International Workshop on Spoken Language Translation 2006 (IWSLT06) campaign as described in (Paul, 2006). In what was refered to as the *OPEN track*, a relatively small subset of approximately $40,000$ Japanese-English sentence pairs was provided to participants in order to train their systems. In this resource, the sentences are quite short, as the figures in the following Table 1 show. Japanese data, which do not usually include spaces, are provided tokenized. As the same sentence may appear several times with different translations, the number of unique sentences in each language is indicated in Table 1.

In order to evaluate our system, we use the 500 sentences test set that was used in the IWSLT06 campaign to evaluate Japanese to English machine translation systems. The quality of translations is then systematically assessed using the well known BLEU and NIST automatic evaluation measures. BLEU (Papineni et al., 2002) is known to show a better correlation with *fluency*, whereas NIST (Doddington, 2002) shows a better correlation with *adequacy*. Each translation is evaluated with 7 reference sentences produced by human translators.

Results and evaluation

Using the bilingual training corpus, we first build a standard phrase-based statistical machine translation system: bidirectional word alignments are obtained

Table 1: Training data statistics.

	Unique sentences	Tokens per line
English	39,602	8.06
Japanese	36,776	10.09

by using the GIZA++ implementation of IBM Model 4 alignments (Och and Ney, 2003), and are used to produce a bilingual phrase table (Koehn et al., 2003). The language model is a standard trigram model with Kneser-Ney smoothing, trained using the SRI language modeling toolkit (Stolcke, 2002). In a first pass, we translate the test set using the Pharaoh decoder (Koehn, 2004), which implements a heuristic beam search for phrase-based translation.

The translated set of 500 sentences includes English tokens, and untranslated Japanese tokens. The translated test has a total size of 5978 tokens (1102 unique tokens), of which 376 are untranslated Japanese tokens. A common practice in order to improve automatic evaluation scores is to remove untranslated tokens: indeed, as they may not be found in the references, removing them can only improve performance. This is shown in the following table, where we evaluate both the raw output containing the unkown words, and the output when unknown words are removed.

	BLEU	NIST
With UWs	0.1638	5.8054
Without UWs	0.1790	5.7627

As can be seen from those figures, removing all unknown words from the output yields a higher BLEU score ($+9.3\%$ relative value), which accounts for a better fluency of sentences. Indeed, intuitively, removing ”noise” may only improve the fluency of sentences. The NIST score however does not vary significantly (-0.7% relative value): again intuitively, removing unknown words should neither significantly improve nor damage the adequacy of sentences (that is, the information they convey).

All untranslated Japanese tokens are then extracted, to be used as the unknown words test set for our proposed unknown word translation method. Table 3 shows the number and length distribution in characters of the extracted unknown words. One unknown word has an average length of 3.175 Japanese characters.

As argued before, in order to analogically translate unknown words, we need aligned examples. In addition to the original training set of aligned sentences, we extract aligned words from the IBM Model 4 generated Japanese to English lexicon. In this first experiment, we retain only word to word alignments which have the highest probability (i .e . one word in Japanese has only one aligned English word). This dictionary of $12,536$ words is then concatenated to the original training data and constitutes the bilingual data that is used to translate unknown words.

Unknown words are then translated and added with their translation to the statistical word alignments, and the statistical machine translation system models are reestimated: the new phrase table now contains the unknown words. We retranslate the test set and evaluate translation quality with objective measures BLEU and NIST. Results are shown on Table 4 at the row entitled *1-Best lexicon*.

When compared to the quality of the translated test set containing unknown words (entitled *Baseline (with UWs)*), there is a significant gain both in terms of BLEU and NIST measure ($+1.09\%$ absolute in BLEU and $+0.2502$ in NIST). However, when compared to the translated test set from which unknown words are automatically removed, there is a small loss in BLEU, but still a significant gain in NIST (-0.43% absolute in BLEU and $+0.2929$ in NIST). If we interpret BLEU and NIST scores in terms of fluency and adequacy as previously assumed, our approach to translation of unknown words always yields a better adequacy (more information is conveyed in the translation), while moderately hurting fluency. This slight decrease in fluency may have two causes: the translation of unknown words can be erroneous in some cases, which should lower both BLEU and NIST scores; if the translation is correct, word to word translation does not allow the statistical machine translation system to cope with the order of words when reestimating models. Because BLEU values more word order (therefore fluency) while NIST values more low-occurence content words (therefore adequacy), in the latter case the BLEU score logically decreases while the NIST score increases.

In order to understand how the addition of translated unknown words contributes to overall translation quality, we sort the translated unknown word list according to the length of the unknown words, to the number of solved analogies, and to the number of attempted analogies in the translation process. We then choose to only include increasing amounts of translated unknown words to the statistical machine translation system's word alignments, in decreasing order (i.e. in the case of the list being sorted by length, we first include the longest unknown words, and end by adding the entire list). The system is then reestimated, and the test set is translated. Figure 2 shows the results in terms of translation quality.

As can be seen from the corresponding BLEU and NIST scores, there indeed appears to be a trade-off between fluency and adequacy. In all three cases, whether unknown words are sorted by length, number of solved

Table 2: Two examples of translation: on the first line is the original sentence, on the second and third line is the baseline translation with and without unknown words, and on the fourth line is the translation produced pby the proposed method. The three lines below are human produced reference translations. Note that in the first example, the proposed method does not allow to translate the first Japanese unknown word.

Original sentence	御 客 様 の 氏名 国籍 職業 を ご 記入 下さい
Baseline with UWs	*your name and* 国籍 *fill in the* 職業
Baseline without UWs	*your name and fill in the*
Proposed method	*your name and fill in the profession*
Reference 1	*your name citizenship and occupation are required here*
Reference 2	*please fill in your name citizenship and occupation*
Reference 3	*please write down your name citizenship and occupation*
Original sentence	ドライ 赤ワイン と ドライ 白ワイン 両方 共 ございます どちら が よろしい です か
Baseline with UWs	*we have red wine* ドライ *and white wine* ドライ *both which would you prefer*
Baseline without UWs	*we have red wine and white wine both which would you prefer*
Proposed method	*we have a dry red wine and dry white wine both which would you prefer*
Reference 1	*we have both dry red wine and white wine which kind would you like*
Reference 2	*we have both dry red wine and white wine which one would you like*
Reference 3	*we serve both dry red and white wine which one would you like*

Table 3: Unknown words statistics.

Length	1	2	3	4	5	6	7	8	9	10+	total
Number	24	172	74	45	19	15	8	6	8	5	376

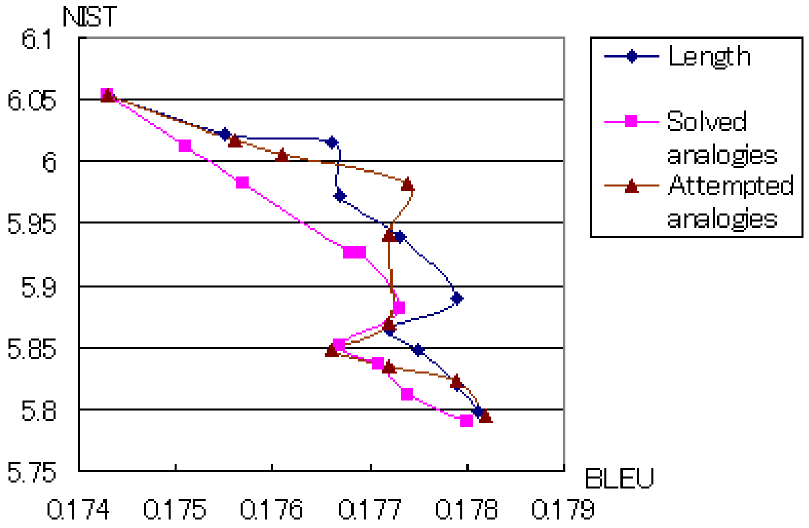

Figure 2: Variation of translation quality in terms of objective measures BLEU and NIST, when the unknown word list is translated decreasingly by length, number of solved analogies, and number of attempted analogies.

or attempted analogies, top-ranked and bottom-ranked words appear to hurt the BLEU score while providing a constant increase in NIST score. However including only unknown words of average length, or number of solved or attempted analogies provides for an increase in NIST score while the BLEU score is maintained. This is particularly true when the list is sorted by number of attempted analogies. Results on the last row of Table 4 show that including only the average third of the unknown words and their translations to the reestimated statistical machine translation system yields scores that are superior to that of the baseline both in terms of BLEU and NIST scores (+0.05% absolute in BLEU and +0.1566 in NIST).

In order to understand how the extracted dictionary influences the translation of unknown words, we perform two additional experiments: in a first experiment, instead of extracting only the alignment with the highest probability (the *1-Best*) from the lexical translation table, we also extract the alignment with the second best probability (the *2-Best*). In a second experiment, we use lexical translation tables that were estimated over a much larger set of $822,000$ lines of CSTAR-type data[4] to extract the dictionary. Only 1-Best alignments are extracted in this case so that this system may be compared with the 1-Best system using $40,000$ sentence pairs. In both cases, the system is then reestimated, and the test set is translated. Figure 3 shows the results in terms of translation quality when increasing amounts of translated unknown words are added to the statistical machine translation system's word alignments, the list being sorted by the number of attempted analogies.

The 2-Best lexicon system is very close in terms of performance to the 1-Best lexicon system while providing a consistent (although small) gain in BLEU score. Extracting a 2-Best lexicon is in fact tantamount to performing paraphrasing on the target language side. On the other hand, unknown word translation does not appear to benefit from a dictionary extracted from a much larger lexicon. Both BLEU and NIST scores are found to be lower than that obtained for the $40,000$ sentence pairs system. This decrease in performance may be explained by the complexity of the algorithm, which is basically square in the amount of data. For this matter, heuristics are used to retrieve sentence pairs from the corpus in order to form analogical equations, and the translation process for one word is given a time limit. As the amount of available examples gets larger, analogy may find it harder to find relevant pairs in the same time limit, or may not have enough time to perform as many recursive translations and therefore be short of finding a suitable translation.

Table 2 shows two practical examples of translated sentences: test sentences are first translated with the baseline, and then with the proposed method. We also display the first three references used by objective evaluation measures.

Discussion and future work

In this work, we proposed to address the problem of translating unknown words by going at a lower level than that of words, and to translate them as strings of characters using proportional analogy. We showed that such analogical translation system could be set in a statistical machine translation framework and be used to alleviate the problem of translating unknown words. Working at the level of characters allows to grasp more linguistic information by capturing commutations that occur inside words.

The performance of the proposed method was evaluated in an experiment in Japanese to English translation using relatively scarce training data, originating from the IWSLT evaluation campaign. Translation quality was then evaluated using objective measures BLEU and NIST and compared to that of a baseline system. The proposed method was found to consistently improve NIST scores while maintaining BLEU scores at a similar level. Assuming that BLEU is more correlated with fluency and NIST with adequacy, the proposed method tends to increase the adequacy of translations (more information is conveyed) while the fluency of sentences remains approximately similar to that produced by the baseline. We also showed that translation quality could benefit from translating only part of the unknown words: when unknown words are sorted according to the number of analogical equations attempted while translating, selecting only average values even allows to slightly improve fluency while consistently improving adequacy.

Future work may be undertaken at two levels: at the lower level of analogical translation of the unknown words, better results should be obtained by investigating more efficient heuristics in order to improve the ratio of solved analogies over the total number of attempted analogies. Furthermore, because word segmentation may be erroneous in the first place it should be interesting to translate series of character substrings and to study recombination schemes at the character-level. At the higher level of integrating analogical translation into a statistical machine translation system, more efficient reordering methods should be investigated. Adding translated unknown words to the original statistical word alignments does not allow for good enough phrase table estimation, therefore having a negative impact on objective evaluation measures such as BLEU or NIST, which basically operate by counting word sequences.

Nevertheless, those preliminary results are encouraging: the proposed method may be easily integrated into a standard phrase-based statistical machine translation system, it is entirely language independent and yields satisfying results by raising the adequacy of translations in terms of objective evaluations measures.

References

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Etienne Denoual. 2006. *Méthodes en caractères pour le traitement automatique des langues*. Thèse de doctorat, Université Joseph Fourier.

[4] This corresponds to a system that was also used in the IWSLT06 campaign, in the *CSTAR track* that included no restriction on the amount and nature of available training data.

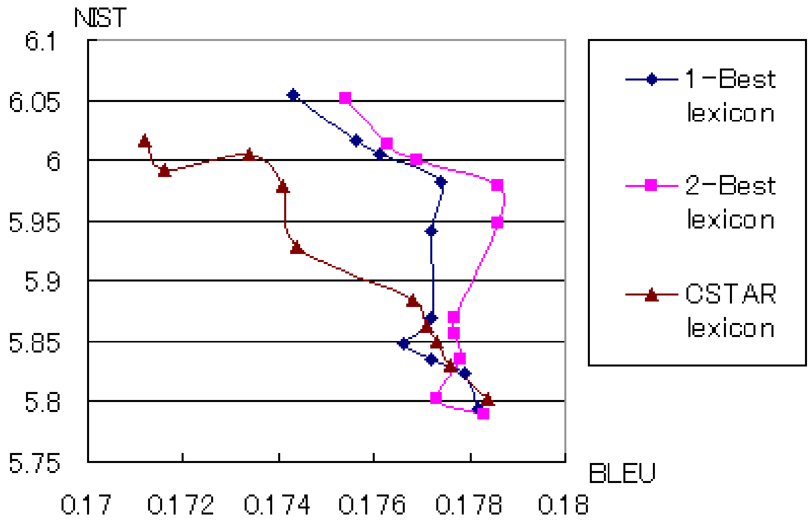

Figure 3: Variation of translation quality in terms of objective measures BLEU and NIST, when the unknown word list is translated decreasingly by number of attempted analogies, and when unknown word translation is performed using a 1-Best lexicon, a 2-Best lexicon, or a larger CSTAR lexicon.

Table 4: Evaluation results.

	BLEU	NIST
Baseline (with UWs)	0.1638	5.8054
Baseline (without UWs)	0.1790	5.7627
1-Best lexicon	0.1747	6.0556
2-Best lexicon	0.1754	6.0507
CSTAR lexicon	0.1712	6.0161
Sorted by attempted analogies	0.1795	5.9193

George Doddington. 2002. Automatic evaluation of machine translation quality using N-gram cooccurrence statistics. In *Proceedings of HLT 2002*, pages 138–145, San Diego.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL'03*, pages 48–54, Morristown. Association for Computational Linguistics.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, volume 3265, pages 115–124, Baltimore.

Yves Lepage and Etienne Denoual. 2006. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19, Numbers 3-4:251–282.

Yves Lepage. 1998. Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume I, pages 728–735, Montréal, August.

Masaaki Nagata. 1999. A part of speech estimation method for japanese unknown words using a statistical model of morphology and context. In *in Proceedings of ACL 1999*, pages 277–284, University of Maryland.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia.

Michael Paul. 2006. Overview of the iwslt06 evaluation campaign. In *Proceedings of IWSLT06*, pages 1–15, Kyoto.

R. Mahesh K. Sinha. 2005. Interpreting unknown words in machine translation from hindi to english. In *Computational Intelligence*, pages 278–282.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, Denver.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The unknown word problem: a morphological analysis of japanese using maximum entropy aided by a dictionary. In *Proceedings of EMNLP 2001*, pages 91–99, Pittsburgh.