Paraphrasing Training Data for Statistical Machine Translation

ERIC NICHOLS[†], FRANCIS BOND[††], D. SCOTT APPLING[†††] and YUJI MATSUMOTO[†]

Large amounts of data are essential for training statistical machine translation systems. In this paper we show how training data can be expanded by paraphrasing one side of a parallel corpus. The new data is made by parsing then generating using an open-source, precise HPSG-based grammar. This gives sentences with the same meaning, but with minor variations in lexical choice and word order. In experiments paraphrasing the English in the Tanaka Corpus, a freely-available Japanese-English parallel corpus, we show consistent, statistically-significant gains on training data sets ranging from 10,000 to 147,000 sentence pairs in size as evaluated by the BLEU and METEOR automatic evaluation metrics.

Key Words: Natural Language Processing, Machine Translation, Paraphrasing, HPSG

1   Introduction

Data-driven machine translation systems such as EBMT and SMT learn how to translate by analyzing aligned bilingual corpora. In general, the more data available the higher the quality of the translation. Unfortunately, there are limits to how much bilingual data exists. In this paper, we propose a method for increasing the amount of parallel text available for training by using a precise, wide-coverage grammar to paraphrase the text in one language.

The novelty in this work is that we are using a hand-crafted grammar to produce the paraphrases, thus adding a completely new source of knowledge to this system. The paraphrases are both meaning-preserving and grammatical, and thus are quite restricted. Possible changes include: changes in word order (*Kim sometimes goes ≡ Kim goes sometimes*), lexical substitution (*everyone ≡ everybody*), contractions (*going to ≡ gonna*) and a limited number of corrections (*the the → the*). We give an example of paraphrasing in (1). The grammar treats all of these sentences as semantically equivalent.

(1) この こと から 、 会社 に は 事故 の 責任 が 無い こと に なる 。

    a. It follows from this that the company is not responsible for the accident.   (= original)

    b. It follows that the company isn't responsible for the accident from this.

    c. It follows that the company is not responsible for the accident from this.

    d. That the company isn't responsible for the accident follows from this.

[†]Nara Institute of Science and Technology, Japan
[††]Nanyang Technological University, Singapore
[†††]Georgia Institute of Technology, United States

We next introduce some related work, then the resources we use in this paper. This is followed by a description of the method and the evaluation. Finally, we discuss the results and our future research plans.

2   Related Work

Approaches to applying paraphrasing in MT can be roughly classified into (1) paraphrasing to expand a machine translation system's coverage, (2) paraphrasing to increase the amount of training or development data, and (3) paraphrasing to increase the similarity between the source and target languages.

In this section, we discuss representative works in each group and compare them with our proposed approach. We limit discussion to *applications* of paraphrasing to machine translation, excluding general discussion of methods of acquiring paraphrases.

2.1   Paraphrasing to Expand Translation Coverage

Callison-Burch et al. (2006) use paraphrases to increase the coverage of unknown source language words in an SMT system. They automatically acquire source language paraphrases from a parallel corpus by using target language phrases as pivots. An example of this approach is given below. By using the German phrase *unter kontrolle* as a pivot, the English phrase *under control* can be paraphrased as *in check*.

(2)   what is more, the relevant cost dynamic is completely <u>under control</u>

     im übrigen ist die diesbezügliche kostenentwicklung völlig <u>unter kontrolle</u>

(3)   wir sind es den steuerzahlern die kosten schuldig <u>unter kontrolle</u> zu haben

     we owe it to the taxpayers to keep the costs <u>in check</u>

New translations are constructed by identifying source language unknown words, finding paraphrases of the unknown words in the system's phrase table, and adding new translation pairs consisting of the unknown word and the translation of its paraphrase. New entries to the phrase table are given the original translation probabilities multiplied by the probability of the source language paraphrases used in their construction. Callison-Burch et al. (2006) showed improvements for sparse datasets for Spanish→English and French→English systems constructed on the Europarl Corpus.

Marton et al. (2009) also use paraphrases to expand an SMT system's phrase table, but they use semantic similarity distribution measures to acquire source language paraphrases from monolingual corpora. They evaluate on Chinese→English and Spanish→English translation tasks, also improving systems trained on sparse datasets, but their approach degrades system performance when trained on 80k or more of data.

Guzmán Herrera and Garrido Luna (2007) take a similar approach and learn new translations for Spanish→English from multi-lingual corpora by using French as a pivot. New translations are added to an existing SMT system's phrase table with probabilities estimated by taking a weighted sum of the combination of paraphrases that produced the new translation. They investigate different weights for the composite paraphrases but do not present evaluations against a baseline system.

The work by Callison-Burch et al. (2006), Marton et al. (2009), and Guzmán Herrera and Garrido Luna (2007) focus on integrating paraphrases acquired from multi- and mono- lingual corpora into an existing SMT system's phrase table with the primary goal of reducing the number of unknown words the system encounters. In order to integrate external paraphrases into an existing phrase table, a measure of translation probability is necessary, and so they develop a series of heuristics to score the artificial alignments produced by pairing paraphrases of a source phrase found in a parallel corpus with the original phrase's alignment in the phrase table.

In contrast, our system produces paraphrases of the sentences in a parallel corpus, and does not alter the phrase table creation process. Rather than learning paraphrases directly from corpora, as in the case of the aforementioned works, our paraphrases are produced from external lexico-syntactic knowledge in the form of an HPSG grammar. English sentences are parsed into a semantic representation that normalizes word order, spelling, and small number of lexical items. Paraphrases are produced from the semantic representation using the same grammar and parser, with paraphrases ranked by a maximum entropy generation model trained on an HPSG treebank. Unlike methods that learn paraphrases directly from corpora, our HPSG paraphrases are limited to grammatical English, eliminating the problem of noisy data.

2.2   Paraphrasing to Increase Translation Data

Nakov (2008) used paraphrases to increase training and parameter tuning data for SMT system. He produced sentence-level paraphrases by using a small set of rules to identify and transform noun phrases in the parallel corpus (e.g., NP_1 of $NP_2 \equiv NP_2$ $'s$ NP_1).

Some examples from Nakov (2008) are:

(4)   of members of the Irish parliament

of Irish parliament members

of Irish parliament's members

(5)   action at community level

community level action

These transformations were structural, not lexical, in nature and limited in scope. Nakov (2008) found that noun phrase-based paraphrases were most effective when applied to training data, achieving a BLEU score gain of about 1 point for limited corpus sizes.

Paraphrases have also been used to enrich the data used for parameter tuning in SMT systems. Madnani et al. (2007) obtained English language paraphrases by identifying paraphrases using a pivot language as in Callison-Burch et al. (2006) and produced sentence-level English paraphrases by training an English→English hierarchical SMT system (Chiang). Experiments showed that paraphrasing the tuning data used for MERT in a Chinese→English hierarchical SMT system performed competitively with increasing human references. Paraphrasing data for parameter tuning is a promising approach, however, evaluating our paraphrasing method in tuning remains future work.

2.3   Paraphrasing to Increase Linguistic Similarity

Another use of paraphrasing is to increase the similarity between source and target languages in order to facilitate translation. The approaches discussed here can be classified into methods that try to simplify the source language vocabulary and those that reorder the source language into a form closer to the word order of the target language.

Simplifying Source Language Vocabulary

One of the earliest applications of paraphrasing to simplify translation input is shown by the rule-based Japanese→English MT system, ALT J/E. Shirai et al. (1993) simplified untranslatable Japanese input into a "pseudo-source language" that, while ungrammatical, was easier for ALT J/E to parse and translate. Yamamoto (2001) adopted a similar approach with his "Sandglass Paradigm" – normalizing input to a rule-based MT system before expanding it again during the translation phase. Watanabe et al. (2002) also used paraphrases to normalize source language text system by detecting paraphrases in a parallel corpus and replacing them with the most commonly occurring variant. Paraphrases were automatically detected with a dynamic programming algorithm, and the normalized data was used to train an SMT system.

Our approach is similar in spirit to these normalization efforts, however, instead of using simple heuristics or identifying paraphrases in a corpus, we apply an external source of knowledge: a rich, lexical grammar. In addition, instead of directly transforming system input, our approach uses paraphrases to enrich the training data, making it more robust by providing instances of lexical and syntactic variants.

Reordering Source Language Text

Overcoming differences in word order is a challenge for translating highly divergent language pairs like Japanese-English or German-English. Recently there has been much work on improving SMT by reordering the source language to closer resemble the word order of the target language.

Nießen and Ney (2001) identify differences in question order and long-distance verbal prefix scrambling as phenomena that cause difficulties for German↔English statistical machine translation and used shallow patterns to reorder "harmonize word order" between the German and English. Collins et al. (2005) made use of parses of source sentences to develop a reordering heuristic as well. Komachi et al. (2006) proposed a reordering model that took into account predicate-argument structure in Japanese and followed a heuristic for reordering sentences in the training data as a preprocessing step. The reordering produces sentences that are not grammatical Japanese, however, they are closer in word order to English, facilitating the SMT alignment process. Katz-Brown and Collins (2008) found that for Japanese→English phrasal SMT a naïve reversal of Japanese source language word order outperformed a dependency-based reordering model. Xu and Seneff (2008) use a rule-based parser to parse English and then generate *Zhonglish*[1]: English reordered to resemble Chinese, with some Chinese function words added. The result is then translated using an SMT system.

Our approach also produces variants in word order, however, they are not artificial reorderings to reduce

_____

[1]A term coined by the paper's authors

word order differences. Rather, these variants are all valid English as defined by the English HPSG grammar. We make the SMT system's training data more robust and representative of English by providing paraphrases that encapsulate the possible positions of adjuncts, such as adverbial and preposition phrases; relative clauses; and other linguistic phenomena in English with variable word order.

3   Resources

In this section we describe the major resources used. For the SMT system we used the open-source Moses system. For paraphrasing we used the open-source English Resource Grammar. We evaluated on the Tanaka Corpus. We chose the Tanaka corpus primarily because of its unencumbered availability (it is in the public domain), making our results easy to reproduce. In the spirit of open science, we have made the paraphrased Tanaka Corpus data as well as the scripts and Moses settings files necessary to reproduce our experiments available online at `http://www3.ntu.edu.sg/home/fcbond/data/`. A summary of all tools used is given in Table 1.

3.1   Moses

Moses (Koehn et al. 2007) is in the words of its creators "a factored phrase-based beam-search decoder for machine translation." It is distributed as open-source software with a collection of utilities that make it easy for users to construct their own SMT system when used with tools for constructing word alignments and language models. For word alignments we used the `giza-pp` branch of GIZA++ (Och and Ney 2003). To construct language models, we used the SRILM Toolkit (Stolcke 2002).

3.2   The English Resource Grammar

The LinGO English Resource Grammar (ERG; (Flickinger 2000)) is a broad-coverage, linguistically precise HPSG-based grammar of English that has been under development at the Center for the Study of Language and Information (CSLI) at Stanford University since 1993. The ERG was originally developed within the Verb*mobil* machine translation effort, but over the past few years has been ported to additional domains and significantly extended. The grammar includes a hand-built lexicon of around 43,000 lexemes. We are using the development release `LinGO (Apr-08)`. Parsing was done with the efficient, unification-based chart parser, PET (Callmeier 2002), and generation with the Linguistic Knowledge Base (Copestake 2002). The ERG and associated parsers and generators are freely available from the Deep Linguistic Processing with HPSG Initiative[2].

For the most part, we use the default settings and the language models trained in the LOGON project both for parsing and generation (Velldal and Oepen 2006). However, we set the root condition, which controls which sentences are treated as grammatical, to be **robust** for parsing and **strict** for generation. This means that robust rules (e.g. a rule that allows verbs to not agree in number with their subject) will apply in parsing but not in generation. The grammar will thus parse *The dog bark* or *The dog barks* but

---

[2]DELPH-IN: `http://www.delph-in.net/`

Tool	Description	Version	Web Page
BLEU Kit	BLEU scores and statistical significance testing	1.03	www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/
ERG	English paraphrasing	LinGo (Apr-08)	www.delph-in.net/erg/
GIZA++	Word-level alignments via IBM models	1.0.3	code.google.com/p/giza-pp/
LKB	HPSG parser/generator	2008/04/13 14:10:44	www.delph-in.net/lkb/
METEOR	MT Evaluation using stemming and synonymy	1.0	www.cs.cmu.edu/~alavie/METEOR/
MeCab	Japanese tokenization	0.97	mecab.sourceforge.net/
Moses	SMT phrasal translation extraction, decoding	20090831svn	statmt.org/moses
NAIST Jdic	Japanese part-of-speech dictionary	0.6.1-20090630	sourceforge.jp/projects/naist-jdic/
PET	Unification-based chart parser	v0.99.14svn	www.delph-in.net/pet/
SRILM	N-gram language models	1.5.9	www.speech.sri.com/projects/srilm/
TreeTagger	English tokenization	3.2	www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Table 1   Tools used for paraphrasing and translation.

	10k	25k	50k	75k	100k	125k	147k
d.0/f.0	35.65	44.47	47.80	50.13	51.55	53.10	53.67
d.2/f.2	37.58 (+1.93)	45.10 (+0.63)	48.92 (+1.12)	50.72 (+0.59)	52.12 (+0.57)	52.85 (-0.25)	53.95 (+0.28)
d.4	38.00 (+2.35)	44.69 (+0.22)	48.34 (+0.54)	49.98 (-0.15)	52.19 (+0.64)	53.26 (-0.16)	53.76 (+0.09)
f.4	36.91 (+1.26)	43.56 (-0.91)	47.58 (-0.22)	50.04 (-0.09)	52.22 (+0.67)	52.96 (-0.14)	53.22 (-0.45)
d.6	37.56 (+1.91)	44.43 (-0.04)	48.22 (+0.42)	50.00 (-0.13)	51.91 (+0.36)	53.16 (+0.06)	53.40 (-0.27)
f.6	37.28 (+1.63)	44.62 (+0.15)	48.55 (+0.75)	50.87 (+0.74)	51.05 (-0.50)	53.28 (+0.18)	53.55 (-0.12)
d.8	37.32 (+1.67)	44.05 (-0.42)	48.94 (+1.14)	50.32 (+0.19)	51.74 (+0.19)	52.65 (-0.45)	53.86 (+0.19)
f.8	37.17 (+1.52)	44.66 (+0.19)	48.15 (+0.35)	50.69 (+0.56)	51.60 (+0.05)	53.05 (-0.05)	53.24 (-0.43)
d.10	37.59 (+1.94)	44.05 (-0.42)	48.65 (+0.85)	50.06 (-0.07)	51.90 (+0.35)	52.94 (-0.16)	53.82 (+0.15)
f.10	37.73 (+2.08)	44.46 (-0.01)	47.65 (-0.15)	49.76 (-0.37)	51.81 (+0.26)	53.23 (-0.13)	53.77 (+0.10)

Table 2   Japanese→English METEOR scores for training data size vs. paraphrases.

only generate *The dog barks*.

$\langle h_1,$

    h_3:person (ARG0 x_4\{PERS *3*, NUM *sg*\}),

    h_5:every_q (ARG0 x_4, RSTR h_6, BODY h_7),

    h_8:_often_a_1 (ARG0 e_9\{TENSE *untensed*\}, ARG1 e_2\{TENSE *pres*\}),

    h_8:_go_v_1 (ARG0 e_2, ARG1 x_4),

    h_8:_to_p (ARG0 e_{10}\{TENSE *untensed*\}, ARG1 e_2, ARG2 x_{11}\{PERS *3*, NUM *pl*, IND *+*\}),

    h_{12}:_the_q (ARG0 x_{11}, RSTR h_{14}, BODY h_{13}),

    h_{15}:_movie_n_of (ARG0 x_{11}, ARG1 i_{16}\{SF *prop*\})

$\{ h_6 =_q h_3, h_{14} =_q h_{15} \} \rangle$

Fig. 1   Semantic Representation of "*Everybody often goes to the the movies.*"

3.3   The Tanaka Corpus

The Tanaka corpus is an open corpus of Japanese-English sentence pairs compiled by Professor Ya-suhito Tanaka at Hyogo University and his students (Tanaka 2001) and released into the public domain. Professor Tanaka's students were given the task of collecting 300 sentence pairs each. After several years, 212,000 sentence pairs had been collected. The sentences were created by the students, often derived from textbooks, e.g. books used by Japanese students of English. Some are lines of songs, others are from popular books and Biblical passages. The original collection contained large numbers of errors, both in the Japanese and English. These are being corrected and added to by volunteers as part of ongoing activity to provide example sentences for the Japanese-Multilingual dictionary JMDict (Breen 2003). Recently, translations into other languages, most notably French, have been added by the TATOEBA project.[3] We give a typical example sentence in (6).

(6)   あの 木 の 枝 に 数 羽 の 鳥 が とまっ て いる 。

"Some birds are sitting on the branch of that tree." (en)

"Des oiseaux se reposent sur la branche de cet arbre." (fr)

The version (2007-04-05) we use has 147,190 sentence pairs in the training split, along with 4,500 sentence pairs reserved for development and 4,500 sentence pairs for testing. After filtering out long sentences (> 40 tokens) as part of the SMT cleaning process, there were 147,007 sentences in the training set. The average number of tokens per sentence is 11.6 for Japanese and 9.1 for English (with the tokenization used in the SMT system).
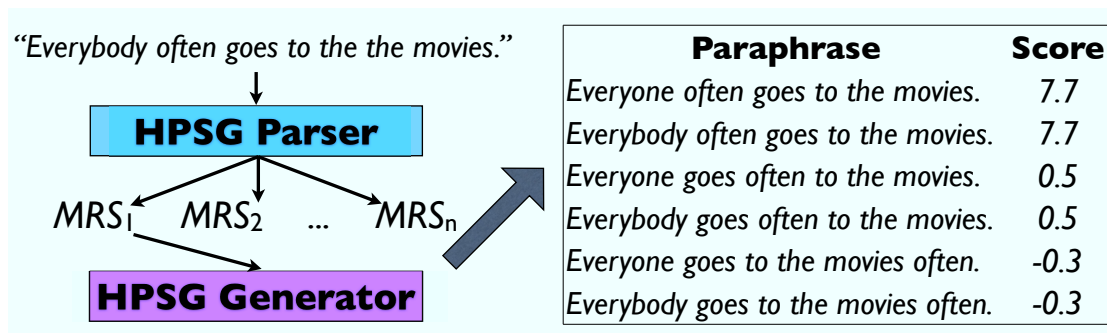
---

[3] http://tatoeba.fr

Fig. 2   Paraphrase process for the sentence "*Everybody often goes to the the movies.*'

4   Method

4.1   Paraphrasing

We paraphrase by parsing a sentence to an abstract semantic representation using the English Resource Grammar then generating from the resultant semantic representation using the same grammar. The semantic representation used is Minimal Recursion Semantics (MRS: (Copestake et al. 2005)). We give an example of the paraphrasing process in Figure 2 that shows three kinds of paraphrasing. The input sentence is "*Everybody often goes to the the movies.*" It is paraphrased to the MRS shown in Figure 1. From that, six sentences are generated. The paraphrased sentences show three changes. Firstly, the erroneous *the the* is corrected to *the*; secondly, *everybody* is optionally paraphrased as *everyone* and finally the adverb *often* appears in three positions (pre-verb, post-verb, post-verb-phrase). We consider the first two to be lexical paraphrases (changes in words) and the latter syntactic paraphrases. Of course, for most sentences there is a combination of lexical and syntactic paraphrases.

"Score" in Figure 2 gives a maximum entropy based likelihood estimate to each of the paraphrases. Note that the highest ranked paraphrase is not in this case the original sentence. The paraphrase is quite conservative: sentence initial *often* is not generated, as that is given a different semantics (it is treated as focused). There are no open class paraphrases like *film* $\equiv$ *movie*. Only a handful of closed class words are substituted, typically those that get decomposed semantically, (e.g., *everybody* $\equiv$ *every(x),person(x)*).

We attempted to parse all sentences of the Tanaka Corpus with the ERG and the PET parser. We got one or more well-formed semantic representation for 87.1% of the sentences (the remainder were rejected as ungrammatical). We selected the top ranked representation and attempted to generate from it, this time using the ERG and the LKB generator. We were able to generate one or more realizations for 83.4% of the original sentences. However, many of these gave only one realization and it was identical to the input sentence. Only 53.4% of the sentences had at least one distinct paraphrase; 31.2% had two, 21.2% had three, dropping down to only 1.1% with ten distinct paraphrases.

We show the distribution of paraphrase types over all of the generated paraphrases in Figure 3. Lexical paraphrases are identified by comparing the set of lexical items in the input with those in the output.

If they are different, then there is a lexical paraphrase (Lex $\neq$). Syntactic paraphrases are identified by comparing the parse trees. Almost a quarter of the sentences generated are the same as the input (Lex =, Syn =). Most variations include some syntactic paraphrasing (Syn $\neq$: 42%), purely lexical paraphrasing is relatively uncommon (8%).
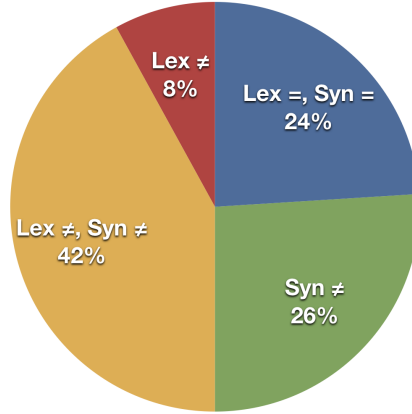

Fig. 3   Types of paraphrases (Lexical and Syntactic)

4.2   Corpus Expansion

Typically when learning translation models, it is assumed that each sentence pair in the parallel corpus is given the same weight. This raises the question of how additional paraphrased data should be weighed. A straight-forward approach would be to simply add each new paraphrase directly to the corpus. However, as sentences can have different numbers of paraphrases, we risk assigning a different weight to each set of original sentence pair and derived paraphrase. A more sophisticated approach would be to assure that each set maintains the same overall weight by distributing it equally among each paraphrase, or by using the paraphrase generation score from Figure 2 to give more likely paraphrases a higher weight. Here, we explore several methods of assigning weights to the paraphrased data by varying the number of times we add each new paraphrase to the parallel corpus.

To make the enhanced training data, we add up to n distinct paraphrases to each unchanged Japanese sentence and original English sentence. We convert all paraphrases to lowercase before checking for uniqueness. If there were m paraphrases, and $n \leq m$ then we just add in the top n ranked paraphrases. If $n > m$ then we produced three test sets:

- (d)istributed: rotate between the original sentence and each paraphrase until the data has been padded out
- (f)irst: after all paraphrases are used, the first (original) sentence is repeated to pad out the data

$$\begin{array}{c|ccccc} d & e_0 & e_1 & e_2 & e_0 & e_1 \\ f & e_0 & e_1 & e_2 & e_0 & e_0 \\ v & e_0 & e_1 & e_2 \end{array}$$

Table 3   Paraphrase distributions ($n = 4, m = 2$)

- (v)arying: add just the paraphrases without padding all entries to the same number of sentences

These variations are shown in Table 3. Both (d and f) keep the distribution close to the original corpus. d puts more weight on the paraphrased sentences and f puts more weight on the original sentence. For v the frequency is distorted; some sentences will be repeated many times. For $n \leq 2$, d and f are the same.

5   Evaluation

In this section, we investigate the effects of supplementing training data with paraphrases on the Tanaka Corpus. We construct phrase-based SMT systems using Moses for the English→Japanese and Japanese→English language pairs, and evaluate systems on various training corpus sizes.

5.1   Moses Baseline

We replicated the baseline in the ACL 2007 Second Workshop on Statistical Machine Translation. The baseline is a factorless Moses system with a 5-gram language model. We followed the online tutorial[4] as-is, with the exception that we used external morphological analyzers to tokenize our data instead of using the provided scripts. We used the Tree Tagger (Schmid 1994) for English and MeCab (Kudo et al. 2004) with NAIST Jdic for Japanese. Part-of-speech information was discarded after tokenization.

All data was tokenized, separating punctuation from words and converted to lowercase prior to training and translation. Translations were detokenized and recased prior to evaluation using the helper scripts distributed as part of the baseline system for the ACL 2007 SMT Workshop.

Prior to evaluation we conducted Minimum Error Rate Training on each system using the development data from the target corpus. We used the MERT implementation distributed with Moses. All results reported in this article are post-MERT BLEU scores.

5.2   Data Preparation

In order to measure the effectiveness of our method, we evaluated the Japanese→English and English→Japanese language pairs over the Tanaka Corpus. Because our HPSG parsers perform best on data that is split on the sentence level, wherever possible we split the corpora into the finest possible sentence pairs. We used the following algorithm to split the evaluation data. However, most of the data in the Tanaka Corpus has already been split at the sentence level as part of the JMDict initiative.

- For each sentence pair:

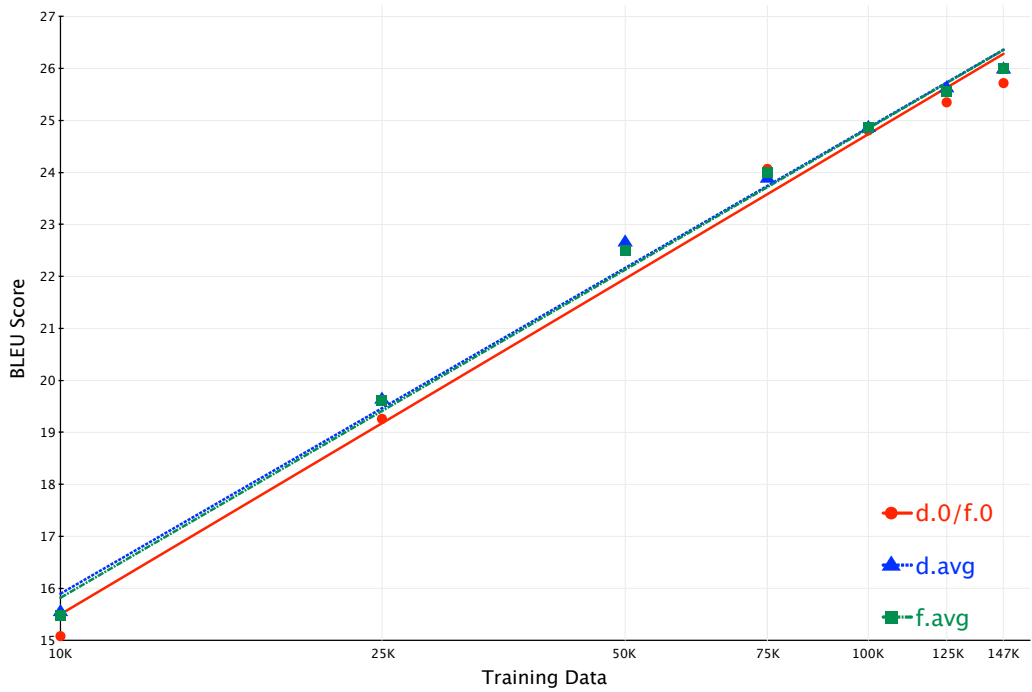[4] www.statmt.org/wmt07/baseline.html

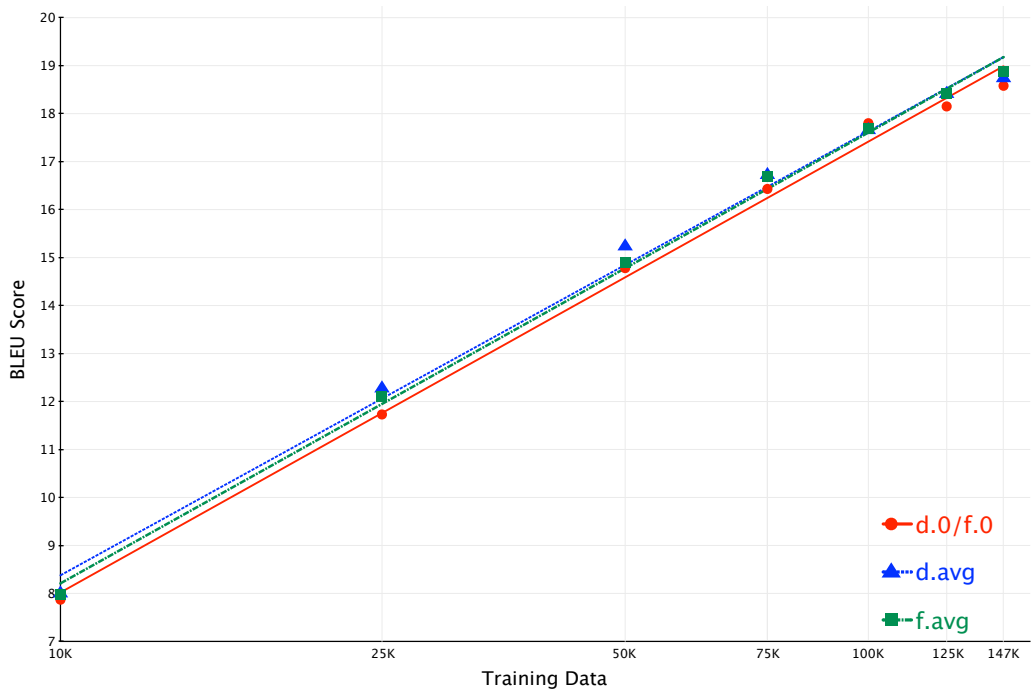Fig. 4   Learning curve for English→Japanese paraphrase distribution averages

Fig. 5   Learning curve for Japanese→English paraphrase distribution averages

11

- split each sentence on sentence-final punctuation (.?!)
- rejoin split on common English titles (Mr./Ms./Mrs./Dr.)
- split sentence pairs with same number of source and target sentences into new pairs
- treat sentence pairs with different number of source and target sentences as a single pair

5.3   Results

We evaluated the effects of adding paraphrases to various initial training data sizes using BLEU and METEOR scores. We compared a baseline of no-paraphrases-added ($d.0/f.0$) to systems with progressively larger numbers of new paraphrased sentence pairs added to each training data size. We tested three distributions (d, f and v). v always gave results below the baseline, so we do not report them in more detail. We give several analyses for d and f below.

Learning Curves

We give learning curves in Figures 4 and 5. The average BLEU scores for *distributed* and *first* paraphrase systems are plotted against training corpus sizes (10k, 25k, 50k, 100k, 125k, and a maximum size of 147k). The training data axis is scaled logarithmically. Best fit lines for the baseline ($d.0/f.0$) and each of the paraphrases show that there is a log-linear relationship in training data size and BLEU score. Paraphrasing almost always outperforms the baseline for small data sets (EJ: 10k-25k, JE: 10k-75k) and large data sets (EJ: 100k-147k, JE: 125k-147k). The region in the middle (EJ: 50k-75k, JE: 100k) is anomalous; the paraphrased averages are below the baseline. We suspect this may be caused by these data sizes containing non-representative samples of data or paraphrases.

BLEU Score

BLEU scores were calculated using the `multi-bleu.perl` implementation distributed with Moses. We measured the statistical significance of BLEU score differences with the bootstrap methods outlined in Koehn (2004) using Jun-ya Norimatsu's MIT-Licensed BLEU Kit. BLEU scores for d and f are given in Table 4, results with an improvement of $p <= 0.05$ over the baseline are shown in **bold**, and the best score for each data size is underlined. We observe a maximum gain of **0.67** BLEU points for English→Japanese at (10k, $d.4$) and a maximum gain of **0.63** for Japanese→Englishat (50k, $d.2/f.2$). Gains appear to peak at 8 paraphrases; $d.10$ and $f.10$ rarely achieve higher scores that can be achieved with fewer paraphrases. The large number of statistically significant BLEU score improvements reinforce our observations made from the learning curves that paraphrasing is beneficial for small data sets and large data sets. We also notice a trend that small numbers of heavily weighted paraphrases like $d.4$ are more effective for small data sets, while larger numbers of lightly-weighted paraphrases like $f.8$ are more effective for large data sets.

Meteor Score

METEOR (Banerjee and Lavie 2005) is an advanced MT evaluation metric that uses stemming and

EJ	10k	25k	50k	75k	100k	125k	147k
d.0/f.0	15.08	19.26	22.58	24.07	24.81	25.35	25.72
d.2/f.2	**15.42** (+0.34)	**19.75** (**+0.49**)	22.64 (+0.06)	23.97 (-0.10)	24.97 (+0.16)	25.59 (+0.24)	25.87 (+0.15)
d.4	**15.73** (**+0.65**)	**19.72** (**+0.46**)	22.74 (+0.16)	23.87 (-0.20)	24.95 (+0.14)	25.56 (+0.21)	**26.08** (**+0.36**)
f.4	**15.54** (**+0.46**)	**19.72** (**+0.46**)	22.23 (-0.35)	24.22 (+0.15)	24.83 (+0.02)	**25.65** (**+0.30**)	26.01 (+0.29)
d.6	**15.59** (**+0.51**)	19.58 (+0.32)	22.72 (+0.14)	23.96 (-0.11)	24.80 (-0.01)	25.54 (+0.19)	25.79 (+0.07)
f.6	**15.66** (**+0.58**)	19.36 (+0.10)	22.53 (-0.05)	23.97 (-0.10)	24.96 (+0.15)	25.58 (+0.23)	25.91 (+0.19)
d.8	15.38 (+0.30)	**19.57** (**+0.31**)	22.56 (-0.02)	23.74 (-0.33)	24.70 (-0.11)	**25.83** (**+0.48**)	**26.13** (**+0.41**)
f.8	15.31 (+0.23)	**19.65** (**+0.39**)	22.62 (+0.04)	24.00 (-0.07)	25.13 (+0.32)	25.59 (+0.24)	**26.15** (**+0.43**)
d.10	**15.64** (**+0.56**)	19.57 (+0.31)	22.65 (+0.07)	23.91 (-0.16)	24.91 (+0.10)	25.62 (+0.27)	**26.08** (**+0.36**)
f.10	**15.45** (**+0.37**)	**19.59** (**+0.33**)	22.45 (-0.13)	23.81 (-0.26)	24.62 (-0.19)	25.63 (+0.28)	**26.05** (**+0.33**)
d.avg	15.55 (+0.47)	19.64 (+0.38)	22.66 (+0.08)	23.89 (-0.18)	24.87 (-0.06)	25.63 (+0.28)	25.99 (+0.27)
f.avg	15.48 (+0.40)	19.61 (+0.35)	22.49 (-0.09)	23.99 (-0.08)	24.87 (-0.06)	25.56 (+0.21)	26.00 (+0.28)

JE	10k	25k	50k	75k	100k	125k	147k
d.0/f.0	7.87	11.73	14.78	16.43	17.80	18.15	18.58
d.2/f.2	8.09 (+0.22)	**12.24** (**+0.51**)	**15.41** (**+0.63**)	16.76 (+0.33)	17.46 (-0.34)	18.15 (+0.00)	18.68 (+0.10)
d.4	7.96 (+0.09)	**12.44** (**+0.71**)	**15.23** (**+0.45**)	16.76 (+0.33)	17.64 (-0.16)	**18.60** (**+0.45**)	18.80 (+0.22)
f.4	7.90 (+0.03)	11.98 (+0.25)	15.01 (+0.23)	16.78 (+0.35)	17.64 (-0.16)	**18.52** (**+0.37**)	18.84 (+0.26)
d.6	7.78 (-0.09)	**12.15** (**+0.42**)	14.91 (+0.13)	16.70 (+0.27)	17.66 (-0.14)	**18.68** (**+0.53**)	18.85 (+0.27)
f.6	8.17 (+0.30)	**12.24** (**+0.51**)	14.83 (+0.05)	**16.95** (**+0.52**)	17.75 (-0.05)	18.37 (+0.22)	18.90 (+0.32)
d.8	**8.31** (**+0.44**)	**12.29** (**+0.56**)	**15.30** (**+0.52**)	16.70 (+0.27)	17.96 (+0.16)	18.33 (+0.18)	18.82 (+0.24)
f.8	7.66 (-0.21)	**12.22** (**+0.49**)	14.89 (+0.11)	16.68 (+0.25)	18.06 (+0.26)	**18.57** (**+0.42**)	**19.04** (**+0.46**)
d.10	7.91 (+0.04)	**12.29** (**+0.56**)	**15.36** (**+0.58**)	16.73 (+0.30)	17.61 (-0.19)	18.34 (+0.19)	18.59 (+0.01)
f.10	8.09 (+0.22)	**12.38** (**+0.65**)	15.01 (+0.23)	16.60 (+0.17)	17.50 (-0.30)	**18.53** (**+0.38**)	18.89 (+0.31)
d.avg	8.01 (+0.14)	12.28 (+0.55)	15.24 (+0.46)	16.73 (+0.30)	17.67 (-0.13)	18.42 (+0.27)	18.75 (+0.17)
f.avg	7.98 (+0.11)	12.11 (+0.38)	14.90 (+0.12)	16.69 (+0.26)	17.68 (-0.12)	18.43 (+0.28)	18.87 (+0.29)

Table 4   English→Japanese (top) and Japanese→English (bottom) BLEU scores for training data size vs. paraphrases.

WordNet synonym matching to relax constraints for English n-gram matches to achieve higher levels of correlation to human judgement than possible with simpler metrics like BLEU and NIST. We calculated all METEOR scores using version 1.0 with the following options: stemming, WordNet stemming, WordNet synonym matching, and "normalization" – stripping of punctuation and conversion to lower case. METEOR score for Japanese→English for d and f are given in Table 2. The METEOR scores do not show as consistent gains as BLEU scores do, but the 10k data set shows great improvements for every paraphrase size. We also note a correlation between statistically significant BLEU score gains and METEOR score improvements; 14/20 paraphrase systems with statistically significant BLEU score gains have increases in METEOR scores.

6   Discussion

Overall, we show significant, consistent improvements on the Tanaka Corpus. Paraphrased SMT systems show statistically significant improvements over the baseline for the majority of the data sizes tested. Furthermore, we observe a log-linear relationship between the size of the system's training data and the BLEU score, with best-fit lines demonstrating the superiority of the paraphrased system over the baseline.

Table 5 shows some examples of how translation output changes with the addition of various amounts of paraphrasing data for the Japanese→English language pair. The example translations contain difficult-to-learn phrasal translations, such as *raining on and off* and *to the point*. As is to be expected from the BLEU scores, the system $f.8$ often gives the best translation. We theorize that the additional data provided by our paraphrases results in better phrasal alignments, which, in turn, improves lexical selection and allows the language model to produce more natural-sounding translations.

Compared to Callison-Burch et al. (2006), Madnani et al. (2007), or Nakov (2008) we are very conservative in our paraphrasing, and this is probably why we get a slightly lower improvement in quality. We could do more extravagant paraphrasing, but would have to retrain our HPSG parser's generation model to effectively rank the new lexical paraphrases. At the moment, it expects fully specified input MRSes. If we were going to allow variation in noun phrase structure or open class lexical variation, then the task could be re-framed as translating between English sentence, and we could build an English→English semantic transfer system to produce richer paraphrases. An example of how to do this (for bilingual transfer of Norwegian→English) is given in Oepen et al. (2007).

Our syntactic reordering is not aimed at matching the target language like Komachi et al. (2006), Xu and Seneff (2008), or Katz-Brown and Collins (2008). We correspondingly get a smaller improvement. On the other hand, because our English paraphrasing method does not depend on a parallel corpus, we expect to get a similar improvement even for different language pairs. Also, our improvement is still there after MERT, whereas the improvement of Komachi et al. (2006) did not make it through the optimization.

We have seen similar increases in SMT system performance for Japanese and English data on the Basic Travel Expression Corpus that is used in the International Workshop for Spoken Machine Translation's translation task. We reported these results in Bond et al. (2008b). Unfortunately, data usage restrictions prevent us from reproducing the results here.

Data Source	Translation
src	猫 は 台所 から 魚 を 持ち出し た 。
ref	The cat made off with a piece of fish from the kitchen.
d.0/f.0	The cat sprang a kitchen fish.
d.2/f.2/d.4/f.4/d.6/d.8/<u>f.8</u>/d.10/f.10	<u>The cat sprang a fish from the kitchen.</u>
f.6	The cat by the fish from the kitchen.
src	一晩 中 雨 が 降っ たり やん だり し て いた 。
ref	It was raining on and off all night long.
d.0/f.0	During the night it has been raining on and off all day.
d.2/f.2/d.4	All night it has been raining on and off.
f.4/<u>f.8</u>/d.10	<u>During the night it has been raining on and off.</u>
d.6	During the night raining off.
f.6	All night, it has been raining on and off.
d.8	All night it has been raining on.
f.10	All night has been raining off and on.
src	彼 の 話 は 短く て 要点 の つい た もの でし た 。
ref	His speech was brief and to the point.
d.0/f.0/d.2/f.2/f.4/f.6/f.10	His speech was brief and to the.
d.4/d.10	His story of the point.
d.6	His story is made of the point.
d.8	His speech was brief and of.
<u>f.8</u>	<u>His story was short and to the point.</u>
src	ビザ の 延長 を お願い し ます 。
ref	Please extend this visa.
d.0/f.0	Do you accept VISA extension of her hierarchical interpersonal relations, please.
d.2/f.2/d.6/d.8	I would like to extend my stay a visa.
d.4/f.6	I'd like to extend my stay visa.
f.4	I'd like to extend my stay with the visa.
<u>f.8</u>	<u>I'd like to stay a visa.</u>
d.10	I'd like to extend my stay the visa.
f.10	The visa I'd like to extend my stay.

Table 5   Example Japanese→English translations from SMT systems trained on 147k of data. The system with the highest BLEU score is <u>underlined</u>.

7   Further Work

There are three areas in which we think the current use of paraphrasing could be improved: (1) increasing the coverage of the grammar (2) adding new classes of paraphrase rules and (3) improving the integration with the SMT process.

To increase the cover of the paraphrasing, we need to improve the handling of unknown words. Currently, the grammar can parse unknown words (which brings the coverage up to almost 95%), but does not pass enough information to the generator to then generate them. We can overcome this with more powerful hybrid parsing, following Adolphs et al. (2008). A more far-ranging increase would be to paraphrase the Japanese side as well. We are working on this using Jacy, an HPSG-based Japanese grammar similar to the ERG (Bond et al. 2008a) and applying the grammatical error tools of Goodman and Bond (2009) to improve the generation coverage of the Japanese grammar.

Before we increase the types of paraphrases we first need to measure which rules (e.g. lexical or syntactic) have the most effect. We then intend to create English rewriting rules using the MRS transfer machinery from the LOGON project, which is already used in an open source Japanese→English MT system (Bond et al. 2005). For example, we can easily write noun phrase rewriting rules of the type used by Nakov (2008). For lexical substitution we will try using WordNet, after first disambiguating the input.

Finally, we would like to enhance Moses (primarily GIZA++) so that input sentences can be weighted. That way, if we have n paraphrases for one sentence and m for another, each can just be entered with a weight of $1/n$ and $1/m$ respectively. If we could do this, we could then experiment with setting a probability based threshold on the number of paraphrases, for example, to select all paraphrases within β of the probability of the original sentence, according to some language model. In this way we could add only "good" paraphrases, and as many as we deem good for each sentence.

8   Conclusion

Large amounts of training data are essential for training statistical machine translation systems. In this paper we showed how training data can be expanded by paraphrasing one side of a parallel corpus. The new data is made by parsing then generating using a precise HPSG-based grammar. This gives sentences with the same meaning, but with minor variations in lexical choice and word order. In experiments paraphrasing the English in the Tanaka Corpus, we showed consistent, statistically-significant gains on training data sets ranging from 10,000 to 147,000 sentence pairs in size as evaluated by the BLEU and METEOR automatic evaluation metrics.

Acknowledgments

Reference

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., and Kiefer, B. (2008). "Some Fine Points of Hybrid Natural Language Parsing." In (ELRA), E. L. R. A. (Ed.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* Marrakech, Morocco.

Banerjee, S. and Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72 Ann Arbor, Michigan. Association for Computational Linguistics.

Bond, F., Kuribayashi, T., and Hashimoto, C. (2008a). "Construction of a Free Japanese Treebank Based on HPSG." In *14th Annual Meeting of the Association for Natural Language Processing*, pp. 241–244 Tokyo. (in Japanese).

Bond, F., Nichols, E., Appling, D. S., and Paul, M. (2008b). "Improving Statistical Machine Translation by Paraphrasing the Training Data." In *Proceedings of IWSLT 2008*, pp. 150–157 Hawaii.

Bond, F., Oepen, S., Siegel, M., Copestake, A., and Flickinger, D. (2005). "Open Source Machine Translation with DELPH-IN." In *Open-Source MT: Workshop at MT Summit X*, pp. 15–22 Phuket.

Breen, J. W. (2003). "Word Usage Examples in an Electronic Dictionary." In *Papillon (Multi-lingual Dictionary) Project Workshop* Sapporo.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). "Improved Statistical Machine Translation Using Paraphrases." In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 17–24.

Callmeier, U. (2002). "Preprocessing and Encoding Techniques in PET." In Oepen, S., Flickinger, D., Tsujii, J., and Uszkoreit, H. (Eds.), *Collaborative Language Engineering*, chap. 6, pp. 127–143. CSLI Publications, Stanford.

Chiang, D. "A Hierarchical Phrase-Based Model for Statistical Machine Translation." In *ACL 2005 Proceedings*.

Collins, M., Koehn, P., and Kǔcerová, I. (2005). "Clause Restructuring for Statistical Machine Translation." In *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 531–540.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). "Minimal Recursion Semantics. An Introduction." *Research on Language and Computation*, **3** (4), pp. 281–332.

Flickinger, D. (2000). "On Building a More Efficient Grammar by Exploiting Types." *Natural Language Engineering*, **6** (1), pp. 15–28. (Special Issue on Efficient Processing with HPSG).

Goodman, M. W. and Bond, F. (2009). "Using Generation for Grammar Analysis and Error Detection." In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pp. 109–112 Singapore.

Guzmán Herrera, F. and Garrido Luna, L. (2007). "Using Translation Paraphrases from Trilingual Corpora to Improve Phrase-Based Statistical Machine Translation: A Preliminary Report." In *Proceedings of the Mexican International Conference on Artificial Intelligence*, pp. 163–172 Los Alamitos, CA,

USA. IEEE Computer Society.

Katz-Brown, J. and Collins, M. (2008). "Syntactic Reordering in Preprocessing for Japanese→English Translation: MIT System Description for NTCIR-7 Patent Translation Task." In *Proceedings of the 7th NTCIR Workshop Meeting* Tokyo, Japan.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In *Proceedings of EMNLP 2004* Barcelona, Spain.

Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., Moran, C., and Birch, A. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 177–180 Prague.

Komachi, M., Matsumoto, Y., and Nagata, M. (2006). "Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure." In *Proceedings of IWSLT 2006*.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). "Applying Conditional Random Fields to Japanese Morphological Analysis." In *EMNLP 2004 Proceedings*, pp. 230–237 Barcelona, Spain. Association for Computational Linguistics.

Madnani, N., Fazil Ayan, N., Resnik, P., and Dorr, B. (2007). "Using Paraphrases for Parameter Tuning in Statistical Machine Translation." In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 120–127 Prague, Czech Republic. Association for Computational Linguistics.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). "Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases." In *EMNLP 2009 Proceedings*, pp. 381–390 Singapore. Association for Computational Linguistics.

Nakov, P. (2008). "Improved Statistical Machine Translation Using Monolingual Paraphrases." In *Proceedings of the European Conference on Artificial Intelligence (ECAI'08)* Patras, Greece.

Nießen, S. and Ney, H. (2001). "Morpho-Syntactic Analysis for Reordering in Statistical Machine Translation." In *Proceedings of MT Summit VIII*, pp. 247–252.

Och, F. J. and Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, **29** (1), pp. 19–51.

Oepen, S., Velldal, E., Løning, J. T., Meurer, P., and Rosen, V. (2007). "Towards Hybrid Quality-Oriented Machine Translation -On Linguistics and Probabilities in MT-." In *TMI 2007 Proceedings* Skövde.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *International Conference on New Methods in Language Processing* Manchester, UK.

Shirai, S., Ikehara, S., and Kawaoka, T. (1993). "Effects of Automatic Rewriting of Source Language within a Japanese to English MT System." In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 226–239 Kyoto, Japan.

Stolcke, A. (2002). "SRILM - An Extensible Language Modeling Toolkit." In *International Conference on Spoken Language Processing*, Vol. 2, pp. 901–904 Denver.

Tanaka, Y. (2001). "Compilation of a Multilingual Parallel Corpus." In *Proceedings of PACLING 2001*, pp. 265–268 Kyushu.

Velldal, E. and Oepen, S. (2006). "Statistical Ranking in Tactical Generation." In *EMNLP 2006 Proceed-*

ings, pp. 517–525 Sydney, Australia. Association for Computational Linguistics.

Watanabe, T., Shimohata, M., and Sumita, E. (2002). "Statistical Machine Translation on Paraphrased Corpora." In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 2074–2081 Las Palmas, Spain.

Xu, Y. and Seneff, S. (2008). "Two-Stage Translation: A Combined Linguistic and Statistical Machine Translation Framework." In *AMTA 2008 Proceedings* Honolulu, Hawaii.

Yamamoto, K. (2001). "Paraphrasing Spoken Japanese for Untangling Bilingual Transfer." In *Proceedings of Natural Language Processing Pacific Rim Symposium 2001*, pp. 203–210.

Eric Nichols: He is currently a Researcher at Nara Institute of Science and Technology, Japan. Eric received a BS in Computer Science and a BA in Japanese from the University of Maryland at College Park, United States. He received an ME and a DrEng in Information Science from NAIST and was a recipient of the Research Student Scholarship from MEXT, Japan. He has done joint research with NTT and NICT. Eric's research interests include machine translation, knowledge acquisition, and information credibility analysis.

Francis Bond: He is currently an Associate Professor at the Division of Linguistics and Multilingual Studies, Nanyang Technological University, Singapore. Francis received a BA in 1988, a BEng (1st) in 1990 and later a PhD in 2001, all at the University of Queensland. He worked on machine translation and natural language understanding at NTT Corp., from 1991 to 2006. From 2006-2009 he worked at NICT, Japan, where his focus was on open source natural language processing. His main research interest is natural language understanding.

D. Scott Appling: He is currently a research scientist at the Georgia Tech Research Institute in Atlanta, Georgia, United States. Scott holds a BS in Computer Science with a Minor in Japanese language and an MS in Computer Science from the Georgia Institute of Technology. His research interests include natural language understanding, machine translation, applied machine learning, and narrative intelligence.

Yuji Matsumoto: He is currently a Professor of Information Science, Nara Institute of Science and Technology. Yuji received his MS and PhD degrees in information science from Kyoto University in 1979 and in 1989. He joined the Machine Inference Section of Electrotechnical Laboratory in 1979. He has then experienced an academic visitor at Imperial College of Science and Technology, a deputy chief of First Laboratory at ICOT, and an associate professor at Kyoto University. His main research interests are natural language understanding and machine learning.