

Improving Statistical Machine Translation by Paraphrasing the Training Data

Francis Bond,¹ Eric Nichols,² Darren Scott Appling,³ Michael Paul¹

¹ National Institute of Information and Communications Technology

² Nara Institute of Science and Technology

³ Georgia Institute of Technology

bond@ieee.org, eric-n@is.naist.jp

darren.scott.appling@gatech.edu, Michael.Paul@nict.go.jp

Abstract

Large amounts of training data are essential for training statistical machine translations systems. In this paper we show how training data can be expanded by paraphrasing one side. The new data is made by parsing then generating using a precise HPSG based grammar, which gives sentences with the same meaning, but minor variations in lexical choice and word order. In experiments with Japanese and English, we showed consistent gains on the Tanaka Corpus with less consistent improvement on the IWSLT 2005 evaluation data.

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 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* \equiv *Kim goes sometimes*), lexical substitution (*everyone* \equiv *everybody*), contractions (*going to* \equiv *gonna*) and a limited number of corrections (*the the* \rightarrow *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。

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

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

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

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

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 how we plan to extend the research.

2. Related Work

Automatic generation and use of paraphrases has been of considerable interest to the MT community in recent years. Because the paraphrase is an alternate representation of the same meaning it can be derived from pre-existing training corpora and automatically aligned with the same target training sentence in a bilingual corpora. Including generated paraphrases as additional training data gives an SMT system the ability to make a richer model and thus positively affecting model quality during evaluation.

There are several areas where paraphrases can be readily introduced into standard phrase-based SMT systems: source and target sides and even during the parameter tuning phase as references. Work has been done to extract paraphrases from bilingual corpora [1] and to extract paraphrase patterns as in [2]. By pivoting on target language phrases, source phrases and potential paraphrases can be found; for extracting patterns the task extends to a generalization of phrases with slots instead

of words. [3] found that it is possible to translate unknown source words by paraphrasing them and then do a translation on the paraphrase. The classic example of the pivot approach from [3] follows:

- (2) what is more, the relevant cost dynamic is completely under control
im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle
- (3) wir sind es den steuerzahlern die kosten schuldig unter kontrolle zu haben
we owe it to the taxpayers to keep the costs in check

By holding the german phrase *unter kontrolle* as a pivot the English phrase *under control* can be paraphrased as *in check*. [2] extend the pivot approach to general patterns by using part of speech as a constraint to slots in their patterns, an example follows:

- (4) all the members of [NNPS_1]
all members of [NNPS_1]

The slot part of speech [NNPS_1] constraints the paraphrase to ensure correct match ups when filling in to make real phrases. Other research involving monolingual corpora, was done in [4] who paraphrased noun phrases, after first parsing sentences to identify the noun phrases. Sentence variants are generated as paraphrases when appropriately structured noun phrases are found. Only six grammar transformation rules were used and there was no lexical paraphrasing per se. The paper also presented a result that paraphrasing entries in the phrase table does not compare to the impact of adding paraphrase directly to the training corpus. As for the BLEU scores, they only ever achieve an increase of about 1 BLEU point and this is on limited corpora sizes.

Some examples from [4] are as follows:

- (5) of members of the Irish parliament
of irish parliament members
of irish parliament's members
- (6) action at community level
community level action

To the extent that paraphrasing techniques are comparable to more implicit methods of language to language manipulation we explore previous research related to reordering models where a form of paraphras-

ing does occur in how translation is done. In [5] 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 pre-processing step. This sort of reordering, while unnatural to native speakers, is still grammatically correct and easier to align to English during model training, it is also a type of paraphrasing. [6] made use of parses of source sentences and then applied a reordering heuristic as well. [7] also discusses the use altering German word order to correspond to English word order there is also some use of annotations on verbs with identifying prefixes to solve the long distance dependency of German verbs types that allow for separation of the prefix from the verb in the sentence structure.

3. Resources

In this section we describe the major resources used. For the SMT system we used the open source Moses system¹. For paraphrasing we used the open source English Resource Grammar. We tested on two Japanese-English corpora, the Tanaka Corpus and the IWSLT corpus. We chose the Tanaka corpus primarily because of its easy availability (it is in the public domain). This will make our results easy to reproduce. We also tested on the IWSLT corpus, as it has been used in several competitions, in order to facilitate comparisons with other systems.

In the spirit of open science, the paraphrased Tanaka Corpus data and our scripts will be put on line at www2.nict.go.jp/x/x161/en/member/bond/data/.

3.1. Moses

Moses [8] is an open-source toolkit for phrase-based statistical machine translation with support for factors. The toolkit is one of the first highly efficient and free SMT decoders and tool kits; it supports building factored statistical models. Factors such as part-of-speech, morphology, and lemmas are applied using translation models and generation models, two features of Moses which in conjunction with the user specified decoding steps help to create stories of how one language might translate best to another.

We used the multi-threaded Giza++ [9] as it fixed a bug in how probabilities are assigned to low frequency events. To construct language models, we used the SRILM

¹We used the 20080711 public release of Moses.

Toolkit [10].

3.2. The English Resource Grammar

The LinGO English Resource Grammar (ERG; [11]) 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 *VerbMobil* 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)`. The ERG and the associated parsers and generators are freely available from the Deep Linguistic Processing with HPSG Initiative (DELPH-IN: www.delph-in.net/).

Generally, we use the default settings and the language models trained in the LOGON project both for parsing and generation [12]. 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 (for example a rule to allow verbs not to 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 only generate *The dog barks*.

3.3. Corpora

We used two corpora, one freely available, and one standard test set.

3.3.1. Tanaka Corpus

The Tanaka corpus is an open corpus of Japanese-English sentence pairs compiled by Professor Yasuhito Tanaka at Hyogo University and his students [13] 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 by volunteers, as part of ongoing activity to provide example sentences

for the Japanese-English dictionary JMDict [14]. Recently, translations in other languages, most notably French, have been added by the TATOEBEA project.² We give a typical example sentence in (7).

- (7) 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).

3.3.2. The IWSLT Corpus

We also tested our system on the **IWSLT** 2005 evaluation corpus [15]. This is a subset of the *Basic Travel Expression Corpus* (BTEC), which contains tourism-related sentences similar to those that are usually found in phrase books for tourists going abroad [16]. Parts of this corpus were already used in previous IWSLT evaluation campaigns [17]. We used the evaluation and development data sets of 2004, although only with the first of the multiple reference translations, for our development corpora and the 500 sentence **IWSLT** 2005 evaluation set, again with only the first of the 16 references, as the evaluation corpus.

The IWSLT corpus has 42,682 sentence pairs. The average number of tokens per sentence is 9.0 for Japanese and 8.0 for English (with the tokenization used in the SMT). The sentences are both shorter and more homogeneous than those in the Tanaka Corpus.

4. Method

4.1. Paraphrasing

We paraphrase by parsing a sentence to an abstract semantic representation using the English Resource Grammar, and then generating from that using the same grammar. The semantic representation used is Minimal Recursion Semantics (MRS; [18]). We give an example in

²www.cyg.utc.fr/tatoeba/

$\langle h_1,$ $h_3:\text{person}(\text{ARG0 } x_4\{\text{PERS } 3, \text{NUM } sg\}),$ $h_5:\text{every_q}(\text{ARG0 } x_4, \text{RSTR } h_6, \text{BODY } h_7),$ $h_8:\text{_often_a_1}(\text{ARG0 } e_9\{\text{TENSE } untensed\}, \text{ARG1 } e_2\{\text{TENSE } pres\}),$ $h_8:\text{_go_v_1}(\text{ARG0 } e_2, \text{ARG1 } x_4),$ $h_8:\text{_to_p}(\text{ARG0 } e_{10}\{\text{TENSE } untensed\}, \text{ARG1 } e_2, \text{ARG2 } x_{11}\{\text{PERS } 3, \text{NUM } pl, \text{IND } +\}),$ $h_{12}:\text{_the_q}(\text{ARG0 } x_{11}, \text{RSTR } h_{14}, \text{BODY } h_{13}),$ $h_{15}:\text{_movie_n_of}(\text{ARG0 } x_{11}, \text{ARG1 } i_{16}\{\text{SF } prop\})$ $\{ h_6 =_q h_3, h_{14} =_q h_{15} \}$	
--	--

Figure 1: Semantic Representation of “Everybody often goes to the the movies.”

Paraphrase	Score
<i>Everyone often goes to the movies.</i>	7.7
<i>Everybody often goes to the movies.</i>	7.7
<i>Everyone goes often to the movies.</i>	0.5
<i>Everybody goes often to the movies.</i>	0.5
<i>Everyone goes to the movies often.</i>	-0.3
<i>Everybody goes to the movies often.</i>	-0.3

Figure 2: Paraphrases of “Everybody often goes to the the movies.”

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* 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).

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, typically those that get decomposed semantically, (like *everybody* \equiv *every person*) are substituted.

For the Tanaka Corpus, 87.1% of the sentences could be parsed and 83.4% paraphrased. However many of these gave only one paraphrase and it was identical to the input sentence. Only 53.4% of 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. The numbers were a few percent higher for the IWSLT corpus. Parsing was done with PET [19] and generation with the LKB [20].

4.2. Corpus Expansion

To make the enhanced training data, we add up to n new sentence pairs, consisting of the unchanged Japanese sentence, the original English sentence and up to n distinct paraphrases. Distinct paraphrases are tested in down-cased form.

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 have been used, the first (original) sentence is repeated to pad out the data
- (v)arying: add just the paraphrases

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		

Table 1: Paraphrase distributions ($n = 4, m = 2$)

These variations are shown in Figure 1. 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 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 present experimental results on two different corpora, evaluating phrase-based SMT systems constructed using Moses for the English→Japanese and Japanese→English language pairs.

Language Pair	Corpus	Paraphrases Added	Bleu	Variance	Delta
EJ	Tanaka Corpus	0	25.96	± 0.71	-
EJ	Tanaka Corpus	d.2	26.10	± 0.74	+0.14
EJ	Tanaka Corpus	d.4	26.25	± 0.71	+0.29
EJ	Tanaka Corpus	d.6	26.63	± 0.72	+0.67
EJ	Tanaka Corpus	d.8	26.16	± 0.71	+0.20
EJ	Tanaka Corpus	f.2	26.10	± 0.77	+0.14
EJ	Tanaka Corpus	f.4	26.28	± 0.73	+0.32
EJ	Tanaka Corpus	f.6	26.13	± 0.68	+0.17
EJ	Tanaka Corpus	f.8	25.83	± 0.65	-0.13
JE	Tanaka Corpus	0	18.75	± 0.82	-
JE	Tanaka Corpus	d.2	19.09	± 0.74	+0.34
JE	Tanaka Corpus	d.4	18.42	± 0.79	-0.33
JE	Tanaka Corpus	d.6	18.71	± 0.83	-0.04
JE	Tanaka Corpus	d.8	18.90	± 0.77	+0.15
JE	Tanaka Corpus	f.2	19.09	± 0.82	+0.34
JE	Tanaka Corpus	f.4	18.92	± 0.81	+0.17
JE	Tanaka Corpus	f.6	19.02	± 0.80	+0.27
JE	Tanaka Corpus	f.8	19.19	± 0.82	+0.44

Table 2: Results of adding paraphrases to Tanaka Corpus training data

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 tutorial³ 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 [21] for English and MeCab [22] 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 paper are post-mert Bleu scores.

5.1. Data Preparation

In order to measure the effectiveness of our method, we evaluated JE and EJ over two data sets: the Tanaka Corpus and the IWSLT 2005 evaluation 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 IWSLT 2005 evaluation corpus, observing no errors in the development data. Once split, the IWSLT 2005 data consisted of 42,699 training sentences, 1,076 development sentences, and 543 test sentences. 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:
 - split each sentence on sentence-final punctuation (.?!)
 - rejoin split on common English titles (Mr./Ms./Mrs./Dr.)
 - split sentence pairs with same # of src and tgt sentences into new sentence pairs
 - treat sentence pairs with different # of src and tgt sentences as a single sentence pair

³www.statmt.org/wmt07/baseline.html

Language Pair	Corpus	Paraphrases Added	Bleu	Variance	Delta
EJ	IWSLT05	0	35.63	± 2.75	-
EJ	IWSLT05	d.2	35.70	± 2.88	+0.07
EJ	IWSLT05	d.4	35.80	± 3.21	+0.17
EJ	IWSLT05	d.6	34.17	± 2.94	-1.46
EJ	IWSLT05	d.8	35.39	± 2.88	-0.24
EJ	IWSLT05	f.2	35.70	± 2.75	+0.07
EJ	IWSLT05	f.4	35.82	± 2.74	+0.19
EJ	IWSLT05	f.6	35.20	± 2.95	-0.43
EJ	IWSLT05	f.8	35.00	± 3.21	-0.63
JE	IWSLT05	0	23.75	± 2.65	-
JE	IWSLT05	d.2	24.36	± 2.82	+0.61
JE	IWSLT05	d.4	24.21	± 2.58	+0.46
JE	IWSLT05	d.6	24.06	± 2.52	+0.31
JE	IWSLT05	d.8	23.60	± 2.71	-0.15
JE	IWSLT05	f.2	24.36	± 2.76	+0.61
JE	IWSLT05	f.4	24.34	± 2.70	+0.59
JE	IWSLT05	f.6	23.78	± 2.51	+0.03
JE	IWSLT05	f.8	23.24	± 2.79	-0.51

Table 3: Results of adding paraphrases to **IWSLT** 2005 training data

5.2. Results

We compared a baseline of no paraphrases added ($d.0$) to systems with progressively larger numbers of new paraphrased sentence pairs added to the training data. We tested three distributions (d , f and v). v always gave results below the baseline, so we do not report them in more detail.

The results for d and f are summarized in Tables 2 and 3 with 2, 4, 6 and 8 paraphrases. All deltas and significance results are calculated against the baseline of no paraphrases (0).

We calculated Bleu score variance and measured statistical significance with the bootstrap methods outlined in [23] using Jun-ya Norimatsu’s MIT-Licensed Bleu Kit.⁴ Variance scores are reported with $p = 0.05$ in Tables 2 and 3. In Tables 2 and 3 results with an improvement of $p < 0.10$ over the baseline are shown in bold.

6. Discussion

The results for $En \rightarrow Ja$ show gains of up to 0.67 Bleu points on the Tanaka Corpus and 0.19 on the **IWSLT**

⁴www.mibel.cs.tsukuba.ac.jp/~norimatsu/bleu_kit/

2005 evaluation data. The results for $Ja \rightarrow En$ show gains of 0.44 on the Tanaka Corpus and 0.61 on the **IWSLT** 2005 evaluation data.

There is a statistically significant improvement for each language pair and paraphrase distribution method on the Tanaka Corpus, but none on the **IWSLT** 2005 evaluation data. We hypothesize this is due to the difference in variance in the two corpora: over ± 2.51 (**IWSLT**) vs. less than ± 0.83 (Tanaka). Changes in Bleu score that would be significant in the Tanaka Corpus, like +0.61 for JE d.2/f.2 are lost in this variance.

The saturation point for EJ tends towards 4 paraphrases, but peaks slightly later at d.6 for the Tanaka Corpus. JE is somewhat inconsistent: for **IWSLT** it peaks at d.2/f.2, but in the Tanaka Corpus it peaks initially at d.2/f.2, before dropping off and then surging to a maximum at f.8⁵.

Overall, we show significant, consistent improvements on the Tanaka Corpus, with less consistent but overall positive results on the **IWSLT** 2005 evaluation data. Our explanation for this difference is that the Tanaka Corpus is a more difficult and heterogeneous

⁵These results contrast with earlier experiments using the 20080525svn snapshot of Moses. There was a clearer saturation point at 4 added paraphrases, although with lower overall Bleu scores and less statistically significant results.

data set, as reflected by its overall Bleu scores, enhancing the effects of new sources of data. Another explanation may be that the IWSLT data already contains many English paraphrases (there are over 2,500 Japanese sentences with more than one English translation and over a 1,000 English sentences with multiple Japanese translations). In contrast for the Tanaka Corpus all the Japanese sentences are unique, although there are over 9,000 English sentences with multiple Japanese paraphrases.

Compared to [3] or [4] 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 re-train the generation model. At the moment, it expects fully specified input MRSs, if we were going to allow variation in, for example, noun phrase structure or open class lexical variation, then we should treat it as a monolingual translation problem, and also train a transfer (paraphrase) model. An example of how to do this (for bilingual transfer (Norwegian-English)) is given in [24].

Our syntactic reordering is not aimed at matching the target language like [5]. We correspondingly get a slighter improvement, but can hope to get a similar improvement even for different language pairs. Also, our improvement is still there after MERT training, whereas theirs did not survive the optimization.

7. Further Work

There are three areas in which we think the current use of paraphrasing could be improved: (1) we can work on increasing the cover of the grammar (2) we could add new classes of paraphrase rules and (3) we could improve 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 are currently working on a fix for this. A more far ranging increase would be to paraphrase the Japanese side as well. We are also working on this, using Jacy, an HPSG-based grammar Japanese of the same type as the ERG [25].

To increase the types of paraphrases we first need to measure which rules (e.g. lexical variation vs. reordering have the most effect). We then intend to make use of the MRS transfer machinery from the LOGON project, which we already use in an open source Japanese-English

MT system [26]. We can easily write noun phrase rewriting rules of the type used by [4]. 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. Conclusions

Large amounts of training data are essential for training statistical machine translations systems. In this paper we show how training data can be expanded by paraphrasing one side. The new data was made by parsing and then generating using a precise HPSG based grammar, which gives sentences with the same meaning, but minor variations in lexical choice and word order. In experiments with Japanese and English, we showed consistent gains on the Tanaka Corpus with less consistent improvement on the IWSLT 2005 evaluation data.

Acknowledgments

Thanks to Kiyotaka Uchimoto, Kentaro Torisawa, Dan Flickinger and the NICT Language Infrastructure group for their valuable discussion.

9. References

- [1] C. Bannard and C. Callison-Burch, “Paraphrasing with bilingual parallel corpora,” in *Association for Computational Linguistics*, 2005, pp. 597–604.
- [2] S. Zhao, H. Wang, T. Liu, and S. Li, “Pivot approach for extracting paraphrase patterns from bilingual corpora,” in *Proceedings of ACL: HLT*, 2008, pp. 780–788.
- [3] C. Callison-Burch, P. Koehn, and M. Osborne, “Improved statistical machine translation using paraphrases,” in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006, pp. 17–24.
- [4] P. Nakov, “Improved statistical machine translation using monolingual paraphrases,” in *Proceedings of the European Conference on Artificial Intelligence (ECAI’08)*, Patras, Greece, 2008.

- [5] M. Komachi, Y. Matsumoto, and M. Nagata, "Phrase reordering for statistical machine translation based on predicate-argument structure," in *Proceedings of IWSLT 2006*, 2006.
- [6] M. Collins, P. Koehn, and I. Kúcerová, "Clause restructuring for statistical machine translation," in *Proceedings of the 43rd Annual Meeting of the ACL*, 2005, pp. 531–540.
- [7] S. Nießen and H. Ney, "Morpho-syntactic analysis for reordering in statistical machine translation," in *Proceedings of MT Summit VIII*, 2001, pp. 247–252.
- [8] P. Koehn, W. Shen, M. Federico, N. Bertoldi, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, O. Bojar, R. Zens, A. Constantin, E. Herbst, C. Moran, and A. Birch, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the ACL 2007 Interactive Presentation Sessions*, Prague, 2007. [Online]. Available: <http://www.statmt.org/moses/>
- [9] Q. Gao, "Multi-threaded giza," 2008, (Source Code Released). [Online]. Available: <http://www.cs.cmu.edu/~qing/>
- [10] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Intl. Conf. Spoken Language Processing*, Denver, 2002, pp. 901–904.
- [11] D. Flickinger, "On building a more efficient grammar by exploiting types," *Natural Language Engineering*, vol. 6, no. 1, pp. 15–28, 2000, (Special Issue on Efficient Processing with HPSG).
- [12] E. Velldal and S. Oepen, "Statistical ranking in tactical generation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 517–525. [Online]. Available: <http://www.aclweb.org/anthology/W/W06/W06-1661>
- [13] Y. Tanaka, "Compilation of a multilingual parallel corpus," in *Proceedings of PACLING 2001*, Kyushu, 2001, pp. 265–268, (<http://www.colips.org/afnlp/archives/pacling2001/pdf/tanaka.pdf>).
- [14] J. W. Breen, "Word usage examples in an electronic dictionary," in *Papillon (Multi-lingual Dictionary) Project Workshop*, Sapporo, 2003, (<http://www.csse.monash.edu.au/~jwb/papillon/dicexamples.html>).
- [15] M. Eck and C. Hori, "Overview of the IWSLT 2005 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Pittsburgh, USA, 2005, pp. 11–32.
- [16] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. of the EUROSPEECH03*, Geneva, Switzerland, 2003, pp. 381–384.
- [17] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the IWSLT04 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 1–12.
- [18] A. Copestake, D. Flickinger, C. Pollard, and I. A. Sag, "Minimal Recursion Semantics. An introduction," *Research on Language and Computation*, vol. 3, no. 4, pp. 281–332, 2005.
- [19] U. Callmeier, "Preprocessing and encoding techniques in PET," in *Collaborative Language Engineering*, S. Oepen, D. Flickinger, J. Tsujii, and H. Uszkoreit, Eds. Stanford: CSLI Publications, 2002, ch. 6, pp. 127–143.
- [20] A. Copestake, *Implementing Typed Feature Structure Grammars*. CSLI Publications, 2002.
- [21] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*. Manchester, UK: unknown, 1994. [Online]. Available: citeseer.ist.psu.edu/schmid94probabilistic.html
- [22] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," in *Proceedings of EMNLP 2004*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 230–237.
- [23] P. Koehn, "Statistical significance tests for machine translation evaluation," 2004. [Online]. Available: citeseer.ist.psu.edu/koehn04statistical.html
- [24] S. Oepen, E. Velldal, J. T. Løning, P. Meurer, and V. Rosen, "Towards hybrid quality-oriented machine translation. on linguistics and probabilities in MT," in *11th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2007*, 2007, pp. 144–153.
- [25] F. Bond, T. Kuribayashi, and C. Hashimoto, "Construction of a free Japanese treebank based on HPSG," in *14th Annual Meeting of the Association for Natural Language Processing*, Tokyo, 2008, pp. 241–244, (in Japanese).
- [26] F. Bond, S. Oepen, M. Siegel, A. Copestake, and D. Flickinger, "Open source machine translation with DELPH-IN," in *Open-Source Machine Translation: Workshop at MT Summit X*, Phuket, 2005, pp. 15–22.