Monolingual Machine Translation for Paraphrase Generation

Chris QUIRK, **Chris BROCKETT** and **William DOLAN**
Natural Language Processing Group
Microsoft Research
One Microsoft Way
Redmond, WA 90852 USA
{chrisq,chrisbkt,billdol}@microsoft.com

Abstract

We apply statistical machine translation (SMT) tools to generate novel paraphrases of input sentences in the same language. The system is trained on large volumes of sentence pairs automatically extracted from clustered news articles available on the World Wide Web. Alignment Error Rate (AER) is measured to gauge the quality of the resulting corpus. A monotone phrasal decoder generates contextual replacements. Human evaluation shows that this system outperforms baseline paraphrase generation techniques and, in a departure from previous work, offers better coverage and scalability than the current best-of-breed paraphrasing approaches.

1 Introduction

The ability to categorize distinct word sequences as "meaning the same thing" is vital to applications as diverse as search, summarization, dialog, and question answering. Recent research has treated paraphrase acquisition and generation as a machine learning problem (Barzilay & McKeown, 2001; Lin & Pantel, 2002; Shinyama et al, 2002, Barzilay & Lee, 2003, Pang et al., 2003). We approach this problem as one of statistical machine translation (SMT), within the noisy channel model of Brown et al. (1993). That is, we seek to identify the optimal paraphrase *T** of a sentence *S* by finding:

$$T^* = \arg\max_T \{\mathrm{P}(T \mid S)\} = \arg\max_T \{\mathrm{P}(S \mid T)\,\mathrm{P}(T)\}$$

T and *S* being sentences in the same language.

We describe and evaluate an SMT-based paraphrase generation system that utilizes a monotone phrasal decoder to generate meaning-preserving paraphrases across multiple domains. By adopting at the outset a paradigm geared toward generating sentences, this approach overcomes many problems encountered by task-specific approaches. In particular, we show that SMT techniques can be extended to paraphrase given sufficient monolingual parallel data.[1] We show that a huge corpus of comparable and alignable sentence pairs can be culled from ready-made topical/temporal clusters of news articles gathered on a daily basis from thousands of sources on the World Wide Web, thereby permitting the system to operate outside the narrow domains typical of existing systems.

2 Related work

Until recently, efforts in paraphrase were not strongly focused on generation and relied primarily on narrow data sources. One data source has been multiple translations of classic literary works (Barzilay & McKeown 2001; Ibrahim 2002; Ibrahim et al. 2003). Pang et al. (2003) obtain parallel monolingual texts from a set of 100 multiply-translated news articles. While translation-based approaches to obtaining data do address the problem of how to identify two strings as meaning the same thing, they are limited in scalability owing to the difficulty (and expense) of obtaining large quantities of multiply-translated source documents.

Other researchers have sought to identify patterns in large unannotated monolingual corpora. Lin & Pantel (2002) derive inference rules by parsing text fragments and extracting semantically similar paths. Shinyama et al. (2002) identify dependency paths in two collections of newspaper articles. In each case, however, the information extracted is limited to a small set of patterns.

Barzilay & Lee (2003) exploit the meta-information implicit in dual collections of news-

[1] Barzilay & McKeown (2001), for example, reject the idea owing to the noisy, comparable nature of their data.

wire articles, but focus on learning sentence-level patterns that provide a basis for generation. Multi-sequence alignment (MSA) is used to identify sentences that share formal (and presumably semantic) properties. This yields a set of clusters, each characterized by a word lattice that captures n-gram-based structural similarities between sentences. Lattices are in turn mapped to templates that can be used to produce novel transforms of input sentences. Their methodology provides striking results within a limited domain characterized by a high frequency of stereotypical sentence types. However, as we show below, the approach may be of limited generality, even within the training domain.

3 Data collection

Our training corpus, like those of Shinyama et al. and Barzilay & Lee, consists of different news stories reporting the same event. While previous work with comparable news corpora has been limited to just two news sources, we set out to harness the ongoing explosion in internet news coverage. Thousands of news sources worldwide are competing to cover the same stories, in real time. Despite different authorship, these stories cover the same events and therefore have significant content overlap, especially in reports of the basic facts. In other cases, news agencies introduce minor edits into a single original AP or Reuters story. We believe that our work constitutes the first to attempt to exploit these massively multiple data sources for paraphrase learning and generation.

3.1 Gathering aligned sentence pairs

We began by identifying sets of pre-clustered URLs that point to news articles on the Web, gathered from publicly available sites such as http://news.yahoo.com/, http://news.google.com and http://uk.newsbot.msn.com. Their clustering algorithms appear to consider the full text of each news article, in addition to temporal cues, to produce sets of topically/temporally related articles. Story content is captured by downloading the HTML and isolating the textual content. A supervised HMM was trained to distinguish story content from surrounding advertisements, etc.[2]

Over the course of about 8 months, we collected 11,162 clusters, comprising 177,095 articles and averaging 15.8 articles per cluster. The quality of these clusters is generally good. Impressionistically, discrete events like sudden disasters, business announcements, and deaths tend to yield tightly focused clusters, while ongoing stories like the SARS crisis tend to produce very large and unfocused clusters.

To extract likely paraphrase sentence pairs from these clusters, we used edit distance (Levenshtein 1966) over words, comparing all sentences pairwise within a cluster to find the minimal number of word insertions and deletions transforming the first sentence into the second. Each sentence was normalized to lower case, and the pairs were filtered to reject:

- Sentence pairs where the sentences were identical or differed only in punctuation;
- Duplicate sentence pairs;
- Sentence pairs with significantly different lengths (the shorter is less than two-thirds the length of the longer);
- Sentence pairs where the Levenshtein distance was greater than 12.0.[3]

A total of 139K non-identical sentence pairs were obtained. Mean Levenshtein distance was 5.17; mean sentence length was 18.6 words.

3.2 Word alignment

To this corpus we applied the word alignment algorithms available in Giza++ (Och & Ney, 2000), a freely available implementation of IBM Models 1-5 (Brown, 1993) and the HMM alignment (Vogel et al, 1996), along with various improvements and modifications motivated by experimentation by Och & Ney (2000). In order to capture the many-to-many alignments that identify correspondences between idioms and other phrasal chunks, we align in the forward direction and again in the backward direction, heuristically recombining each unidirectional word alignment into a single bidirectional alignment (Och & Ney 2000). Figure 1 shows an example of a monolingual alignment produced by Giza++. Each line represents a unidirectional link; directionality is indicated by a tick mark on the target side of the link.

We held out a set of news clusters from our training data and extracted a set of 250 sentence pairs for blind evaluation. Randomly extracted on the basis of an edit distance of $5 \leq n \leq 20$ (to allow a range of reasonably divergent candidate pairs while eliminating the most trivial substitutions), the gold-standard sentence pairs were checked by an independent human evaluator to ensure that

[2] We hand-tagged 1,150 articles to indicate which portions of the text were story content and which were advertisements, image captions, or other unwanted material. We evaluated several classifiers on a 70/30 test train split and found that an HMM trained on a handful of features was most effective in identifying content lines (95% F-measure).

[3] Chosen on the basis of ablation experiments and optimal AER (discussed in 3.2).

Figure 1. An example Giza++ alignment

they contained paraphrases before they were hand word-aligned.

To evaluate the alignments, we adhered to the standards established in Melamed (2001) and Och & Ney (2000, 2003). Following Och & Ney's methodology, two annotators each created an initial annotation for each dataset, subcategorizing alignments as either SURE (necessary) or POSSIBLE (allowed, but not required). Differences were highlighted and the annotators were asked to review their choices on these differences. Finally we combined the two annotations into a single gold standard: if both annotators agreed that an alignment should be SURE, then the alignment was marked as SURE in the gold-standard; otherwise the alignment was marked as POSSIBLE.

To compute Precision, Recall, and Alignment Error Rate (AER) for the twin datasets, we used exactly the formulae listed in Och & Ney (2003). Let A be the set of alignments in the comparison, S be the set of SURE alignments in the gold standard, and P be the union of the SURE and POSSIBLE alignments in the gold standard. Then we have:

$$\text{precision} = \frac{|A \cap P|}{|A|} \quad \text{recall} = \frac{|A \cap S|}{|S|}$$

$$\text{AER} = \frac{|A \cap P + A \cap S|}{|A + S|}$$

Measured in terms of AER[4], final interrater agreement between the two annotators on the 250 sentences was 93.1%.

Training Data Type:	L12
Precision	87.46%
Recall	89.52%
AER	**11.58%**
Identical word precision	89.36%
Identical word recall	89.50%
Identical word AER	**10.57%**
Non-identical word precision	76.99%
Non-identical word recall	90.22%
Non-identical word AER	**20.88%**

Table 1. AER on the Lev12 corpus

Table 1 shows the results of evaluating alignment after trainng the Giza++ model. Although the overall AER of 11.58% is higher than the best bilingual MT systems (Och & Ney, 2003), the training data is inherently noisy, having more in common with analogous corpora than conventional MT parallel corpora in that the paraphrases are not constrained by the source text structure. The identical word AER of 10.57% is unsurprising given that the domain is unrestricted and the alignment algorithm does not employ direct string matching to leverage word identity.[5] The non-identical word AER of 20.88% may appear problematic in a system that aims to generate paraphrases; as we shall see, however, this turns out not to be the case. Ablation experiments, not described here, indicate that additional data will improve AER.

3.3 Identifying phrasal replacements

Recent work in SMT has shown that simple phrase-based MT systems can outperform more sophisticated word-based systems (e.g. Koehn et al. 2003). Therefore, we adopt a phrasal decoder patterned closely after that of Vogel et al. (2003).

We view the source and target sentences S and T as word sequences $s_1..s_m$ and $t_1..t_n$. A word alignment A of S and T can be expressed as a function from each of the source and target tokens to a unique *cept* (Brown et al. 1993); isomorphically, a cept represents an aligned subset of the source and target tokens. Then, for a given sentence pair and word alignment, we define a *phrase pair* as a subset of the cepts in which both the source and target tokens are contiguous.[6] We gathered all phrase

[4] The formula for AER given here and in Och & Ney (2003) is intended to compare an automatic alignment against a gold standard alignment. However, when comparing one human against another, both comparison and reference distinguish between SURE and POSSIBLE links. Because the AER is asymmetric (though each direction differs by less than 5%), we have presented the average of the directional AERs.

[5] However, following SMT practice of augmenting data with a bilingual lexicon, we did append an identity lexicon to the training data.

[6] While this does preclude the usage of "gapped" phrase pairs such as *or* → *either … or*, we found such map-

pairs (limited to those containing no more than five cepts, for reasons of computational efficiency) occurring in at least one aligned sentence somewhere in our training corpus into a single replacement database. This database of lexicalized phrase pairs, termed *phrasal replacements*, serves as the backbone of our channel model.

As in (Vogel et al. 2003), we assigned probabilities to these phrasal replacements via IBM Model 1. In more detail, we first gathered lexical translation probabilities of the form P(*s* | *t*) by running five iterations of Model 1 on the training corpus. This allows for computing the probability of a sequence of source words *S* given a sequence of target words *T* as the sum over all possible alignments of the Model 1 probabilities:

$$P(S \mid T) = \sum_{A} P(S, A \mid T) = \prod_{t \in T} \sum_{s \in S} P(s \mid t)$$

(Brown et al. (1993) provides a more detailed derivation of this identity.) Although simple, this approach has proven effective in SMT for several reasons. First and foremost, phrasal scoring by Model 1 avoids the sparsity problems associated with estimating each phrasal replacement probability with MLE (Vogel et al. 2003). Secondly, it appears to boost translation quality in more sophisticated translation systems by inducing lexical triggering (Och et al. 2004). Collocations and other non-compositional phrases receive a higher probability as a whole than they would as independent single word replacements.

One further simplification was made. Given that our domain is restricted to the generation of monolingual paraphrase, interesting output can be produced without tackling the difficult problem of inter-phrase reordering.[7] Therefore, along the lines of Tillmann et al. (1997), we rely on only monotone phrasal alignments, although we do allow intra-phrasal reordering. While this means certain common structural alternations (e.g., active/passive) cannot be generated, we are still able to express a broad range of phenomena:

- **Synonymy**: *injured* → *wounded*
- **Phrasal replacements**: *Bush administration* → *White House*
- **Intra-phrasal reorderings**: *margin of error* → *error margin*

Our channel model, then, is determined solely by the phrasal replacements involved. We first assume a monotone decomposition of the sentence pair into phrase pairs (considering all phrasal decompositions equally likely), and the probability P(*S* | *T*) is then defined as the product of the each phrasal replacement probability.

The target language model was a trigram model using interpolated Kneser-Ney smoothing (Kneser & Ney 1995), trained over all 1.4 million sentences (24 million words) in our news corpus.

3.4 Generating paraphrases

To generate paraphrases of a given input, a standard SMT decoding approach was used; this is described in more detail below. Prior to decoding, however, the input sentence underwent preprocessing: text was lowercased, tokenized, and a few classes of named-entities were identified using regular expressions.

To begin the decoding process, we first constructed a lattice of all possible paraphrases of the source sentence based on our phrasal translation database. Figure 2 presents an example. The lattice was realized as a set of $|S| + 1$ vertices $v_0..v_{|S|}$ and a set of edges between those vertices; each edge was labeled with a sequence of words and a real number. Thus a edge connecting vertex v_i to v_j labeled with the sequence of words $w_1..w_k$ and the real number p indicates that the source words s_{i+1} to s_j can be replaced by words $w_1..w_k$ with probability p. Our replacement database was stored as a trie with words as edges, hence populating the lattice takes worst case $O(n^2)$ time. Finally, since source and target languages are identical, we added an identity mapping for each source word s_i: an edge from v_{i-1} to v_i with label s_i and a uniform probability u. This allows for handling unseen words. A high u value permits more conservative paraphrases.

We found the optimal path through the lattice as scored by the product of the replacement model and the trigram language model. This algorithm reduces easily to the Viterbi algorithm; such a dynamic programming approach guarantees an efficient optimal search (worst case $O(kn)$, where n is the maximal target length and k is the maximal number of replacements for any word). In addition, fast algorithms exist for computing the n-best lists over a lattice (Soong & Huang 1991).

---

pings to be both unwieldy in practice and very often indicative of poor a word alignment.

[7] Even in the realm of MT, such an assumption can produce competitive results (Vogel et al. 2003). In addition, we were hesitant to incur the exponential increase in running time associated with those movement models in the tradition of Brown el al (1993), especially since these offset models fail to capture important linguistic generalizations (e.g., phrasal coherence, headedness).

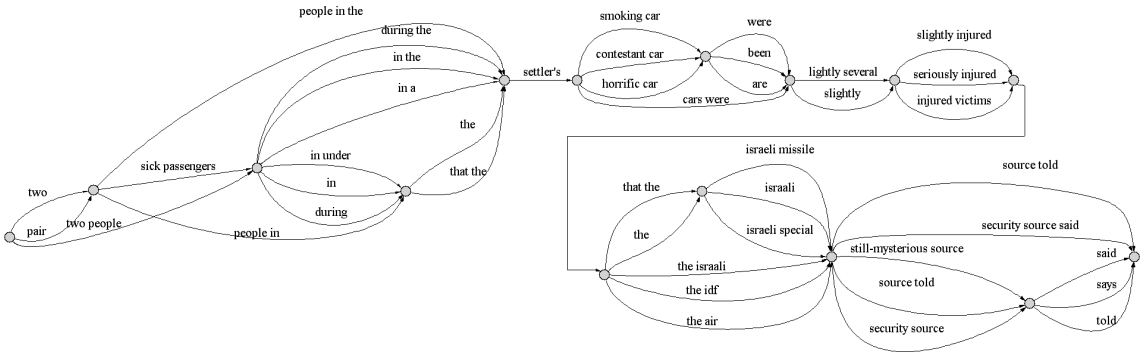

Figure 2. A simplified generation lattice: 44 top ranked edges from a total 4,140

Finally the resultant paraphrases were cleaned up in a post-processing phase to ensure output was not trivially distinguishable from other systems during human evaluation. All generic named entity tokens were re-instantiated with their source values, and case was restored using a model like that used in Vita et al. (2003).

3.5 Alternate approaches

Barzilay & Lee (2003) have released a common dataset that provides a basis for comparing different paraphrase generation systems. It consists of 59 sentences regarding acts of violence in the Middle East. These are accompanied by paraphrases generated by their Multi-Sequence Alignment (MSA) system and a baseline employing WordNet (Fellbaum 1998), along with human judgments for each output by 2-3 raters.

The MSA WordNet baseline was created by selecting a subset of the words in each test sentence—proportional to the number of words replaced by MSA in the same sentence—and replacing each with an arbitrary word from its most frequent WordNet synset.

Since our SMT approach depends quite heavily on a target language model, we presented an alternate WordNet baseline using a target language model.[8] In combination with the language model described in section 3.4, we used a very simple replacement model: each appropriately inflected member of the most frequent synset was proposed as a possible replacement with uniform probability. This was intended to isolate the contribution of the language model from the replacement model.

Given that our alignments, while aggregated into phrases, are fundamentally word-aligned, one question that arises is whether the information we learn is different in character than that learned from much simpler techniques. To explore this hypothesis, we introduced an additional baseline that used statistical clustering to produce an automated, unsupervised synonym list, again with a trigram language model. We used standard bigram clustering techniques (Goodman 2002) to produce 4,096 clusters of our 65,225 vocabulary items.

4 Evaluation

We have experimented with several methods for extracting a parallel sentence-aligned corpus from news clusters using word alignment error rate, or AER, (Och & Ney 2003) as an evaluation metric. A brief summary of these experiments is provided in Table 1. To evaluate the quality of generation, we followed the lead of Barzilay & Lee (2003). We started with the 59 sentences and corresponding paraphrases from MSA and WordNet (designated as WN below). Since the size of this data set made it difficult to obtain statistically significant results, we also included 141 randomly selected sentences from held-out clusters. We then produced paraphrases with each of the following systems and compared them with MSA and WN:

- **WN+LM**: WordNet with a trigram LM
- **CL**: Statistical clusters with a trigram LM
- **PR**: The top 5 sentence rewrites produced by Phrasal Replacement.

For the sake of consistency, we did not use the judgments provided by Barzilay and Lee; instead we had two raters judge whether the output from each system was a paraphrase of the input sentence. The raters were presented with an input sentence and an output paraphrase from each system in random order to prevent bias toward any particular judgment. Since, on our first pass, we found inter-rater agreement to be somewhat low (84%), we asked the raters to make a second pass of judgments on those where they disagreed; this significantly improved agreement (96.9%). The results of this final evaluation are summarized in Table 2.

[8] In contrast, Barzilay and Lee (2003) avoided using a language model for essentially the same reason: their MSA approach did not take advantage of such a resource.

Method	B&L59	B&L59 + 141
PR #1	54 / 59 = **91.5%**	177 / 200 = **89.5%**
PR #2	53 / 59 = **89.8%**	168 / 200 = **84.0%**
PR #3	46 / 59 = **78.0%**	164 / 200 = **82.0%**
PR #4	49 / 59 = **83.1%**	163 / 200 = **81.5%**
MSA	46 / 59 = **78.0%**	46 / 59 = **78.0%**
PR #5	44 / 59 = **74.6%**	155 / 200 = **77.5%**
WN	23 / 59 = **39.0%**	25 / 59 = **37.9%**
WN+LM	30 / 59 = **50.9%**	53 / 200 = **27.5%**
CL	14 / 59 = **23.7%**	26 / 200 = **13.0%**

Table 2. Human acceptability judgments

5 Analysis

Table 2 shows that PR can produce rewordings that are evaluated as plausible paraphrases more frequently than those generated by either baseline techniques or MSA. The WordNet baseline performs quite poorly, even in combination with a trigram language model: the language model does not contribute significantly to resolving lexical selection. The performance of CL is likewise abysmal—again a language model does nothing to help. The poor performance of these synonym-based techniques indicates that they have little value except as a baseline.

The PR model generates plausible paraphrases for the overwhelming majority of test sentences, indicating that even the relatively high AER for non-identical words is not an obstacle to successful generation. Moreover, PR was able to generate a paraphrase for all 200 sentences (including the 59 MSA examples). The correlation between acceptability and PR sentence rank validates both the ranking algorithm and the evaluation methodology.

In Table 2, the PR model scores significantly better than MSA in terms of the percentage of paraphrase candidates accepted by raters. Moreover, PR generates at least five (and often hundreds more) distinct paraphrases for each test sentence. Such perfect coverage on this dataset is perhaps fortuitous, but is nonetheless indicative of scalability. By contrast Barzilay & Lee (2003) report being able to generate paraphrases for only 59 out of 484 sentences in their test set, a total of 12%.

One potential concern is that PR paraphrases usually involve simple substitutions of words and short phrases (a mean edit distance of 2.9 on the top ranked sentences), whereas MSA outputs more complex paraphrases (reflected in a mean edit distance of 25.8). This is reflected in Table 3, which provides a breakdown of four dimensions of interest, as provided by one of our independent evaluators. Some 47% of MSA paraphrases involve significant reordering, such as an active-passive alternation, whereas the monotone PR decoder precludes anything other than minor transpositions within phrasal replacements.

	MSA	PR#1
Rearrangement	28 / 59 = **47%**	0 / 100 = **0%**
Phrasal alternation	11 / 59 = **19%**	3 / 100 = **3%**
Info added	19 / 59 = **32%**	6 / 100 = **6%**
Info lost	43 / 59 = **73%**	31 / 100 = **31%**

Table 3. Qualitative analysis of paraphrases

Should these facts be interpreted to mean that MSA, with its more dramatic rewrites, is ultimately more ambitious than PR? We believe that the opposite is true. A close look at MSA suggests that it is similar in spirit to example-based machine translation techniques that rely on pairing entire sentences in source and target languages, with the translation step limited to local adjustments of the target sentence (e.g. Sumita 2001). When an input sentence closely matches a template, results can be stunning. However, MSA achieves its richness of substitution at the cost of generality. Inspection reveals that 15 of the 59 MSA paraphrases, or 25.4%, are based on a single high-frequency, domain-specific template (essentially a running tally of deaths in the Israeli-Palestinian conflict). Unless one is prepared to assume that similar templates can be found for most sentence types, scalability and domain extensibility appear beyond the reach of MSA.

In addition, since MSA templates pair entire sentences, the technique can produce semantically different output when there is a mismatch in information content among template training sentences. Consider the third and fourth rows of Table 3, which indicate the extent of embellishment and lossiness found in MSA paraphrases and the top-ranked PR paraphrases. Particularly noteworthy is the lossiness of MSA seen in row 4. Figure 3 illustrates a case where the MSA paraphrase yields a significant reduction in information, while PR is more conservative in its replacements.

While the substitutions obtained by the PR model remain for the present relatively modest, they are not trivial. Changing a single content word is a legitimate form of paraphrase, and the ability to paraphrase across an arbitrarily large sentence set and arbitrary domains is a desideratum of paraphrase research. We have demonstrated that the SMT-motivated PR method is capable of generating acceptable paraphrases for the overwhelming majority of sentences in a broad domain.

6 Future work

Much work obviously remains to be done. Our results remain constrained by data sparsity, despite the large initial training sets. One major agenda item therefore will be acquisition of larger (and more diverse) data sets. In addition to obtaining greater absolute quantities of data in the form of clustered articles, we also seek to extract aligned sentence pairs that instantiate a richer set of phenomena. Relying on edit distance to identify likely paraphrases has the unfortunate result of excluding interesting sentence pairs that are similar in meaning though different in form. For example:

The Cassini spacecraft, which is en route to Saturn, is about to make a close pass of the ringed planet's mysterious moon Phoebe

On its way to an extended mission at Saturn, the Cassini probe on Friday makes its closest rendezvous with Saturn's dark moon Phoebe.

We are currently experimenting with data extracted from the first two sentences in each article, which by journalistic convention tend to summarize content (Dolan et al. 2004). While noisier than the edit distance data, initial results suggest that these can be a rich source of information about larger phrasal substitutions and syntactic reordering.

Although we have not attempted to address the issue of paraphrase identification here, we are currently exploring machine learning techniques, based in part on features of document structure and other linguistic features that should allow us to bootstrap initial alignments to develop more data. This will we hope, eventually allow us to address such issues as paraphrase identification for IR.

To exploit richer data sets, we will also seek to address the monotone limitation of our decoder that further limits the complexity of our paraphrase output. We will be experimenting with more sophisticated decoder models designed to handle reordering and mappings to discontinuous elements. We also plan to pursue better (automated) metrics for paraphrase evaluation.

7 Conclusions

We presented a novel approach to the problem of generating sentence-level paraphrases in a broad semantic domain. We accomplished this by using methods from the field of SMT, which is oriented toward learning and generating exactly the sorts of alternations encountered in monolingual paraphrase. We showed that this approach can be used to generate paraphrases that are preferred by humans to sentence-level paraphrases produced by other techniques. While the alternations our system produces are currently limited in character, the field of SMT offers a host of possible enhancements—including reordering models—affording a natural path for future improvements.

A second important contribution of this work is a method for building and tracking the quality of large, alignable monolingual corpora from structured news data on the Web. In the past, the lack of such a data source has hampered paraphrase research; our approach removes this obstacle.

Acknowledgements

We are grateful to Mo Corston-Oliver, Jeff Stevenson, Amy Muia, and Orin Hargraves of the Butler Hill Group for their work in annotating the data used in the experiments. This paper has also benefited from discussions with Ken Church, Mark Johnson, and Steve Richardson. We greatly appreciate the careful comments of three anonymous reviewers. We remain, however, solely responsible for this content.

References

R. Barzilay and K. R. McKeown. 2001. Extracting Paraphrases from a parallel corpus. In *Proceedings of the ACL/EACL*.

R. Barzilay and L. Lee. 2003. Learning to Paraphrase; an unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*.

P. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation. *Computational Linguistics* 19(2): 263-311.

W. Dolan, C. Quirk and C. Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. To appear in *Proceedings of COLING-2004*.

C. Fellbaum, ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. Goodman. 2002. JCLUSTER. Software available at http://research.microsoft.com/~joshuago/

A. Ibrahim. 2002. *Extracting Paraphrases from Aligned Corpora.* Master of Engineering Thesis, MIT.

A. Ibrahim, B. Katz, and J. Lin. 2003. Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceedings of the Second International Workshop on Paraphrasing (IWP 2003)*. Sapporo, Japan.

R. Kneser and H. Ney. 1995. Improved backing-off for N-gram language modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*: 181-184. Detroit, MI.

P. Koehn, F. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL*.

V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physice-Doklady* 10: 707-710.

D. Lin and P. Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: 323-328.

I. D. Melamed. 2001. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press.

R. Mihalcea and T. Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *HLT/NAACL Workshop: Building and Using Parallel Texts*: 1-10.

F. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the ACL*: 440-447. Hong Kong, China.

F. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1): 19-52.

B. Pang, K. Knight, and D. Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. *Proceedings of HLT/NAACL*.

Y. Shinyama, S. Sekine and K. Sudo. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of NAACL-HLT*.

F. K. Soong and E. F. Huang. 1991. A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* 1: 705-708. Toronto, Canada.

E. Sumita. 2001. Example-based machine translation using DP-matching between work sequences. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*: 1–8.

C. Tillmann, S. Vogel, H. Ney, and A. Zubaiga. 1997. A DP Based Search Using Monotone Alignments in Statistical Translation. In *Proceedings of the ACL*.

L. Vita, A. Ittycheriah, S. Roukos, and N. Kambhatla. 2003. tRuEcasing. In *Proceedings of the ACL*: 152-159. Sapporo, Japan.

S. Vogel, H. Ney and C. Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the ACL*: 836-841. Copenhagen Denmark.

S. Vogel, Y. Zhang, F. Huang, A. Venugopal, B. Zhao, A. Tribble, M. Eck, and A. Waibel. 2003. The CMU Statistical Machine Translation System. In *Proceedings of MT Summit IX*, New Orleans, Louisiana, USA.