2006-12-01

# Generating Paraphrases with Greater Variation Using Syntactic Phrases

Rebecca Diane Madsen
*Brigham Young University - Provo*

# GENERATING PARAPHRASES WITH GREATER VARIATION

# USING SYNTACTIC PHRASES

by

Rebecca Madsen

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computer Science

Brigham Young University

December 2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Rebecca Madsen

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

| | |
|---|---|
| _____ | _____ |
| Date | Eric Ringger, Chair |
| _____ | _____ |
| Date | Deryle Lonsdale |
| _____ | _____ |
| Date | Eric Mercer |

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Rebecca Madsen in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill the university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

| | |
|---|---|
| Date | Eric Ringger<br>Chair, Graduate Committee |

Accepted for the Department

| | |
|---|---|
| Date | Parris Egbert<br>Graduate Coordinator |

Accepted for the College

| | |
|---|---|
| Date | Thomas Sederberg<br>Associate Dean, College of Physical and Mathematical Sciences |

# ABSTRACT

## GENERATING PARAPHRASES WITH GREATER VARIATION USING SYNTACTIC PHRASES

Rebecca Madsen

Department of Computer Science

Master of Science

Given a sentence, a paraphrase generation system produces a sentence that says the same thing but usually in a different way. The paraphrase generation problem can be formulated in the machine translation paradigm; instead of translation of English to a foreign language, the system translates an English sentence (for example) to another English sentence. Quirk et al. (2004) demonstrated this approach to generate almost 90% acceptable paraphrases. However, most of the sentences had little variation from the original input sentence.

Leveraging syntactic information, this thesis project presents an approach that successfully generated more varied paraphrase sentences than the approach of Quirk et al. while maintaining coverage of the proportion of acceptable paraphrases generated. The ParaMeTer system (Paraphrasing by MT) identifies syntactic chunks in paraphrase

sentences and substitutes labels for those chunks. This enables the system to generalize movements that are more syntactically plausible, as syntactic chunks generally capture sets of words that can change order in the sentence without losing grammaticality. ParaMeTer then uses statistical phrase-based MT techniques to learn alignments for the words and chunk labels alike. The baseline system followed the same pattern as the Quirk et al. system – a statistical phrase-based MT system.

Human judgments showed that the syntactic approach and baseline both achieve approximately the same ratio of fluent, acceptable paraphrase sentences per fluent sentences. These judgments also showed that the ParaMeTer system has more phrase rearrangement than the baseline system. Though the baseline has more within-phrase alteration, future modifications such as a chunk-only translation model should improve ParaMeTer's variation for phrase alteration as well.

# ACKNOWLEDGMENTS

My sincere thanks to my advisor, Eric Ringger, for helping me focus on the details of this project and for encouraging me to pursue my ideas and interests and giving me guidance once I had chosen a topic. Thanks also go to Deryle Lonsdale for his continued support and enthusiasm in satisfying my other research interests and for his faith in my abilities to accomplish whatever I set out to do.

Thanks go to those who allowed the use of their toolkits and software and data in creating the ParaMeTer system. Without these my project would be nothing; I indeed stood on the shoulders of giants. I would also like to thank Chris Quirk, Chris Brockett, and Bill Dolan for the description of their system, and for responding to my questions.

I am indebted to all forty of those who participated in the survey; without them I would have been unable to evaluate my system. You know who you are, even though I do not. I would also like to thank Neil Mayo and the WebExp2 developers for the survey software, and my brother Karl for his helpful cgi suggestions.

Most of all I would like to thank my husband, Wayne, for everything he does, for all the mischief abounding. Without him, I am sure I would have eaten nothing but grilled cheese and pizza while working on my thesis. His antics and excursions to distract me when I needed distractions, and his efforts to encourage me to work every other moment, helped see me through to the end.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

## 1. Introduction

The ability to restate a given phrase in a different way is one way we as humans can try to help others understand what we *really* mean, or to show that we understood what someone else said to us. Two statements are said to be paraphrases of one another when they use different words to achieve the same semantic equivalence – when they "mean the same thing." Being able to generate paraphrases using a computer program would be useful for multiple applications: query reformulation in information retrieval (e.g. web search or help system searches), confirming understanding in a dialogue, and generating a shorter paraphrase in summarization, among other tasks.

Generally the paraphrase identification (or recognition) problem in natural language processing (NLP), on the other hand, is formulated such that we are given two sentences and we must decide if the two sentences are equivalent or not. The majority of paraphrase research so far has focused on identifying paraphrases rather than generating paraphrases. In part this is due to the fact that it is easier in general to achieve better results on the discrimination problem than on the generation problem. However, the issue of generating possible paraphrases is interesting because there are different types of applications that can take advantage of a system that can create a new way of saying the same thing.

One recent approach to the paraphrase generation problem suggests formulating the problem of generating paraphrases of sentences as a machine translation (MT) problem (Quirk et al., 2004). Instead of translating a sentence from French (for example) to English, the system translates from *English to English* – possibly using different words or a different ordering for the resulting sentence.

Research in statistical MT (Koehn et al., 2003) has shown that phrasal MT, using alignments between sequences of words, is a better approach than word-based MT, aligning single words. While some work has explored the effects of leveraging syntactic methods and structure for statistical MT (e.g. Chiang, 2005; Quirk et al., 2005; Riezler and Maxwell, 2006; Zhang et al., 2006), most continue to focus on statistical word-based or phrase-based approaches.

This thesis presents ParaMeTer (Paraphrasing by MT), a syntax-based statistical MT approach to sentential paraphrase which performs better than an approach using statistically identified phrases alone. The evidence for improvement of performance is that humans judged the syntactic MT system to generate more unique paraphrases than a statistical phrase-based MT system. In brief, the system developed for this thesis employs the following method: syntactic phrases are identified and replaced by a label before using statistical phrase-based MT methods, so that whole syntactic phrasal movement can be learned; thus general types of movements may be learned instead of memorizing specific phrases.

## 2. Statistical Machine Translation

Machine translation is typically viewed as a noisy channel problem (or source-channel problem). In this formulation, a source produces a sentence in one language (L1). Figure 1 demonstrates how the L1 sentence passes through a (possibly noisy) channel that transforms it into a sentence in a second language (L2).

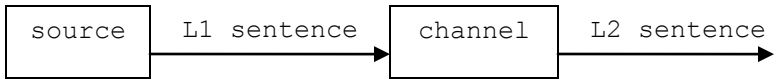


**Figure 1**: The source-channel model of Machine Translation.

To apply the model to a given sentence in L2, a decoder must construct the L1 sentence most likely to have produced the L2 sentence, as shown in Figure 2.


**Figure 2**: Translation as a decoding process.

Perhaps contrary to intuition, L1 is called the "target language," and L2 is called the "source language" for MT since the sentences we observe come from L2, and we are translating into L1. Thus, we usually have a language model (describing the source) to encourage fluency and grammaticality in the target language, a translation model (describing the channel) of how words and phrases are changed while "passing through the channel", and a decoder that builds the most likely sentence in the target language (i.e., the best translation) according to the statistical models.

### *2.1.* Language model

Language modeling is the task of assigning a probability to any word string in a language. The probability of word sequence $\overline{w} = w_1 \ldots w_n$, or sentence, can be expressed

(factored by the chain rule of joint probability from left to right) as the probability of each word given everything that came before it:

$$(1)\quad P(\overline{w}) = P(w_1)P(w_1 \mid w_2)P(w_3 \mid w_1, w_2)\cdots P(w_i \mid w_1, \cdots, w_{i-1})$$

The right-hand side of Equation 1 can be rewritten as the following:

$$(2)\quad P(\overline{w}) = \prod_{i=1}^{n} P(w_i \mid w_1 \cdots w_{i-1})$$

When we talk about trigram language models, we simplify the product with the assumption that the probability of a word is conditionally independent of all earlier words given the previous two words. This is expressed in Equation 3:

$$(3)\quad P(\overline{w}) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1}, w_{i-2})$$

A language model is commonly used to identify more probable sentences; it is trained from a corpus of sentences in the desired language. Since a well-constructed language model, such as a smoothed version of the above trigram language model, is able to assign a non-zero probability to any given sentence, it is useful for machine translation even in cases where sentences (or their parts) have not been seen before.

*2.2.* Alignment model

The word alignment problem consists of trying to find the possible connections between words, given a parallel corpus (i.e. pairs of corresponding sentences). In translation, this problem can be thought of as trying to automatically learn possible translations of a word in a foreign language, though word alignment systems are designed to learn the statistical strength of links between words, not to learn a foreign language dictionary. Equation 4 shows how translation probabilities are calculated. The probability of a translation (or target) sentence *t*, given a foreign (source) sentence *s*, can be computed by summing the probability of *s* and an alignment *a*, given *t*, over all possible alignments.

$$(4)\quad P(s \mid t) = \sum_{a} P(s, a \mid t)$$

Summing over all possible alignments depends, of course, on the ability to *enumerate* all alignments. By making assumptions about which alignments are possible, the problem is greatly simplified. Brown et al. (1993) at IBM Research proposed a series of models, known as IBM Models 1 through 5, that allow for increasingly complex alignments.

For present purposes, phrases are sequences of consecutive words in a sentence (Koehn et al., 2003). A phrase translation model can be trained from a word alignment model by collecting aligned phrase pairs that are consistent with the word alignment model. A phrase is consistent with the word alignment model if the words in the phrase are only aligned to each other and not to words outside the phrase. Once possible phrases

have been determined, phrase translation probabilities are calculated by the relative frequency of the phrase with no smoothing (i.e. maximum likelihood estimation).

*2.3.* Decoder

For a given source sentence, the decoder consults the language model and the translation model as it enumerates possible target sentences in an efficient manner (typically by dynamic programming). One way to do this is for the decoder to construct a lattice of possible translations (or paraphrases) of the source sentence in the test data and then to find the optimal path through this lattice. The decoder looks for the highest probability of a translation sentence *t*, given the source (or foreign) sentence *s*, as shown in Equation 5.

$$\hat{t} = \arg\max_{t} P(t \mid s) \tag{5}$$

Using Bayes rule, the probability of a translation given the source can be found by combining the alignment model score $P( s \mid t )$ and language model score $P( t )$ as in Equation 6.

$$\hat{t} = \arg\max_{t} \frac{P(s \mid t)P(t)}{P(s)} = \arg\max_{t} P(s \mid t)P(t) \tag{6}$$

Then the sentence *t* that maximizes the product of the language model and translation model is the optimal target sentence $\hat{t}$, according to the models.

# 3. Related Work

Research in using MT for sentential paraphrase is the basis of this thesis project, but syntax in the paraphrase task and in MT are associated research topics. Related work includes general uses of noisy channel models in NLP, general approaches to paraphrase generation, MT approaches to paraphrase generation, the use of syntax in MT, and paraphrase identification and recognizing textual entailment using syntactic information.

## *3.1.* Other noisy channel problems

Noisy channel models are used not just in MT, but in a variety of other NLP problems as well. Ringger (2000) employed a noisy channel model to post-process speech recognition output and correct the errors. This technique also helps when coupled with spoken dialogue systems to improve understanding and robustness of the dialogue system. Optical character recognition (Kolak et al., 2003) can also use a noisy channel model to correct output errors. Other noisy channel problems include handwriting recognition, speech recognition, and spelling correction.

## *3.2.* Paraphrase generation

### 3.2.1. MT approaches

Quirk et al. (2004) developed a paraphrase generation system that uses statistical MT techniques. They used a trigram language model, a word alignment translation model constructed using Giza++ (Och and Ney, 2003), and a phrasal decoder they built. The phrases they identified were sets of contiguous words which appeared in both the source and target sentences. Human evaluators found that almost 90% of the results from the

200 sentence test set were acceptable paraphrases, but that there was little or no rearrangement, phrasal alternation, or information added. In other words, though acceptable paraphrases, the results were generally exact matches, or simple word or phrase substitutions. I am not aware of any other work using a statistical MT approach to paraphrase generation.

### 3.2.2. General approaches

Quirk et al. used a paraphrase generation system by Barzilay and Lee (2003) as a point of comparison. Though Barzilay and Lee did not use MT techniques, they did use multi-sequence alignment to identify clusters of sentences that share similar properties. From these clusters they created word lattices, which constitute templates that can be used to generate novel sentences given the input. Quirk et al. found the Barzilay and Lee approach to be severely limited in domain and thus limited in generality. Essentially, the technique fails to allow for generalization. Other approaches to paraphrase are for paraphrase extraction (Barzilay and McKeown, 2001; Shinyama et al. 2002; Bannard and Callison-Burch, 2005) or identification, not generation; some methods focus on paraphrases of phrases (Callison-Burch et al., 2006; Kauchak and Barzilay, 2006) rather than sentences.

Pang et al. (2003) used syntax to build finite state automata (FSAs) from multiple reference translation sets. The FSAs can then be said to represent paraphrases, and extracting paths from these FSAs can create novel paraphrases. Any path in the FSA that starts at the start node and ends at the end node is considered a valid path allowed by the FSA, though it may not necessarily be grammatically well-formed.

### *3.3.* Syntactic phrases for MT

Though Koehn et al. (2003) found their statistical phrase-based MT system outperforms an approach that restricts phrasal alignments to those which are syntactically motivated, it is likely that this result depends on the particular approach taken in leveraging syntax. It is also possible that this result is specific to MT between two languages and would not hold for paraphrase generation. Note that the approach used in this thesis does not make Koehn et al.'s restriction on syntactically motivated phrasal alignments, but attempts to get the best of both worlds. ParaMeTer replaces syntactic phrases by their labels and then uses the more general and flexible statistically identified phrases, covering both words and syntactic phrasal labels.

Koehn and Knight (2003) demonstrated success in separating noun phrase (NP) and prepositional phrase (PP) translation as a sub-task of MT; they also showed improvement to translation accuracy when using maximum entropy to re-rank translation candidates based on syntactic features.

Quirk et al. (2005) also found improvement to be gained in using syntactic information in the form of dependency treelets in MT. Riezler and Maxwell (2006) presented a dependency-based statistical MT model that uses dependency-structure snippets in a grammar-based generator, and then uses an ordering model to improve grammaticality of translations. They found this approach improved grammaticality without losing coverage for translations. Zhang et al. (2006) demonstrated a method of binarization to move from the exponential complexity of systems based on synchronous grammars and tree transducers, to linear complexity for factoring syntactic reorderings in statistical MT.

Chiang (2005) shows that using hierarchical phrase pairs (phrases that contain subphrases) and learning a synchronous context-free grammar from an aligned corpus even without linguistically motivated syntax improves upon phrase-based techniques. Chiang distinguishes between formal syntax (using a context-free phrase structure grammar) and linguistic syntax (motivated by linguistic theory and assumptions).

### *3.4.* Recognizing textual entailment

PASCAL's Recognizing Textual Entailment challenge (RTE) (Dagan et al., 2005) is a related problem to paraphrase – or, rather, paraphrase detection is a special case of recognizing entailment: when both sentences entail the other they are said to be paraphrases of each other. Several systems competing in RTE employed some measure of linguistic knowledge, such as word overlap or syntactic structure or propositional logic when comparing a text and a hypothesis to determine if the text entailed the hypothesis (Dagan et al., 2005). Most of the systems typically reported F1 levels and accuracy levels between 0.5 and 0.6. Many of the authors of the competing systems also commented that the knowledge-based approaches needed more knowledge to perform better. It is not immediately obvious how these types of systems would be used to *generate* a hypothesis.

Bayer et al. (2005) compared their traditional linguistic-based approach to an approach inspired by MT techniques. The MT-inspired system significantly outperformed the traditional approach in terms of recall, but the authors drew the conclusion in the end that they expect results to improve if they could leverage the strengths and potential of both systems.

Using a set of heuristics to align logical forms of the text and hypothesis, Snow, Vanderwende, and Menezes (2006) demonstrated the effectiveness of using syntactic heuristics to recognize false entailment. A dependency graph-based approach is not unusual, but most systems focus on recognizing true entailment instead of false entailment.

### *3.5.* Applications for paraphrases

Applications for paraphrase generation include question answering, information retrieval, statistical machine translation, dialogue systems, and summarization tasks. Query reformulation in information retrieval could use paraphrase information. Being able to formulate other possible related queries using a paraphrase system could be used to automatically supplement a user's query, or to suggest other possible searches to a user through an interaction.

Dialogue systems are another area of NLP that could use the ability to generate paraphrases. Humans express understanding of a conversation by restating portions of the dialogue. A dialogue system could coordinate understanding of the other participant's dialogue move using paraphrase generation. Another way paraphrase could aid a dialogue system is in what are known as frame-based dialogue systems; frames are used to specify the semantics for system actions. User requests can vary in structure from the semantics represented in frames and fail to match the appropriate frames, but paraphrases of user requests could be used to map the requests to frames and allow the system to successfully interpret and handle them.

This is similar to another application of paraphrases – revising the question in the question-answering task (Duclaye et al., 2003; Rinaldi et al., 2003; Duboue and Chu-Carroll, 2006). By revising the question asked using paraphrase methods, recall can be improved and better answers found by the system.

Callison-Burch et al. (2006) used paraphrases at the phrase level to extend the coverage in statistical MT. When encountering an unknown phrase, substituting a paraphrase for the phrase allows the SMT system to generate a better translation.

Zhou et al. (2006) used paraphrases, also at the phrase level, for automatic summary evaluation. They extract a phrasal paraphrase lookup table from Chinese-English translations with multiple English translations. All English translations for a common Chinese phrase are considered to be paraphrases of each other. The paraphrase lookup table then is used to find the optimal paraphrase alignments between two summaries and determine the score.

Kauchak and Barzilay (2006) leveraged paraphrases to refine automatic evaluation techniques for machine translation and summarization tasks (e.g. BLEU and ROUGE). Revising the reference sentence to be closer to the machine-generated sentence by substituting words and phrases as they are admissible allows the reference sentence to be more useful in evaluation. They demonstrate results using this method that correlate better to human judgments.

## 4. Corpus Information

A paraphrase corpus consists of pairs of corresponding sentences, just as an MT parallel corpus (or bitext) does. The pairs of sentences in this case are not “translations” of each

other but “paraphrases” of each other, where each of the sentences in the pair is in the same language and has the same meaning as the other sentence.

### *4.1.* Description of the data

The Microsoft Research Paraphrase corpus (Dolan et al., 2004) was constructed by gathering news articles from Internet news sources that may have different authors, but cover the same stories. “Covering the same stories” means that there is significant content overlap or minor editorial differences, particularly when referring to factual information about a story. To extract pairs of sentence-level paraphrases, Dolan, et al. compared possible pairs using Levenshtein edit distance (at the word level) to measure the smallest number of word insertions and deletions it would take to convert one sentence to another. Words in the sentences were lowercased, and pairs that were duplicates or differed only in punctuation were automatically rejected. Pairs where one sentence was less than two-thirds as long as the other were also rejected, as were pairs with a Levenshtein distance greater than 12.0. In order to focus on more interesting paraphrases – that differed in more than just one or two word substitutions – pairs were also rejected that had a Levenshtein distance less than 8.0. The final corpus contains a total of 5,801 sentence pairs. The sentences in the final corpus are good examples of paraphrase, but were ultimately still not that different from each other. It seems an interesting conundrum to attempt to learn how to generate very different paraphrase sentences when the training sentence pairs are generally structured similarly.

The training data is made up of 70% of the sentence pairs (4076) and the test data is the remaining 30% (1725). Using human judges to label the sentence pairs, 67% (3900

sentence pairs) of the entire corpus was judged to be semantically equivalent (paraphrases); the rest of the dataset consists of sentence pairs that are not semantically equivalent. Two human judges were used to determine semantic equivalence, and a third was used in cases where they disagreed. The judges had 83% inter-rater agreement.

This means, however, that 33% of the training corpus cannot be included when training the alignment model – or the system would learn from incorrect examples of paraphrase. The language model can still train from the entire corpus since the language model just needs to have examples of English sentences in the same domain (i.e. the news domain), not a parallel paraphrase corpus.

For this thesis, all 3900 pairs of sentences from the MSR Paraphrase corpus that were judged as true paraphrases were used to train the alignment model, whether they were designated as training or test data by MSR. The remainder of the test data (consisting of non-paraphrase pairs) was further divided into development and test sets; 141 sentences were reserved for the test set, and 435 sentences were used for development. 59 sentences from Barzilay and Lee (2003) were also used in the test set, following Quirk et al. (2004).

The language model was supplemented by interpolating the 5225 sentence pairs (10,450 sentences) – 3900 from the true paraphrase examples and the remainder from the rest of the sentences in Microsoft's original training set – with a selection of 864,906 sentences (20,053,714 words) from the North American News Corpus.[1]

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T21

*4.2.* Chunks in the data

How often chunks of interest (e.g. prepositional phrases, noun phrases, relative clauses, complement clauses) occur, and at what level of granularity, helped to familiarize myself with what types of phrases I could expect to work with during later experiments. Looking at the types of phrases helped make it obvious that there were several levels of granularity, or different ways of looking for matches, that were possible. Some of the matching schemes I experimented with for this thesis, I report on here to help familiarize the reader with this perspective of the data: general phrase type, matching left-most word, matching right-most word, and exact phrase matches.

Of the 6459 occurrences of prepositional phrases (PPs) in the sentences on the left (i.e. sentence #1 as labeled in the MSR Paraphrase training corpus), 4700 of them occur when there is a prepositional phrase in the parallel sentence as well. This refers to any prepositional phrase (hence the general phrase type granularity level), not necessarily a matching phrase. Only 1720 of the PPs occurring in the sentence on the right (i.e. sentence #2 as labeled in the corpus) occur without a phrase in the parallel sentence.

At a more specific granularity, PPs with the left-most word "of" occur with a matching PP with "of" 1035 times; 78 different left-most words (of, with, on, from, between all occurring frequently) occur in a Zipf's law relation, as shown in Figure 3. The same trend is seen when considering the right-most word as a possible match; "percent" is the most common right-most word (54 instances), followed by "years" at 44 instances. More than 2800 different right-most words occur in PPs, again following Zipf's law (see Figure 4).

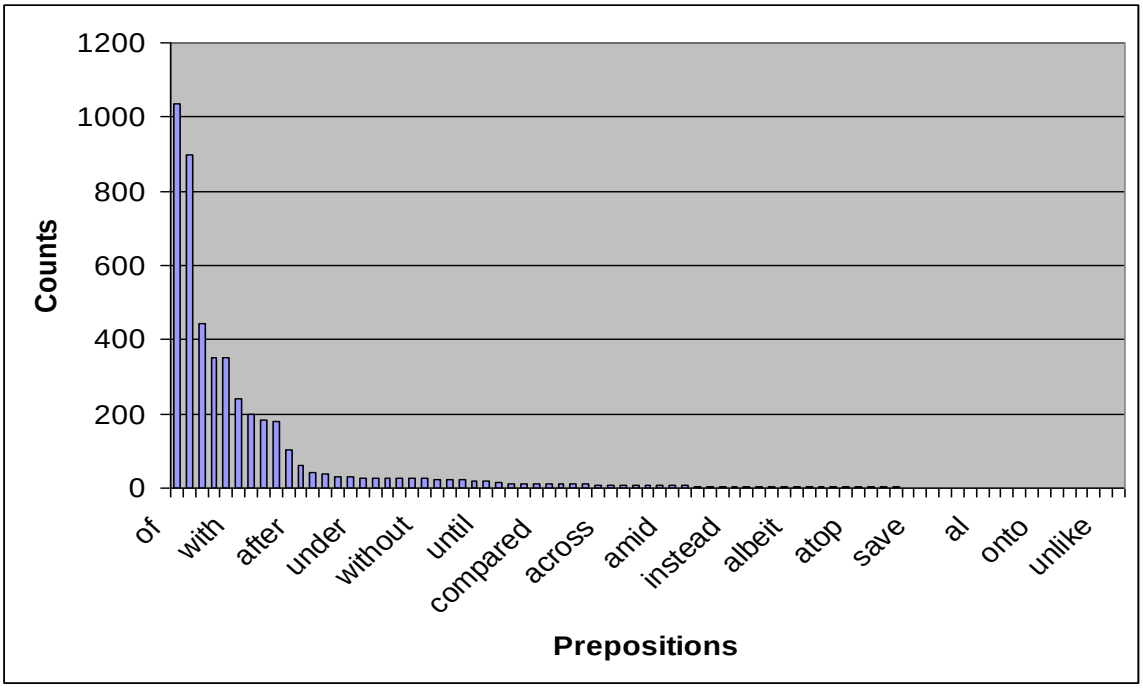


**Figure 3**: Counts of left-most words (generally the preposition) in PPs. Prepositions are sampled to show every fifth word in the set.

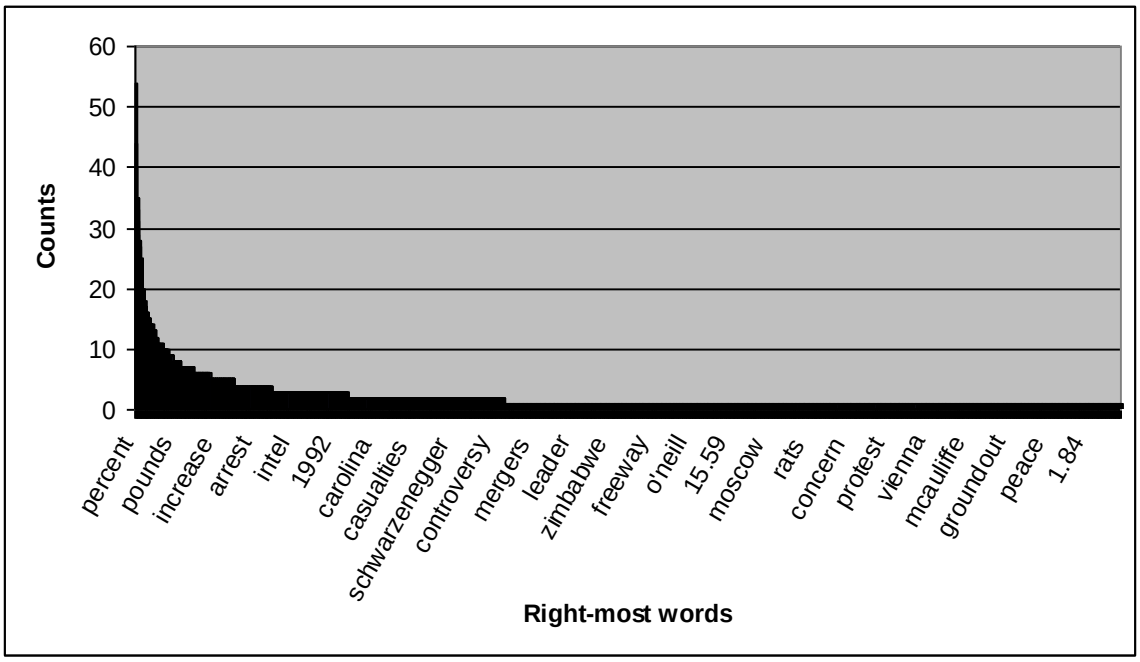


**Figure 4**: Counts of right-most words (generally the head noun) in PPs. Again, the right-most words are sampled, showing a subset of the words.

There are 3037 unique PPs in the corpus, 2745 of them occurring only once. The most common PP is “on the New York Stock Exchange,” occurring 14 times; “in Iraq” occurs 13 times, and “in prison” 11 times.

## 5. Baseline System

The baseline was modeled after the statistical phrase-based MT system Quirk et al. (2004) described. The same corpus was not available in this case, so some alterations and optimizations were made as described in this section.

### *5.1.* Training the language model

The language model used in this thesis was built using the SRI language model toolkit[2] (SRILM) (Stolcke, 2002). I employed the SRILM toolkit to build a trigram model using interpolated Kneser-Ney smoothing. Quirk et al. (2004) used 1.4 million sentences from their entire news corpus (not just the sentences identified as paraphrases) to train their language model. My language model, trained from the 5225 sentence pairs (10,450 sentences) in my training set, was interpolated with a model trained using 864,906 sentences (20,053,714 words) from the North American News Corpus using a weight of 0.5. Sentence breaking within paragraphs in the North American News Corpus was done using Alias-I’s LingPipe toolkit.[3]

These settings were chosen for two reasons: Quirk et al. (2004) used a trigram language model using interpolated Kneser-Ney smoothing, and experimentation with different settings demonstrated that this approach yielded the lowest perplexity when

[2] http://www.speech.sri.com/projects/srilm/
[3] http://www.alias-i.com/lingpipe/demos/tutorial/sentences/read-me.html

tested on the development test set. Using the SRILM default of Good-Turing smoothing, the language model built using the MSR corpus got a perplexity of 168.207. Kneser-Ney smoothing improved this to 144.106. The language model built from 864,906 sentences of the North American News Corpus got a perplexity of 212.780 (the perplexity steadily improved with each addition of more data). Mixing these two language models together with a weight of 0.5 got the lowest perplexity of these experiments, at 131.443.

Note that even though both the North American News Corpus and the MSR Paraphrase corpus come from the news domain, there is still a mismatch in years (and thus topics and people and events) the data covers. Quirk et al.'s LM data came from the same news sources and the same timeframe as their translation model data – 2003; the North American News Corpus pulls from the New York Times News Service, the Los Angeles Times – Washington Post News Service, and the English language portion of the Associated Press Worldstream newswire service from the years 1994-1998.

*5.2.* Phrasal alignment model

Giza++[4] implements common word-alignment models known as IBM Models 1-5 (Brown et al., 1993) and an HMM alignment model, along with modifications and improvements by Och. I applied the Giza++ tool (Och and Ney, 2003) to compute the alignment models using 5 iterations each of IBM Models 1 and the HMM model, and 3 iterations each of Models 3 and 4. The word alignment model was learned for both directions (L1 → L2, and L2 → L1), and Koehn's phrase translation scripts[5] were used to extract a statistical phrase translation model.

[4] http://www.fjoch.com/GIZA++.html
[5] http://www.iccs.informatics.ed.ac.uk/~pkoehn/

### *5.3.* Decoder settings

The decoder in statistical MT combines the language model and the translation model to produce paraphrase sentences; I employed the Pharaoh[6] decoder (Koehn, 2004) for this purpose. Settings on Pharaoh included a weight of 0.2 on the translation table, 0.5 on the language model, 0.2 for distortion (reordering), and -1 for sentence length penalties. Pharaoh can also return the top n-best sentences instead of just the top ranked sentence, but I use just the top sentence.

Pharaoh is capable of handling word alignments or phrasal alignments in generating and searching a lattice. The decoding process constructs translations of the test data, and uses a beam search (I used a beam size of 100) to determine the most likely translations (i.e. to find the best path(s) through the lattice).

## 6. ParaMeTer: Leveraging Chunks for MT

This thesis identifies relevant syntactic chunks for movement in paraphrase. These phrases are not just based on statistics of consecutive words but on syntax and abstract syntactic labels. The advantage of this is that syntactic phrases are coherent units which can be moved together as licensed by the syntax of the language; a statistically identified "phrase" or word sequence, on the other hand, cannot always move without scrambling the sentence nonsensically. Instead of identifying a set of consecutive words to "memorize," the ParaMeTer system can use syntactic chunk labels and learn possible reorderings that are more linguistically viable during training. Thus, ParaMeTer is better able to generalize possible movements in a translation during decoding.

[6] http://www.isi.edu/licensed-sw/pharaoh/

The primary aspect of paraphrase generation I was interested in improving was in generating more unique paraphrases. Quirk et al. (2004) noted that paraphrases generated by their MT model tended to lack variation. The mean edit distance of their generated paraphrases was 2.9 on the top ranked sentences produced by their system; this means that most of the paraphrases were simple word substitutions or short phrase substitutions at best. These sentences generally lacked any phrasal alternation or rearrangement. I use a phrasal chunker to identify syntactic phrases, to hide these phrases behind their labels, and to employ statistical phrase alignment models to discover syntactically interesting patterns of movement in order to help improve the diversity of paraphrases produced. Details of the ParaMeTer system are described in this section.

### *6.1.* ParaMeTer algorithm

The complete training process for ParaMeTer is illustrated in Figure 5. After formatting data so one sentence occurs on each line, the paraphrase training data, test data, and the North American News data were each tagged for part-of-speech and chunked for phrase type. Labels were heuristically chosen for each phrase type given the tags and chunks, using general phrase type labels (e.g. NP for noun phrases and PP for prepositional phrases) for the language model data and the test data. The aligned paraphrase data was labeled using the same heuristics for each phrase type and using different matching schemes to decide which phrase in each sentence aligned with which phrase in the paraphrase sentence. All data was uniformly lowercased after being labeled. The language model was learned for each phrase type separately, using both the North

American News data and the paraphrase training data. MT phrase alignment tables were learned for each phrase type and matching scheme, using the paraphrase training data.

The complete runtime process for ParaMeTer is illustrated in Figure 6. Paraphrase sentences were generated from the test sentences (which were labeled using the same process described during training) by employing the Pharaoh decoder and with the language model and translation model. Words from the original chunks in the test data were substituted for the labels, and finally the best sentence for each test sentence was selected based on the translation score. The process for training and runtime will be described in further detail in subsequent sections.

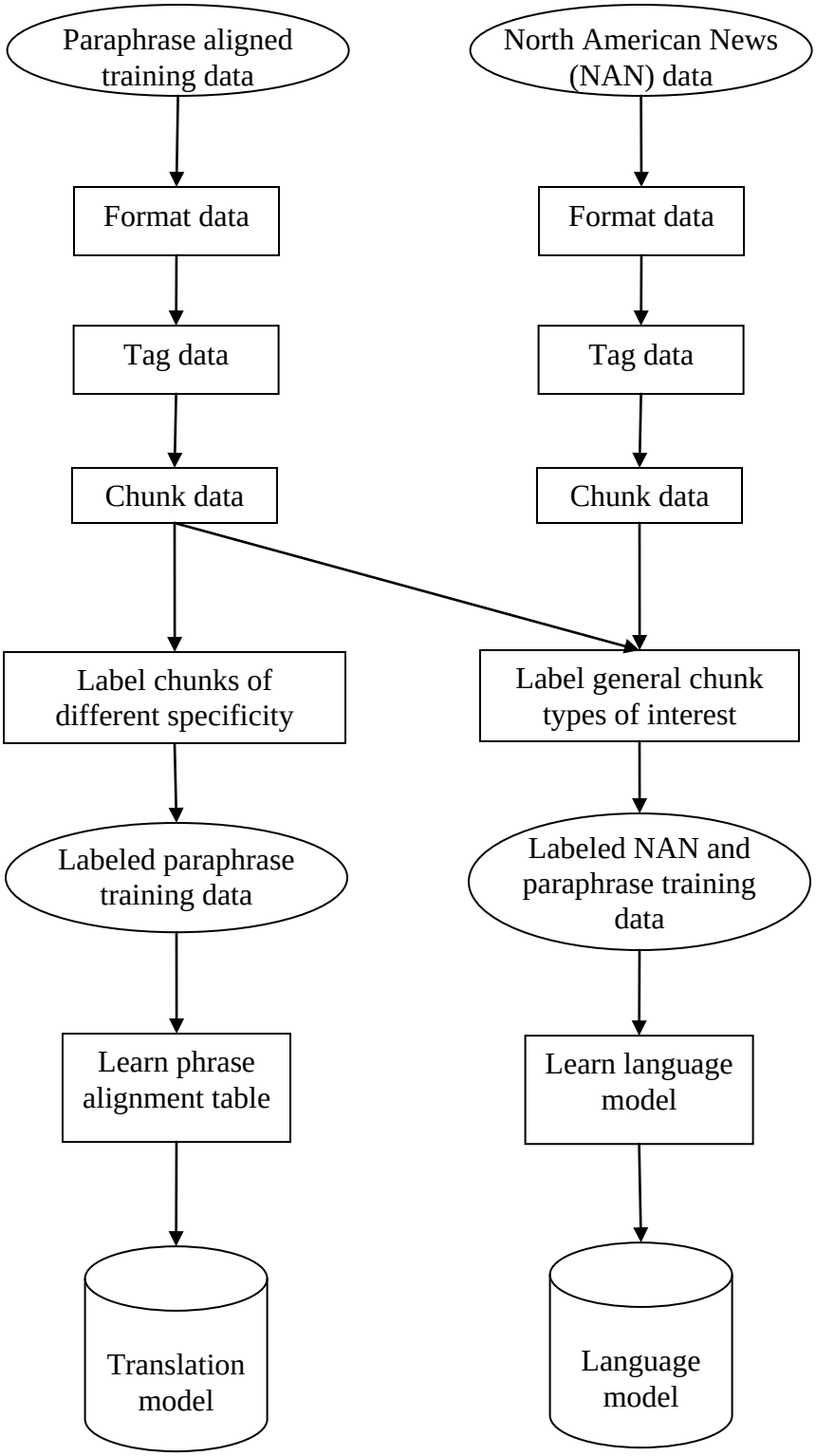


**Figure 5**: The training process for the syntax-based MT system, ParaMeTer.

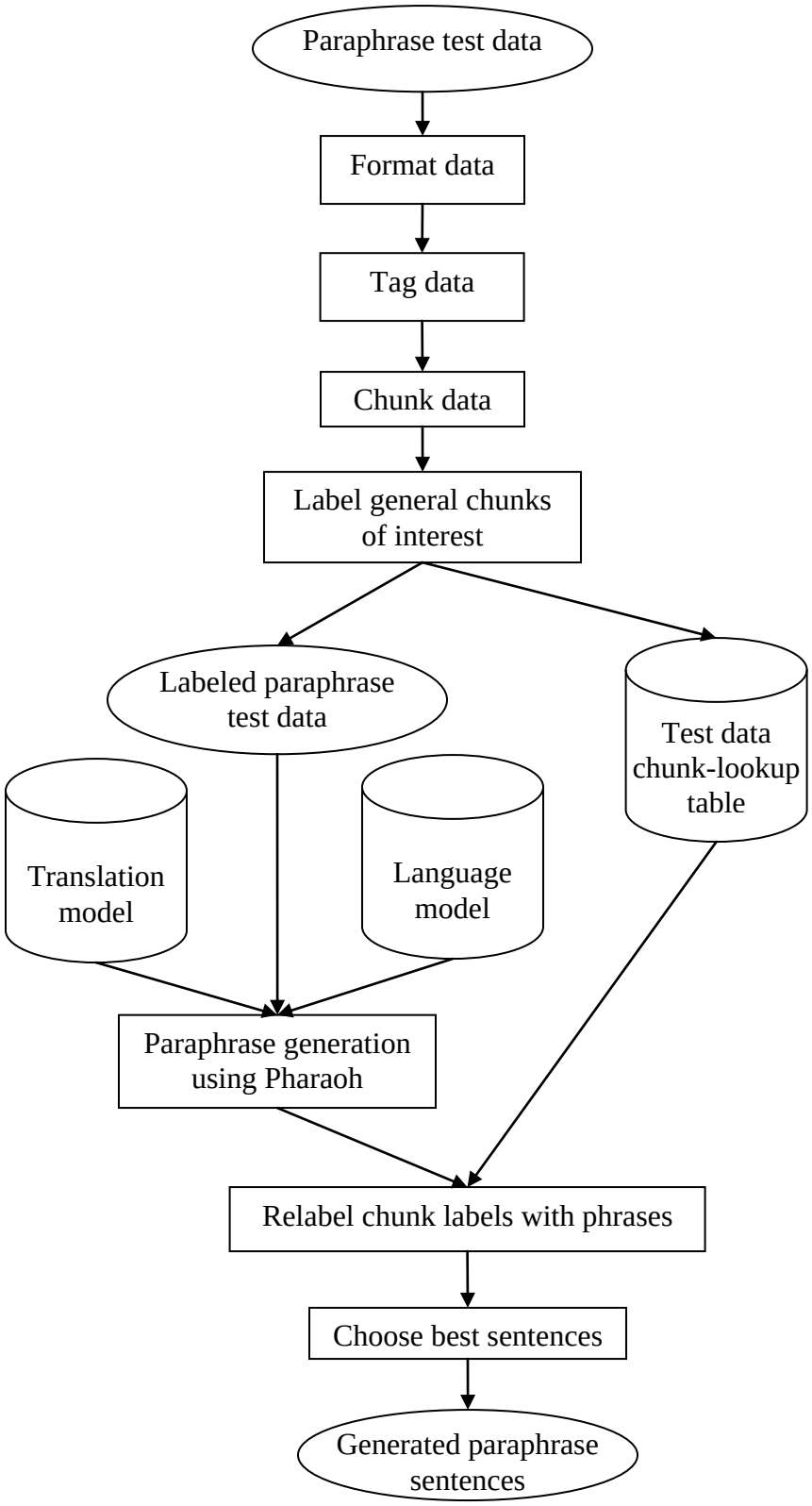


**Figure 6**: The runtime process for the syntax-based MT system, ParaMeTer.

### *6.2.* Part-of-speech tagging

Stanford's part-of-speech tagger[7] (Toutanova and Manning, 2000; Toutanova et al., 2003) was used to first label the words in each sentence with a part-of-speech tag as the first step in the pipeline. The tagger by default tokenizes sentences before tagging. These part-of-speech tags are then used as features in the YamCha chunker[8] (Kudo and Matsumoto, 2000). I used the wsj3t0-18-bidirectional model that is included with Stanford's distribution set, trained from the Wall Street Journal corpus.

### *6.3.* Examples of chunks used

In order to identify patterns of phrasal chunks in the data, an off-the-shelf chunker, YamCha, was used with a model kindly provided by Kudo.[9] A chunker is a shallow parser which attempts to segment words in a sentence into different chunks based on the phrasal categories – noun phrases (NP) or prepositional phrases (PP), for example. Kudo and Matsumoto train support vector machines (SVMs) in a Conditional Markov Model to chunk sentences with their YamCha system. Their system uses features such as part-of-speech tags, the words themselves, surrounding contexts, and neighboring chunk labels to train the SVM to classify the beginning, inside, and outside ("B-I-O") of a chunk. The "outside" label is used to specify that a word belonged to none of the labels of interest.

This method performs well, especially for chunks that occur fairly often (like noun phrases, verb phrases, and prepositional phrases). The example in Figure 7 shows a sentence from the training corpus marked with labels in the described method.

[7] http://nlp.stanford.edu/software/tagger.shtml
[8] http://chasen.org/~taku/software/yamcha/
[9] http://chasen.org/~taku/software/sec/CoNLL.model.zip

```
CHUNK:    B-NP I-NP       B-VP  I-VP      B-PP B-NP I-NP B-PP
POS TAG:  DT   NNS        VBD   VBN       IN   NNP  NNP  IN
WORDS:    The  processors were  announced in   San  Jose at


CHUNK:    B-NP I-NP  I-NP      I-NP  O
POS TAG:  DT   NNP   NNP       NNP   .
WORDS:    the  Intel Developer Forum .
```

**Figure 7**: An example of chunks in a sentence from the training set. The part-of-speech tags are given above each word, and the chunk label is given above each part-of-speech tag.

Notice that the chunker tends to "forget" the higher-level phrases in favor of the more immediate chunk label. In the example above, though "San" and "Jose" are part of the PP "in San Jose," the chunker labels them as the more immediate phrase type, B-NP and I-NP, respectively, instead of I-PP. This necessitates that, after the chunker has labeled the data, ParaMeTer uses specific heuristics to identify the whole phrase; it could not simply identify a prepositional phrase as beginning with a word labeled B-PP and continuing just until a different type of phrase began. A PP using my heuristics is a word labeled B-PP, followed by an NP, verb phrase, adverbial phrase, adjective phrase, or PP, but only including one NP. An NP is a word labeled B-NP followed by anything labeled I-NP. A complement clause (CC) is a word labeled B-SBAR, followed by all words until one is labeled an O chunk. A relative clause (RC) is a word labeled with part-of-speech tag WDT, followed by all words until one is labeled an O chunk. Of course, these heuristics are not perfect and rely heavily on the POS tags and chunk labels to be correct; a particle (e.g. kick things *off* on Monday) is only distinguished from a preposition (e.g. live *off* Main Street) if the POS tags and chunk labels are correct.

*6.4.* Integration with MT model

Providing YamCha with a pre-trained model provided by Kudo (trained on other data already labeled with chunk tags, such as the Wall Street Journal corpus) and with part-of-speech tags for the paraphrase sentence pairs (labeled using the Stanford part-of-speech tagger), chunks in the data were identified and labeled. ParaMeTer then substitutes the labels for the chunks of interest in the sentences.

These preprocessed sentences were then aligned by statistical phrase translation methods – using the chunk labels in place of the subsumed word sequences and allowing the labels to masquerade as words – in order to pinpoint common movements of syntactic phrases. For example, PPs often can move from the beginning of a sentence to the end in creating a paraphrase sentence. Figure 8 demonstrates how two PPs ("in San Jose" and "at the Intel Developer Forum") switch positions in these two paraphrases from the training corpus. Learning these types of syntactic movements allowed the MT system then to create more unique paraphrase sentences, since abstract labels helped the system to generalize better.

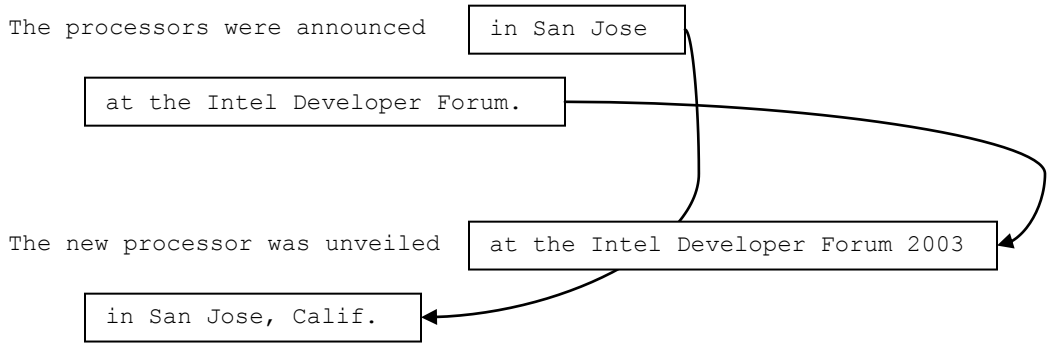


**Figure 8**: An example of PP movement in a pair of paraphrase sentences.

To identify this movement, different types of matching schemes (discussed in section 6.5) were used to learn which PP matches with which. At least one of the PP translation models learned this very movement: that "in san jose pp" aligns with "pp in san jose."

The chunks of interest are any types of chunks for which we expect to see movement. The chunks that I experimented with were:

- noun phrases
- prepositional phrases
- relative clauses
- complement clauses

Using labels for these phrases allowed the translation model to represent how phrases move in general, and the language model to represent how these phrase types occur in context in English. As chunks were identified and labeled, a lookup table was kept listing the original words from the phrase that each label replaced. Then, after generating a paraphrase for the test sentence, the system substituted the original words again in place of the label.

Since more than one phrase of the same type received the same label, only one phrase could be substituted in a given sentence at the same time. Otherwise, after runtime the system would not know which label to replace with which phrase (since ideally the phrases would have shifted around in some way). This means that for each phrase in the sentence, I duplicated the sentence so that each phrase appeared as the only phrase labeled in a single copy of the sentence. One sentence was also duplicated without any

phrases labeled so the system could leverage information about alignments without the labels as well.

Duplicating the sentence pairs for each matching phrase likely affected the statistics for some word/phrase combinations in the language model and the translation model. However, this did not seem to negatively affect performance of the system, and in fact may have contributed to the syntax-based system achieving more fluent paraphrases than the baseline system. This approach did allow alignments for all possible phrases in the dataset to be learned, rather than just a subset. I discuss an alternative way for accomplishing the same goal in the Future Work section.

### *6.5.* Chunk label matching methods

As each sentence only has one phrase labeled at once, the aligned data used for training the translation model poses a question. How should the system determine which phrase in one sentence should be associated with which phrase in the other sentence? I wanted the system to learn how phrases move, but if it labeled the wrong phrase as a match in the aligned sentence, it would learn inaccurate movement patterns.

I experimented with several different types of matching schemes for each of the four different types of phrases to determine the best approach for identifying matching phrases, so the phrases in the parallel sentences would correspond. Matching schemes, from most general to most specific, included:

- same general phrase types
- identical left-most word
- identical right-most word
- lowest edit distance score
- best score from word-alignment tables
- exact phrase match

The most general method, *same-general-phrase-types*, just considered whether the phrase in one sentence was the same type as the phrase in another sentence. For example: if both phrases were PPs, they were valid candidates as matching. Both sentences in the Figure 9 have three PPs identified by the chunking method: "on the internet," "on June 10," and "for sale." In the *same-general-phrase-types* method, all three of these are potential matches with all three of the PPs in the other sentence.

```
They had published an advertisement [ on the internet ] [
on June 10 ], offering the cargo [ for sale ], he added.

[ On June 10 ], the ship's owners had published an
advertisement [ on the internet ], offering the explosives
[ for sale ].
```

**Figure 9**: Two sentences from the training set. PPs are marked within square brackets.

The *identical-left-most-word* scheme produced matches if, for example, the two PPs had the same preposition. In Figure 9, "on the internet" could match with either "on the internet" or with "on June 10" in the second sentence. The *identical-right-most-word* scheme validated a match if the two NPs, for instance, had the same right-most word

(often this would be the head-word of an NP, but not always). In the above example, “an advertisement” would match “an advertisement,” but “the cargo” and “the explosives” would fail to match.

The *lowest-edit-distance-score* method calculated the minimum edit distance for each phrase of the same type, and the lowest score (below a certain threshold) was considered the best match. This allows PPs like “of its million-plus domestic workers” to match “of its domestic employees”, since the edit distance between the two phrases is low (one substitution at a cost of 2 and one deletion at a cost of 1 is an edit distance of 3). The most specific method, *exact-phrase-match*, only allowed for matches to a phrase in the second sentence of the same type if it contained exactly the same words.

The *best-score-from-word-alignment-tables* scheme was between the *lowest-edit-distance-score* and *exact-phrase-match* in generality. It was more generous in allowing a match between two phrases than *exact-phrase-match* since phrases did not need to contain exactly the same words in the exact order; but it was less generous than *lowest-edit-distance-score* since all the words had to have some alignment score with the other phrase to justify a match. The *best-score-from-word-alignment-tables* scheme used Giza++ word-alignment tables trained using IBM Model 1 to find the strength of association between phrases, as justified by the scores for possible word alignments in the two phrases. Phrases longer than 9 words could not be matched using this scheme, as enumerating all possible alignments between two phrases to calculate the score was a problem of exponential complexity. The probability of a source phrase given a target phrase can be determined by adding the probability of the source (*s*) and an alignment (*a*) given the target (*t*), over all possible alignments:

$$(7)\ P(s \mid t) = \sum_{a} P(s, a \mid t)$$

The right-hand side of Equation 7 can further be expressed as the following:

$$(8)\ \sum_{a} P(s, a \mid t) = \sum_{a} \left( \prod_{x, y \in a} P(x \mid y) \right)$$

In Equation 8, *x* and *y* are words in the source and target phrases, respectively. For each word *x* in the source phrase, the probability of *x* given a word *y* in the target phrase was determined by looking up the value in the word-alignment table. This allowed relative clauses like "that expires in August" to match "which expires in August," but rejected matches with a zero probability that edit distance would have allowed like "that affects millions around the world" and "that affects millions of Americans." Two phrases would have a zero probability if even one of the words in the phrases had a zero probability of aligning with each of the words in the other phrase. The edit distance score, on the other hand, was not nearly as picky about which words it allowed to substitute for each other. A smoothing scheme for translation tables could be used (Foster et al., 2006) to avoid having a zero probability when using the word-alignment table matching method.

Since the matching schemes could elicit multiple possible matches for each phrase, I used greedy methods to choose only the "best" match for each scheme instead of opting to duplicate sentences further for each possible match (and having to concoct a way to mix the score for each match with the language model and translation model score

at the end to choose the best paraphrase sentence). The *lowest-edit-distance-score* and *best-score-from-word-alignment-tables* methods used the best score as the deciding factor for the best match. Phrases were only allowed to be used in a match once, though, so phrases were matched using the best score on unused phrases. The other four schemes used the first (leftmost) qualifying match that had not been used already in a match with a previous phrase in the sentence. In general the greedy choice was correct, since the edit distance between sentence pairs in the corpus was limited and sentences were not structured too differently. Though in some cases the greedy match will align phrases incorrectly, the benefit that sentences were duplicated once for each phrase in the sentence, but only once, meant that proliferation of incorrect matches did not get out of hand.

### *6.6.* Selecting the best models

As the language models did not depend on matching schemes, there were only four different language models – one for each phrase type. The NP language model then was used in conjunction with each of the six NP translation models (one translation model for each matching scheme). The trends in perplexity for the language models seen with the baseline system held when building a language model on chunked data, though perplexity tended to be higher. Higher perplexity scores likely arose because more words can follow a label like “NP” or “PP” than could follow any individual word in those phrases. Perplexity on the MSR data when building the noun phrase language model was 250.388, but mixing it with extra data brought perplexity down to 148.220. The prepositional phrase language model went from a perplexity of 163.838 to 147.036; the relative clause

language model went from 218.670 to 205.758; and the complement clause language model went from 213.913 to 202.312.

Given four different types of phrases, and six different phrase matching schemes for each method, I created twenty-four different translation models. To select the best possible translation model for each phrase type, I ran the development sentences through the decoder using each model. Pharaoh returns a translation score (the log probability) for each sentence, which is a combination of the score from the translation model and the score from the language model for the sentence generated. This does not ensure that a sentence with a lower score is more fluent or a better translation; that is left to optimization of the models themselves. But the score can serve as a guide, where a lower log probability score is preferred by the decoder. Table 1 shows the translation scores for each phrase type for the relevant matching schemes. The matching schemes without a score were not kept because the results were uninteresting.

| | **CC** | **NP** | **PP** | **RC** |
|---|---|---|---|---|
| *Same general phrase types* | x | x | x | x |
| *Identical left-most word* | -40.1799 | x | -28.4973 | -50.3786 |
| *Identical right-most word* | -45.3415 | **-24.3185** | -27.8164 | -40.5922 |
| *Lowest edit distance score* | -50.8906 | -33.7741 | -42.0871 | **-50.3786** |
| *Best score from word-alignment tables* | -41.9795 | -36.1317 | **-28.4973** | -46.5379 |
| *Exact phrase match* | **-36.2129** | x | -28.1469 | -49.6944 |

**Table 1**: Translation scores for the development set for each phrase type and matching scheme. The x's in the table indicate results that were not kept.

I compared the output from using the six different models for each phrase type, considering both the translation score Pharaoh calculated and how fluent the sentences

read. This guided me in choosing the best matching scheme to use for each phrase type. The best translation model for NPs seemed to be using the *identical-right-most-word* scheme. Intuitively this makes sense, as in general the right-most word in a noun phrase is the head word. So intuitively the best matches would come from assuming noun phrases with the same head word are matches. This choice was clear in both the readability of sentences and in the translation scores.

The matching scheme that seemed to produce the best results (lowest translation scores and most readable sentences generated) for PPs was the *best-score-from-word-alignment-tables* method. This was a harder choice, as the scores for the left-most word (generally the preposition for PPs), the same right-most word (generally matching noun phrases in the PPs), and exact phrase match were all very close. However, the Giza++ word-alignment tables seemed to give better sentences for PPs.

```
(A) he allowed two runs in seven innings and struck out six .

(B) he allowed two runs in seven innings and struck out six .

(C) he allowed two runs , struck out six in seven innings .
```

**Figure 10**: Sentence results from the development set. Sentence A is the input sentence, sentence B is the identical right-most word output, and sentence C is the best score from word-alignment tables output.

In Figure 10, though the *identical-right-most-word* gave a good sentence, it was exactly the same as the input. The *best-score-from-word-alignment-tables* at least gave some variation to the input, which is what I was hoping for ultimately – more varied sentences.

RCs made another difficult choice. Though the translation score was the worst, they had more readable sentences when using the *lowest-edit-distance-score* matching scheme, as shown in Figure 11, sentence C.

```
(A) ballmer has been vocal in the past warning that linux
is a threat to microsoft .

(B) in the past warning of a threat to the ballmer has been
vocal ' linux is , microsoft said .

(C) ballmer had been vocal in the past , warning linux is
threat to microsoft .
```

**Figure 11**: Another set of sentence results from the development set. Sentence A is the input sentence, sentence B is the identical right-most word output, and sentence C is the lowest edit distance output.

In further examining the look up table of matched phrases, the lowest edit distance scheme allowed relative clauses with some variation to match, without over-matching as the more general schemes did, and under-matching, as the *exact-phrase-match* and *best-score-from-word-alignment-tables* methods did.

Another surprising result was that CCs performed best using the *exact-phrase-match* scheme. I would have assumed that, since complement clauses tend to be longer, exact-matches would be less common and thus under-match; and indeed in studying the CC phrases that were matched and the ones that were not, this is the case. However, perhaps because of the way the training data was constructed, learning CC movement does not help in translation. CCs do not move positions in the training sentences; this likely would make the edit distance large enough to have eliminated a sentence pair from the MSR paraphrase corpus. Instead, CCs in the data seem to be altered somewhat rather than moved. In Figure 12, the CCs in both sentences (“that the kidnappers fiercely

resisted the army assault this morning, firing Kalashnikov rifles" and "that the kidnappers put up fierce resistance during the army assault, firing Kalashnikov rifles") appear in the same position, just with new information or slight variations in wording.

```
El Watan, an Algerian newspaper, reported that the
kidnappers fiercely resisted the army assault this morning,
firing Kalashnikov rifles.

El Watan, an Algerian newspaper, reported that the
kidnappers put up fierce resistance during the army
assault, firing Kalashnikov rifles.
```

**Figure 12**: Two sentences from the training set.

I imagine the edit distance involved in moving a phrase that long would easily overshoot the restrictions involved in constructing the corpus originally.

After selecting the best translation model for each of the four phrase types, ParaMeTer used a script to select the best scoring translation among the four phrase-types' models for each sentence in the development data. This produces one "best" translation for each test sentence. This means that each "best" translation ultimately came from the test sentence being substituted with one phrase label of one of the four types. While this approach makes substituting the original words in the phrase after the paraphrase sentence has been generated an unambiguous problem, it severely limited the amount of phrasal movement ParaMeTer could model in any given sentence. An approach to overcome the ambiguity of multiple phrases in the same sentences is presented in the Future Works section of this document, which may improve the amount phrasal movement that can be modeled.

*6.7.* Experimenting with alternative paraphrase data

Since 3900 pairs of sentences pale in comparison to the 139,000 pairs of sentences used by Quirk et al. (2004), I experimented with supplementing the translation model with additional data, just as I supplemented the language model. The additional data was derived from multiple human translations in English of the same Chinese and Arabic sentences in LDC corpora[10]. Since the human translations mean the same thing, they can be considered paraphrases of each other. Over 27,000 additional paraphrase sentences were extracted in this way.

Without functionality in the tools for statistical phrase training to interpolate between translation models, I trained a model using both parallel corpora concatenated together. The MSR paraphrase corpus is limited to an edit distance of 12 and sentences cannot be less than 2/3 the length of the other. The human translations had no such restrictions and overall had different style choices as well (e.g. some translators used colons before a quotation instead of quotation marks around the statement). Given that the two datasets are quite different, concatenating the two corpora together hardly seems like an adequate approach. Alternatives to make better use of the translation sentences are discussed in the Future Works section.

In fact, not only did the results from this experiment produce sentences with much worse translation scores, the results were much less readable in general. Compare the average translation score for *identical-right-most-word* NPs. The MSR Paraphrase corpus alone got an average score of -24.3185; with the MT data, the best average was still a dismal -36.1317. For the best CC matching scheme, the MSR Paraphrase corpus alone

[10] Multiple Translation Arabic, Parts 1-2; and Multiple Translation Chinese, Parts 1-4: http://www.ldc.upenn.edu/Catalog/byType.jsp

got an average score of -36.2129; with the MT data, the best average was -40.1799. Aligning words that have moved much further (by relaxing the limit on edit distance) seems to be a more difficult task. Perhaps having the paraphrases so different complicates the paraphrase generation task nearer to the level of difficulty involved in translating to a different language. Attempts to interpolate human translation paraphrases into the language model only worsened the language model perplexity and (by association) translation scores.

For the purposes of this thesis, final results were produced without the human translation data. It would be interesting in future work to attempt other ways of including this data so the results are not limited to simple structural paraphrases like that in the training corpus. Perhaps normalization techniques of style and of Chinese and Arabic words, or a method of interpolation between translation tables, or some other technique would produce more fluent paraphrases.

## 7. Comparison with Baseline

Evaluation of a paraphrase generation system is a difficult problem. Automatic techniques for paraphrase identification are not yet confident enough to be of use. Employing human evaluators is a costly venture (whether in time or money). However, if we consider a subset of the test data, this approach is plausible. Mean edit distance intuitively should be helpful to consider as well.

*7.1.* Human judgments

Following in the pattern set by Quirk et al. (2004), one level of success is human acceptability judgments. For this thesis, forty judges each evaluated 30 pairs of sentences, randomly selected from the 200 test sentences generated by the syntax-based system and from the 200 generated by the baseline (statistical MT) system. Each sentence pair was judged by three human judges (so that a tie-breaker was built in). The judges were presented with two sentences at a time: the first sentence was always the original test sentence used as input to the systems, and the second was one of the two systems' generated paraphrase. Six questions followed each pair in order to determine whether the sentences generated by the system were acceptable paraphrases – whether they were roughly semantically equivalent – and to gauge variability of the sentences:

- Were you able to understand the meaning of sentence B?
- Do these two sentences mean the same thing?
- Does sentence B have phrases that have moved from the order in sentence A?
- Does sentence B have phrases that are reordered internally, or substituted with different, equivalent phrases?
- Was there information lost in sentence B?
- Was there information gained in sentence B?

The results from the human evaluations are summarized in Tables 2-4.

The first question helps enlighten us about how readable, or fluent, the sentences were. I use the term *fluent* somewhat loosely, to indicate that the human judges could understand the meaning of the sentence. Not surprisingly, with such an impoverished translation model both systems have some trouble producing sentences that make sense. It is interesting that ParaMeTer, trained on the same amount of data, produces more than twice as many fluent sentences.

| | **Baseline (SMT-only)** | **ParaMeTer (Syntax-based)** |
|---|---|---|
| *Fluent* | 53 / 201 = 26.37% | 119 / 201 = 59.20% |
| *Paraphrase* | 27 / 201 = 14.14% | 60 / 201 = 29.85% |
| *Rearrangement* | 161 / 201 = 84.29% | 167 / 201 = 83.08% |
| *Phrasal alternation* | 140 / 201 = 73.30% | 63 / 201 = 31.34% |
| *Info lost* | 145 / 201 = 75.9% | 125 / 201 = 62.19% |
| *Info added* | 75 / 201 = 39.26% | 41 / 201 = 20.40% |

**Table 2**: Human judgment results.

Whatever the reason for this improved fluency – because of the data duplication, or because the syntax-based approach had shorter sentences to generate (replacing a multi-word chunk with one label shortens the sentences), or because the syntactic approach is able to generalize better and thus more robustly – more experiments would help determine why I see this effect. Since the other raw counts in Table 2 showing sentence variability are likely the result of word scrambling from non-fluent sentences, Table 3 isolates the analysis to those sentences which are fluent, and Table 4 looks more specifically at variability in those sentences which are fluent paraphrases.

The second question in the survey determined how many paraphrases were generated. Table 3 shows more precisely, among all the sentences that were fluent, how many were good paraphrases of the input sentence.

| | **Baseline** | **ParaMeTer** |
|---|---|---|
| *Fluent paraphrases* | 23 | 55 |
| *Fluent* | 53 | 119 |
| *Percent* | 43.40% | 46.22% |

**Table 3**: Percent of paraphrases among fluent sentences.

As hoped, ParaMeTer seems to produce approximately the same percentage of paraphrases as the baseline. The baseline produces just under 45 percent fluent paraphrases, and the syntax-based system just over 45 percent.

The next table, Table 4, examines the variability among acceptable paraphrases. The last four questions were intended to capture qualitative measures, including: rearrangement, phrasal alternation, information lost, and information added (Quirk et al., 2004).

| | **Baseline** | **ParaMeTer** |
|---|---|---|
| *Rearrangement* | 7 / 23 = 30.43% | 34 / 55 = 61.18% |
| *Phrasal alternation* | 12 / 23 = 52.17% | 17 / 55 = 30.91% |
| *Info lost* | 3 / 23 = 13.04% | 11 / 55 = 20.00% |
| *Info gained* | 7 / 23 = 30.43% | 5 / 55 = 9.09% |

**Table 4**: Variability among acceptable paraphrase sentences.

The baseline, which was intended to reproduce the approach described by Quirk et al., exhibits much more alternation and variability than Quirk et al. reported of their system. Table 5 compares the percentages of variability among acceptable paraphrase sentences to those reported by Quirk et al.

| | **Baseline** | **ParaMeTer** | **Quirk et al.** |
|---|---|---|---|
| *Fluent paraphrase* | 43.40% | 46.22% | 89.5% |
| *Rearrangement* | 30.43% | 61.18% | 0% |
| *Phrasal alternation* | 52.17% | 30.91% | 3% |
| *Info lost* | 13.04% | 20.00% | 31% |
| *Info gained* | 30.43% | 9.09% | 6% |

**Table 5**: Variability among acceptable paraphrase sentences, comparing also to Quirk et al.

This discrepancy of variability results between Quirk et al. and the baseline could be due to a couple of reasons. Either the human judges in this thesis were not primed in the same manner and thus more liberal in assessing how different two sentences were, or the sparsity of data for the translation model produces more word scrambling in general, in both acceptable and unacceptable paraphrases. Whatever the reason, the differences between the human judgments on the baseline and the syntax-based approach remain noteworthy.

As shown in Tables 4 and 5, the syntax-based system has over 60% of the fluent paraphrases demonstrating movement of phrases, while the baseline system has just over 30%. An example from the development results illustrates this type of movement. In

Figure 13, Sentence A is the input sentence, and sentence B is the paraphrase sentence generated by ParaMeTer. Sentence C is the sentence generated by the baseline.

As noted previously, the within-phrase alternation results are not surprising, either, since the chunk-based system merely replaced whole chunks with the original words after translation. A translation model or some other solution for replacing chunks with something other than the original words would need to be implemented to see more improvement on phrasal alternation. This is a matter taken up in the section on Future Work.

```
(A) west nile virus -- which is spread through infected
mosquitoes -- is potentially fatal .

(B) west nile virus -- which is potentially fatal -- is
spread through infected mosquitoes .

(C) the west nile virus , which is spread through the
mosquitoes , is potentially fatal changes -- infected --
```

**Figure 13**: Example of results from the development set. Sentence A is the input sentence, sentence B is the syntax-based system's output, and sentence C comes from the SMT-only system.

The difference between information lost in the two systems is less noteworthy, though both ParaMeTer and the baseline lost less information than Quirk et al., which is an improvement (see Table 5). Information gained is almost as low in ParaMeTer as Quirk et al.'s system; the information gained in the baseline seems vastly greater, but this might be more due to the translation model than anything else. At least one judge noted the insertion of random words in sentences that later proved to be from the SMT-only system; the judge chose to reflect this in the evaluation by marking "information gained." For instance:

```
(A) worldwide , 7,183 sars cases and 514 deaths have been
reported in 30 countries .

(B) 7,183 sars cases have been reported in 30 countries
worldwide , and 514 deaths .

(C) when 7,183 and to have been reported cases of severe
514 slayings of the 14 lithuania gained worldwide
```

**Figure 14**: Another example of results from the development set. As before, A = input, B = syntax-based system output, and C = SMT-only system output.

In the above example, Figure 14, sentence C is again the SMT-only system's result. Information such as "14 lithuania gained" and that the deaths/slaying were "severe" has been added to the resulting sentence.

### *7.2.* Mean edit distance

The mean edit distance was used in originally building the paraphrase corpus, so it was hoped that this measure would produce interesting commentary on the success of the system. Since the goal of using syntactic chunks in the paraphrase generation system was to create more interesting, more novel constructions, I hoped for higher edit distance scores. A higher edit distance measure in the paraphrases would indicate more variation in general between the sentences. The human judgments would ensure that this variability was not achieved by sacrificing readability and accuracy in generating true paraphrases.

However, the actual edit distance averages were not as fruitful as hoped. Though ParaMeTer scores higher than the 2.9 Quirk et al. (2004) reported of their system, so does the SMT-only baseline implementation meant to duplicate their approach. The chunk-based approach has an average of 5.64 edit distance among the 55 fluent paraphrases,

while the baseline has an average of 6.17 edit distance among the 23 fluent paraphrases it produced.

This increase in edit distance is perhaps related to the reasons for general inflation of sentence variability in both the baseline and syntactic system (see Table 5) from the numbers reported by Quirk et al. (2004). The higher edit distance may be because the human judges in this thesis were more liberal in declaring a sentence "fluent" or because the sparsity of data licensed more word scrambling in both the baseline and syntactic systems. Whichever the cause, the higher edit distance seems to give support to the higher human judgments on variability found in this thesis. Perhaps the word-scrambling effect better explains why, though the mean edit distance is greater for the baseline, greater syntactic variability (as judged by humans) was found in the syntactic system.

## 8. Future Work

Further experiments will consider basic methods to improve results from both the baseline and the syntax-based system, including: removing punctuation, reinserting capitalization, replacing numbers and proper nouns with more general labels (e.g. "<year>" or "<dollar_amount>"), and optimizing parameters throughout the paraphrase generation process. These efforts should improve understandability and generalization for both systems. A multivariate optimization procedure – such as Powell's algorithm (Press et al., 1988) – could optimize the parameters in the pipeline; a possible metric against which to optimize might combine edit distance, some measure of paraphrase recognition, and grammatical correctness.

Experimentation with other methods of including the multiple reference translations as paraphrase data will be explored to see what gains may be possible from using more data for the translation model. Normalization of style (such as colons versus quotation marks to offset quotes in an article) and of Chinese and Arabic words would likely bring the data closer in similarity to the MSR paraphrase data. Edit distance might even be used to ensure the corpora match better (the MSR paraphrase data was restricted to sentence pairs between 8.0 and 12.0 edit distance). A method of interpolation between the MSR paraphrase data's translation table and the Chinese/Arabic translation data's translation table might help to give more weight to the data that best resembles the test set; perhaps concatenating copies of the data in proportion to the interpolation constant instead of in the original proportions would be a method of simulating interpolation. If I instead move away from the MSR paraphrase data as a test set, interpolation and edit distance restrictions, of course, would not be necessary. Experiments using other data alone, without combining with the MSR paraphrase data, are worth consideration; the multiple reference translation corpora or various translations of the Bible could provide a larger corpus of paraphrases than is possible with the MSR paraphrase data. In fact, experiments with another type of paraphrase corpus might enlighten the effects of restricting paraphrases using edit distance.

Investigations specific to the syntactic approach are also useful to consider. One experiment would be to include building a phrasal-only translation model to allow for reordering within phrases and phrasal substitutes. This experiment would hopefully improve the within-phrase alternation score for the syntactic system, so a translation model trained just from the phrase matches could learn possible phrase alternations.

Then, words other than the original words in a given phrase could be substituted for the phrase labels after the paraphrase sentences have been generated.

More experimentation could be done with the proliferation of matches to determine whether a greedy/best-first matching scheme is better than merely duplicating sentences further for all possible matches. I could also experiment with duplicating the baseline data, to help isolate the effect of the data duplication on the syntactic system. This could help answer the question of why the syntax-based system is able to generate more understandable sentences.

Using detailed labels for different types of phrases based on the matching scheme may help the translation models generalize more accurately. Instead of labeling all PPs the same, PPs might be labeled PP-with or PP-for in the *identical-left-most-word* matching scheme; prepositional phrases or relative clauses that begin with different words may move and behave differently. This would involve modifying how Pharaoh accesses the language model, however, so the language model may learn where phrases may appear without learning the detailed information. One obvious advantage to this approach would be that it would eliminate, or greatly reduce, the need to artificially duplicate data. The detailed information would make it less common that two phrases with the same label would occur in the same sentence, and if enough detail is preserved, I could ensure that different phrases were always labeled differently. Being able to label phrases differently means that multiple phrases can be labeled in each sentence, removing existing restrictions on modeling variability.

Finally, there may be other possibilities for choosing the best translation for a sentence at runtime. To some extent, the need for merging (or selecting among) multiple

translations may disappear as I experiment with detailed labels, eliminating the duplication of sentences and the resultant multiple translations for each test sentence. However, it may still make sense to keep the models for different phrase types separated, and thus still have 4 sentences to choose from.

## 9. Conclusions

This work demonstrates a feasible method to not just generate paraphrase sentences, but to generate more unique structures by taking into account syntactic information and generalizing how to move phrases. It also suggests possible benefits to leveraging syntax in MT applications by demonstrating an application which improves when using syntactic chunks.

Only 59.20% of the test sentences from the syntax-based system were found to be understandable, however this is an improvement over the 26.37% understandable sentences the baseline generated. Both systems generated about the same rate (approximately 45%) of valid paraphrases when considering just the fluent sentences.

Overall results are promising, clearly demonstrating increased rearrangement by leveraging syntactic chunk information. Of the 55 fluent paraphrases from ParaMeTer, 61.18% were judged to have phrasal rearrangement. Only 30.43% of the baseline's 23 fluent paraphrases show phrasal rearrangement. This shows that the syntax-based approach is able to better generalize in order to produce more fluent sentences with more variation in grammatical structure.

The baseline, however, has more within-phrase alternation, with 52.17% of the 23 sentences. ParaMeTer has only 30.91% of the 55 sentences showing phrase alternation.

This disparity would most likely be remedied with a translation model for the phrases themselves, so the system does not necessarily substitute the original words for the phrase labels after paraphrases were generated.

It would be interesting to see if the results found here carry over to models with more translation data at their disposal, or models trained from data with greater variability like the English-Chinese and English-Arabic multiple human translation data. Given the results from these experiments, however, it seems evident that the syntactic approach presented successfully gains sentence variability without losing the meaning behind the sentences.

**Bibliography**


Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, Ann Arbor, MI, pages 597-604.

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, Canada, pages 16-23.

Regina Barzilay and Kathleen McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL)*, Toulouse, France, pages 50-57.

Samuel Bayer, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh. 2005. MITRE's Submissions to the EU Pascal RTE Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, U.K.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(*2*):263-311.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 17-24.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, pages 263-270.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, U.K.

William B. Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland, pages 350-356.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: the impact of paraphrasing for Question Answering. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC.

Florence Duclaye, François Yvon, and Olivier Collin. 2003. Learning paraphrases to improve a question-answering system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) Workshop on Natural Language Processing for Question-Answering*, pages 35-41.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, July 2006, to appear.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 455-462.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 115–124.

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, pages 311-318.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, Canada, pages 48-54.

Okan Kolak, William Byrne, and Philip Resnik. 2003. A Generative Probabilistic OCR Model for NLP Applications. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, Canada, pages 55-62.

Taku Kudo and Yuji Matsumoto. 2000. Use of Support Vector Learning for Chunk Identification. In *Proceedings of the 4th Conference on CoNLL-2000 and LLL-2000*, Lisbon, Portugal, pages 142-144.

William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical Recipes in C.* Cambridge University Press, Cambridge.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(*1*):19-51.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based Alignment of Multiple Translations: Extracting Paraphrases and Generating New Sentences. In *Proceedings of the 2003 Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL)*, Edmonton, Canada, pages 102-109.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona Spain, pages 142-149.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, pages 271-279.

Stefan Riezler and John T. Maxwell III. 2006. Grammatical Machine Translation. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 248-255.

Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a Question Answering system. In *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP)*, pages 25-32.

Eric K. Ringger. 2000. Correcting Speech Recognition Errors. *University of Rochester Computer Science Department Technical Report TR-731 and Ph.D. Dissertation*.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of the 2nd International Conference on Human Language Technology (HLT)*, San Diego, CA, pages 40-46.

Rion Snow, Lucy Vanderwende, and Arul Menezes. 2006. Effectively Using Syntax for Recognizing False Entailment. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 33-40.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, pages 901-904.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, Hong Kong, pages 63-70.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, Canada, pages 252-259.

Hao Zhang, Daniel Gildea, Liang Huang, and Kevin Knight. 2006. Synchronous Binarization for Machine Translation. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 256-263.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using Paraphrases to Evaluate Summaries Automatically. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, NYC, pages 447-454.

**Appendix A** – Subset of paraphrase results with judgments

A subset (the first 56 sentence groups) of paraphrase results from the test set, with average human judgment included.

(A0) both studies are published in the journal of the american medical association .
(B0) the studies are published in the journal of the american medical association .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C0) the studies are being published today in the journal of the american medical knowledge gained
*Fluent=**y**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**n**, Lost=**n**, Gain=**y***

(A1) hilsenrath and klarman each were indicted on three counts of securities fraud .
(B1) hilsenrath and klarman were indicted on three counts each of securities fraud .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C1) hilsenrath and to klarman both , were indicted on three counts of securities fraud .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A2) from broadway comedies like `` the seven year itch '' (1952) , `` will success spoil rock hunter ? ''
(B2) `` the seven year itch '' (1952) , `` will success spoil rock hunter ? '' from broadway comedies like
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C2) will success spoil rock hunter ? broadway comedies like the `` the seven year , itch too (1952) had received a romance between
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A3) mcgill also detailed the hole that had been cut in the caprice 's trunk .
(B3) mcgill , detailed in a hole sold .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C3) mcgill , detailed in the hole was cut the caprice of the changes trunk
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A4) today , he will find out whether he is still in the running to buy back the u.s. entertainment business of troubled vivendi universal sa .
(B4) today , he will find out whether he is still in the running to buy back the u.s. entertainment business of vivendi universal troubled sa .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***
(C4) `` he will find that out , he said he is still running to buy back to the u.s. entertainment enterprise in the troubled vivendi universal sa gained whether the
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A5) dixon 's win moved him into second place in the points standings , 49 behind kanaan .
(B5) dixon 's win moved into second place behind him in the points standings , 2 kanaan .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(C5) has moved him into the second place in the standings ) gained 2 points behind kanaan scott get
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A6) tennessee titans quarterback steve mcnair apologized thursday for his arrest hours earlier in nashville on suspicion of drunken driving and illegal possession of a handgun .
(B6) apologized for his arrest on suspicion of drunken driving illegal possession of a handgun and tennessee titans quarterback steve mcnair thursday in nashville hours earlier .
*Fluent=**n**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C6) tennessee quarterback steve mcnair apologized for the arrest hours earlier in the drunken driving and illegal possession of a handgun gained titans today nashville on suspicion of
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(A7) the two democrats on the five-member fcc panel held a news conference to sway opinion against powell .
(B7) the two democrats held a news conference to sway opinion against powell on the five-member fcc panel .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C7) the two democrats on the five-member he called a news conference being held sway opinion against powell .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A8) a senior whitehall official said : `` it devalued the currency , there is no question about that .
(B8) a senior whitehall official , a senior whitehall official , said : `` there is no question about that , it devalued the currency .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**y***
(C8) a government official replied that `` it 's devalued currency said there was no question about the changes in the whitehall
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A9) she was taken by ambulance to charing cross hospital in hammersmith .
(B9) she was taken to charing cross hospital by ambulance in hammersmith .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C9) she was taken by ambulance to the charing cross hospital in hammersmith gained
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(A10) the three grocery chains were relying on store managers and replacement workers to keep their stores open .
(B10) the three grocery chains were managers and replacement workers to keep their stores open relying on store .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C10) three of the grocery store chains , were relying on the managers and to replacement people to keep their stores to open windows
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A11) on thursday , a washington post article argued that a 50 basis point cut from the fed was more likely , contrary to the wall street journal 's line .

(B11) a washington post article argued that a 50 basis point cut from the fed was more likely , contrary to the wall street journal on thursday , 's line .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C11) the washington post article argued that a were more likely to the wall street journal is planned changes on thursday , a 50-basis-point cut out ellison said that contrary to the
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A12) a chinese court yesterday jailed two people for life and imprisoned 12 others for organising a sex party involving hundreds of japanese tourists .
(B12) a chinese court yesterday jailed for two people and imprisoned for 12 others organising a sex party involving hundreds of japanese life tourists .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**y***
(C12) in a court and the people imprisoned for life and others for the leader of a sex that hundreds of japanese tourists gained chinese jailed 12 organising
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A13) the bluecore3-multimedia includes a 16-bit stereo audio codec with dual adc and dac for stereo audio .
(B13) the bluecore3-multimedia includes a 16-bit stereo audio codec with dual adc dac for stereo and audio .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C13) in the bluecore3-multimedia had a 16-bit stereo audio codec with the dual adc and to dac for the stereo audio gained
*Fluent=**n**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A14) she first went to a specialist for initial tests last monday , feeling tired and unwell .
(B14) she first went to a specialist for initial tests last monday , feeling tired , unwell .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C14) goodrem first , went to a specialist for the past monday feeling tired and to unwell changes to the initial tests )
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**y***

(A15) the first health-care worker in the country to die of sars was a filipina-canadian who contracted the disease at north york general hospital , the site of the second outbreak .
(B15) the first health-care worker who contracted the disease , the site of the second outbreak in the country to die of a filipina-canadian sars was at north york general hospital .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***
(C15) health-care worker in russia is a contracted to the public in general , the site of the changes in the first to die of sars , filipina-canadian last york hospital in the second outbreak
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A16) a floating airfield with a flight deck covering 4.5 acres , the ship took about five years to build .
(B16) a floating airfield , the ship took about five years to build a flight with deck covering 4.5 acres .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(C16) in a airfield flight deck of five years to build changes , the ship took off with a covering were down about 1 acres of
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=y, Gain=**n***

(A17) dixon was otherwise the class of the field at pikes peak international raceway .
(B17) the class of the dixon was otherwise field at pikes peak international raceway .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C17) scott was in the field at pikes peak international raceway changes in the class otherwise
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A18) halabi 's military attorney , air force maj. james key , denied the charges , which could carry a death penalty .
(B18) air force maj. james key , denied the charges , which could carry a death penalty , halabi 's military attorney .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C18) the military lawyer , air force maj. out of the charges , which could carry out of the death penalty gained halabi is james key is denied
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A19) garner said the self-proclaimed mayor of baghdad , mohammed mohsen al-zubaidi , was released after two days in coalition custody .
(B19) garner , mohammed mohsen al-zubaidi , said the mayor , mohammed mohsen al-zubaidi was released after two days in custody self-proclaimed of baghdad , coalition .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C19) garner said the falluja mayor of the mohammed mohsen al-zubaidi ) were released two days of the changes long-anticipated custody in baghdad , self-proclaimed
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A20) unit volumes also set a record as notebooks accounted for more than 40 percent of sales .
(B20) unit volumes and set in a record as notebooks accounted for more than 40 percent of sales .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C20) volumes set a record , as notebooks accounted for more than the 40 percent of the unit sales .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(A21) apple computer 's new online music service sold more than 1 million songs during its first week of operation , the company said monday .
(B21) apple computer 's new online music service sold more than 1 million songs during its first week of operation , the company , monday , said monday .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C21) the new online music subscription sold over one million , the songs in its first week of operation , the company said monday gained apple 's church and more
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A22) klarman was arrested by fbi agents in the hamptons , an exclusive summer resort enclave east of new york city .

(B22) klarman was arrested by fbi agents in the east of new york city hamptons , an exclusive summer resort enclave .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C22) klarman had been arrested by the fbi agents in hamptons summer resort , an exclusive enclave east of the new york gained
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A23) the daily hurriyet said the raid aimed to foil a turkish plot to kill an unnamed senior iraqi official in kirkuk .
(B23) the turkish daily hurriyet , an unnamed senior iraqi official , said a plot to kill an unnamed senior iraqi official in the raid aimed to foil kirkuk .
Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n**
(C23) the turkish daily hurriyet foil to a plot to kill a government official , in kirkuk changes in the raid aimed at the unnamed iraqi , u.s.
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**y***

(A24) the last month has been a whirlwind for an 18-year-old who led his high school team to the ohio state championship .
(B24) the last month has been a whirlwind an 18-year-old high school who led his team for to the ohio state championship .
*Fluent=**n**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C24) in the past month has been a whirlwind for an 18-year-old who headed the ohio , state championship team to the high school in the changes
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A25) box cutters were used as a weapon by the sept. 11 , 2001 , hijackers and have since been banned as carry-on items .
(B25) the sept. 11 , 2001 , and have since been banned as carry-on items box cutters were used as a weapon by hijackers .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C25) box cutters , were used as a weapon of the sept. 11 , 2001 and to have carry-on items windows and was banned , as hijackers
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A26) murdoch 's family owns about 30 percent of news corp. shares .
(B26) murdoch 's family owns about 14 percent of news corp. shares .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**y***
(C26) murdoch family 's owns , about 30 percent of threat corp. shares .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**y***

(A27) the best-performing stock was altria group inc. , which rose more than 27 percent to close at $ 42.31 a share .
(B27) the stock rose more than 27 percent to close at $ best-performing was altria group inc. , which 42.31 a share .
*Fluent=**n**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C27) in the stock was ) , which rose more than 27 percent , at $ 42.31 of a share . best-performing its group inc.
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A28) `` the gpl violates the u.s. constitution , together with copyright , antitrust and export control laws , and ibms claims based thereon , or related thereto , are barred . ''

(B28) `` the gpl violates the u.s. constitution , together with copyright , antitrust and export control laws , and ibms claims , or related thereto , based thereon are barred , related thereto said .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**y***
(C28) the `` the gpl violates the u.s. constitution , together with copyright , antitrust and export control laws , saying thereon ) , or related thereto , are too barred windows and to ibms under
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A29) lurd 's ja ' neh also called for an interim government and the deployment of a u.s.-led western peacekeeping force .
(B29) lurd also called for an interim government and the deployment of a u.s.-led western peacekeeping force 's ja ' neh .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C29) the government and the deployment of a u.s.-led western peacekeeping force . lurd is ja ' neh called for an .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A30) the cgt warned of prolonged industrial action unless raffarin agreed .
(B30) the cgt warned of industrial action prolonged unless raffarin agreed .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C30) in the cgt warned of prolonged industrial action unless raffarin agreed gained
*Fluent=**n**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A31) `` there are a number of locations in our community , which are essentially vulnerable , '' mr ruddock said .
(B31) `` there are a number of locations in our community , which are essentially vulnerable . ''
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C31) furthermore , there are in a number of locations in our community ) , which are essentially the vulnerable , a romance ruddock june .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A32) scientists believed stardust trapped thousands of particles of dust .
(B32) scientists believed trapped stardust thousands of particles of dust .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C32) the scientists believed stardust thousands of particles of the changes dust
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A33) mr pollard said : `` this is a terrible personal tragedy and a shocking blow for james 's family .
(B33) mr pollard , a shocking blow , said : `` this is a terrible tragedy for a shocking blow 's personal and family james .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C33) the pollard said in a terrible tragedy and shocking blow for the family 's james changes that `` the is all personal
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A34) the results beat the 54 cents loss per share consensus estimate of 23 analysts polled by thomson first call .

(B34) the results consensus estimate of analysts polled by thomson first call beat the 54 cents loss per share 23 .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C34) the results of the 54 cents per share , research unit of 21 analysts polled by thomson first call . beat loss
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A35) prosecutors did an about-face in may and asked that the autopsy reports be unsealed after portions of conner peterson 's autopsy report favorable to the defense were leaked to the media .
(B35) prosecutors did a about-face in may and asked that the autopsy reports be unsealed after portions of conner peterson 's autopsy report favorable to the defense were leaked to the media .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C35) the government had to be in the portions of the conner peterson 's favorable to the survey , were leaked to the changes in the media about-face , in and asked that the autopsy results unsealed autopsy the defense
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A36) the tech-laced nasdaq composite index . ixic eased 5.16 points , or 0.32 percent , at 1,590.75 , breaking a six-day string of gains .
(B36) the tech-laced nasdaq composite index . ixic eased 5.16 points , or 0.32 percent , at breaking a six-day string of gains , 1,590.75 .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C36) in the 5.16 points , or 0.32 percent , to tech-laced nasdaq composite index ( . ixic ) slipped to 1,590.75 breaking out of a six-day spate of a gained
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A37) however , in the new jersey case , a panel of the u.s. court of appeals for the 3rd circuit upheld the government by a 2-1 vote .
(B37) however , a u.s. court of appeals for the 3rd circuit upheld the government in the new jersey , panel of the case by a 2-1 vote .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C37) however , in the new jersey , the panel of federal appeals court for the 3rd circuit ruled that the government by 62-43 whopping changes in the case
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**y***

(A38) by state law , 911 calls are not public information and were not released .
(B38) by state law , 911 calls are not public information , were not released .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C38) by state law ( 911 calls are not the public and were not released .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A39) eight firefighters also suffered minor injuries , including burns , heat exhaustion , bruises and muscle strain , and were treated and released from the hospital , fire officials said .
(B39) eight firefighters also suffered minor injuries , including heat exhaustion , bruises and muscle strain , and were treated and released from the hospital , fire officials , burns , eight firefighters , said .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(C39) firefighters suffered low heat exhaustion ) bruises and to muscle strain ) and were treated and to release from the hospital , officials said launch ) gained 8 , and had injuries burns
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A40) suvs parked on residential streets in monrovia were tagged with `` elf '' and other slogans , and another was set ablaze in front of a house , sgt. tom wright said .
(B40) residential streets and other slogans , set ablaze in front of a house were tagged with elf , sgt. tom wright said , and another was `` in suvs parked on monrovia , sgt. tom wright , sgt. tom wright , said .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**y***
(C40) suvs parked on the streets , in monrovia , were tagged with the second underground group was set ablaze , in front of a house , said the sergeant laboratory was quoted as other slogans ) and another residential june .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A41) on tuesday , the nikkei average dropped 107.08 points , or 1.3 percent , to close at 8,120.24 .
(B41) on tuesday , the nikkei average dropped points , or 1.3 percent , to close 107.08 at 8,120.24 .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C41) in the 107.08 points , or 0.4 percent , at 8,120.24 ) , the nikkei on tuesday , having gained ,
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***

(A42) the highest-paid washington private-college president is susan resneck pierce , who heads the university of puget sound .
(B42) the highest-paid washington private-college president is pierce , who heads the university of puget sound susan resneck .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C42) the president of the heads of the university of puget sound changes in the highest-paid private-college , is susan resneck pierce ) ,
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A43) of 24 million phoned-in votes , 50.28 percent were for studdard , putting him 130,000 votes ahead of aiken .
(B43) of 24 million phoned-in votes , were ahead of 130,000 , putting him 50.28 percent for studdard votes aiken .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***
(C43) 5.1 million , the phoned-in votes ) and putting him , 130,000 ahead of the changes aiken votes for the studdard out of the 50.28 percent , to
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A44) the exotic animal trade is enormous , and it continues to spiral out of control .
(B44) the exotic animal trade is enormous , and continues to spiral out of control it .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C44) in the exotic animal trade , is enormous ) and to the continues to spiral out of control .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(A45) in a mixture of ancient pagan and modern christian rites , the villagers have staged a series of ceremonies hoping to erase the misfortunes they believe have kept them poor .
(B45) in a mixture of modern christian rites have staged a series of ceremonies hoping to erase the misfortunes believe they have kept them poor and the villagers , ancient pagan .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***
(C45) in mixture of the ancient pagan and to have a series of ceremonies hoping to erase misfortunes have kept them , they believed modern christian rites , the villagers rallied in the poor gained
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A46) part of the accord was the implementation of a special health council that would monitor health spending and progress in reforming the health system .
(B46) the accord was part of the implementation of a special health council that would monitor the progress in reforming health spending and health system .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***
(C46) the accord was implementation of the health council that will monitor the health spending and to progress , in reforming health system changes in the special within
*Fluent=**n**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**y**, Lost=**n**, Gain=**n***

(A47) yes , from today flash memory purchased from amd or fujitsu will be branded spansion .
(B47) yes , will be spansion flash memory purchased from today from amd or fujitsu branded .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C47) `` purchased from the amd , or fujitsu will be branded gained spansion flash memory yes , from the
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

(A48) yee is a chinese-american who converted to islam after graduating from the u.s. military academy at west point .
(B48) after graduating from the u.s. military academy at west point yee is a chinese-american who converted to islam .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**y***
(C48) the who converted to the islam in the military academy in the west point changes in the graduating from the yee is chinese-american
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**y***

(A49) ending the picketing at ralphs frees up about 18,000 union members .
(B49) ending ralphs frees up the picketing at about 18,000 union members .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C49) ralphs frees up the end of the picketing , at a 18,000 union members .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A50) shares of allergan fell 14 cents to close at $ 78.12 on the new york stock exchange ( news - web sites ) .
(B50) shares of allergan fell 14 cents to close at $ 78.12 ( news - web sites ) on the new york stock exchange .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***

(C50) shares of allergan fell 12 cents to close at $ 78.12 on the new york stock exchange ( news - web sites said .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**y***

(A51) three southern politicians who `` stood up to ancient hatreds '' were honored monday with profile in courage awards from the john f. kennedy library and museum .
(B51) three southern politicians who `` stood up to ancient hatreds '' were honored monday with profile in courage awards from the john f. kennedy library , museum .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C51) three southern politicians who stood up the second to the ancient hatreds too , were honored monday with profile in courage awards from the john f. kennedy library and to gained from the museum
*Fluent=**n**, Paraphrase=**n**, Rearrange=**n**, Phrase Alter=**y**, Lost=**n**, Gain=**y***

(A52) the magazine glorifies the soldiers but not necessarily the bush administration , which put them in iraq .
(B52) the bush administration , which put them in the magazine glorifies the soldiers but not necessarily iraq .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C52) the soldiers , but not necessarily the administration ) , which put them in the nation . rosie glorifies in the
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A53) in recent years , hampton kept in touch with friends and stayed in trouble : he faced charges of fare-beating and credit card theft .
(B53) in recent years , hampton kept in touch with friends , stayed in trouble : he faced charges of fare-beating and credit card theft .
*Fluent=**y**, Paraphrase=**y**, Rearrange=**n**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C53) in recent years , hampton kept in touch with them , in the faced charges of fare-beating and credit-card theft windows and to stayed that dogged
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A54) young has 28 days to file a response and ask the nasd for a hearing .
(B54) young 28 days to file a response and ask for a hearing has the nasd .
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**y**, Gain=**n***
(C54) female and 28 days to file in a response and ask for a hearing , changes in the nasd
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**y***

(A55) hovan did not speak , but his lawyer , john speranza , said his client `` did not wake up that day '' intending to hurt anyone .
(B55) hovan did not speak , his lawyer , john , hovan , said his client `` did not wake up that day intending to hurt anyone , but speranza , hovan said .
*Fluent=**y**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**n**, Lost=**n**, Gain=**n***
(C55) hovan 's lawyers said his client had not received wake up the day that the book intending to the cross-border anyone changes , but did n't speak , said in the john speranza
*Fluent=**n**, Paraphrase=**n**, Rearrange=**y**, Phrase Alter=**y**, Lost=**y**, Gain=**n***

**Appendix B** – Human evaluation results

Human judgments were obtained by administering the survey (see Appendix C) to each person, and randomly selecting sentences that needed judgment to present to the participant. Three human judgments were collected for each sentence pair from each system. The judgments were then averaged to obtain the final results reported in this thesis. Full human evaluation results can be obtained by contacting the author or BYU's NLP research group.

**Appendix C** – Survey materials

## *Welcome to the BYU Paraphrase Project Survey*

This survey is being conducted by a BYU Computer Science Master's student to evaluate differences in paraphrases. A paraphrase of a sentence is a rewording -- possibly using different words -- which roughly means the same thing; a paraphrase of a sentence does not have to be longer or shorter than the original sentence.

You will receive one survey consisting of 30 pairs of sentences. You will be asked to answer a few questions about each pair of sentences. The questions are the same for each sentence pair and will help evaluate:

1. whether or not the sentences mean roughly the same thing, and
2. how different the two sentences are from each other.

Please expect this survey to take approximately 20-30 minutes.

There are no anticipated risks or benefits to participating in this study. **Participation in this research study is voluntary.** You may withdraw at any time during the experiment without penalty or refuse to participate entirely. To withdraw from the study, simply close your web browser.

This study is anonymous; your survey will not be linked to your name or identification number. *Your completion and submission of the survey is considered your consent to participate in the research project.*

Participants were chosen from the BYU student body population by recruiting volunteers via email, flyer, or classroom announcement where the professor and/or departments approved such contact.

If you have questions regarding this study you may contact Rebecca Madsen at phone number: (801) 362-9588, or email her at: rmadsen@byu.net.

If you have questions regarding your rights as a participant in research projects, you may contact Dr. Renea Beckstrand, Chair of the Institutional Review Board for Human Subjects, 422 SWKT, Brigham Young University, Provo, UT 84602; phone, (801) 422-3873; email, renea_beckstrand@byu.edu.

Click here to continue

## *Instructions*

You will be given 30 sentence pairs to evaluate, which should take about half an hour. Each pair of sentences will be followed by the same 6 questions. These questions are meant to help evaluate two things:

1. whether or not the pair of sentences are paraphrases of each other (whether they mean roughly the same thing), and
2. how different the two sentences are from each other.

**Important!** Because of the way the sentences were generated, they lack capitalization and proper punctuation. Focus on the words in the sentences instead of punctuation, and try to excuse minor errors in how fluent the English is when answering the survey questions. (Major errors will, of course, affect meaning, so answer questions accordingly.) We are primarily interested in the meaning of the sentences and how varied they are from each other. There are no right or wrong answers, though, so just use your best judgment.

### Consider the following example

A: nationally , the federal centers for disease control and prevention recorded 4,156 cases of west nile , including 284 deaths .

B: the federal centers for disease control and prevention recorded 4,156 cases of west nile , including 284 deaths , nationally .

- Were you able to understand the meaning of sentence B? *(yes/no)*
- Do these two sentences mean the same thing? *(yes/no)*
- Does sentence B have phrases that have moved from the order in sentence A? *(yes/no)*
- Does sentence B have phrases that are reordered internally, or substituted with different, equivalent phrases? *(yes/no)*
- Was there information lost in sentence B? *(yes/no)*
- Was there information gained in sentence B? *(yes/no)*

### Explanation of questions

### The first two questions are asking about the meaning (more or less) of the sentences:

*Were you able to understand the meaning of sentence B?*
Sentence B makes sense for the most part; if there was anything that was not exactly fluent English, it is not noticable enough to bar understanding of meaning. If you think this is the case, you would mark the first question **yes.**

*Do these two sentences mean the same thing?*
Since sentences A and B are saying basically the same thing, you would mark the second question **yes.**

## The next four questions are meant to evaluate how different the two sentences are:

*Does sentence B have phrases that have moved from the order in sentence A?*
The sentences are structured in the same way, with subject and verb and prepositional phrase in the same positions in both sentences, but the phrase "nationally" has moved from the beginning of the sentence to the end, so you would mark the third question **yes.**

*Does sentence B have phrases that are reordered internally, or substituted with different, equivalent phrases?*
The fourth question is asking whether the ordering **within a phrase** is different, or whether one phrase has been replaced by another that means the same thing. "Water bottle" and "bottle of water" would be one example of phrasal alternation. Another example is "big screen" being replaced by "movie theater." In the above example, all phrases are the same in both sentences, so you would mark this question **no.**

*Was there information lost in sentence B?*
*Was there information gained in sentence B?*
The last two questions deal with the amount of information each sentence contains. In the above example, both sentences have the same information so you might mark both questions **no.**

Click here to begin the survey. This should open a new window, so you may refer to these instructions throughout your survey. Please plan on the survey taking about 30 minutes to complete.

Sentence A: another suicide bomber killed three israel soldiers six days later in an attack on the jewish settlement of ariel , north of ramallah , in the west bank .
Sentence B: another suicide bomber was a week later , in an attack on the jewish settlement of ariel ) north of ramallah in the west bank changes , in the soldiers and three of israel

Were you able to understand the meaning of sentence B?

Yes No

Do these two sentences mean the same thing?

Yes No

Does sentence B have phrases that have moved from the order in sentence A?*

Yes No

Does sentence B have phrases that are reordered internally, or substituted with different, equivalent phrases?**

Yes No

Was there information lost in sentence B?

Yes No

Was there information gained in sentence B?

Yes No

Next

*Completed 0 of 30.*

* e.g. "the boy bought candy for his sister from the store" -> "the boy bought candy from the store for his sister"; or "the cat chased the mouse" -> "the mouse was chased by the cat"

** e.g. "water bottle" -> "bottle of water"; or "big screen" -> "movie theater"

**Appendix D** – IRB application information

# Application for the Use of Human Subjects

## Part A Application Information

| | | | |
|---|---|---|---|
| 1. Title of the Study:<br>Generating Paraphrases with Greater Variation using Syntactic Phrases | | | |
| 2. Principal Investigator:<br>Rebecca Madsen | | 3. Contact Person:<br>(if different from PI): | |
| Title: graduate student | Dept: Computer Science | Title: | Dept: |
| Address (+ ZIP):<br>1169 W 500 N, Provo, UT 84601 | | Address (+ Zip): | |
| Phone: 801.362.9588 | Email:<br>rmadsen@byu.net | Phone: | Email: |
| 4. Co-Investigator(s): Eric Ringger<br>(Name & Affiliation): Assistant Professor, Computer Science department, Brigham Young University | | | |
| 5. Research Originated By: (Check One) ~ Faculty X Student ~ Staff | | | |
| 6. Research Purpose: ~ Grant ~ Dissertation X Thesis ~ ORCA Scholarship<br>(Check All that Apply) ~ Other ~ Honors Thesis ~ Course Project: Which Course? | | | |
| 7. Correspondence Request: X Mail ~ Call for Pick-Up | | | |

## Part B Research Study Synopsis

| |
|---|
| 1. Short Study Description:<br>This project will explore a method for generating paraphrases of sentences, or creating a novel sentence that means the same thing as the input sentence using different words. Human subjects will help to evaluate the results of the project by participating in an online survey. They will be given pairs of sentences – the input sentence and the sentence generated by the computer system – and will answer a question about the validity and variety of sentences. |
| 2. Study Length<br>The program is being constructed presently. Evaluation (the surveys) should be conducted in March so the thesis may be completed in April. |
| 3. Location of Research<br>a. Where will the research take place?<br>I am planning on a web-based survey; subjects would be recruited on campus.<br>b. Will the PI be conducting and/or supervising research activity at any sites not under the jurisdiction of the BYU<br>IRB? ~ Yes X No If Yes, please list sites: |
| 4. Subject Information:<br>a. Number of Subjects: 39-65 b. Gender of Subjects: either c. Ages of Subjects: college-aged |

| |
|---|
| 5. Potentially Vulnerable Populations: (Check All that Apply)<br>~ Children ~ Pregnant Women ~ Cognitively Impaired ~ Prisoners ~ Institutionalized<br>~ Faculty's Own Students ~ Other. Please describe: None of the above. |
| 6. Non-English Speaking Subjects<br>a. Will subjects who do not understand English participate in the research: ~ Yes X No<br>b. If yes, describe your resources to communicate with the subjects: NA, subjects need to know English since the data they need to evaluate is in English.<br>c. Into what language(s) will the consent form be translated: None. |
| 7. Additional Subject Concerns<br>a. Are there cultural attitudes/beliefs that may affect subjects in this study? ~ Yes X No<br>b. If yes, please describe attitudes and how they may affect subjects. |
| 8. Dissemination of Research Findings<br>a. Will the research be published? X Yes ~ No If yes, where if known? thesis, unknown if other<br>b. Will the research be presented? X Yes ~ No If yes, where if known? thesis, unknown if other |
| 9. External Funding<br>a. Are you seeking external funding? ~ Yes X No What agency?<br>b. Have you received funding? ~ Yes X No c. Dollar amount? |
| 10. Method of Recruitment: (Check All that Apply)<br>X Flyer X Classroom Announcement ~ Letter to Subjects ~ Third Party ~ Random X Other (Email) |
| 11. Payment to Subjects<br>a. Will subjects be compensated for participation? ~ Yes X No If yes, please indicate amount: NA<br>b. Form of Payment: ~ Cash ~ Check ~ Gift Certificate ~ Voucher ~ 1099 ~ Other<br>c. Will Payment be prorated? ~ Yes ~ No If yes, please explain: NA<br>d. When will the subject be paid? ~ Each Visit ~ Study Completion ~ Other NA |
| 12. Extra Credit<br>a. Will subjects be offered extra credit? ~ Yes X No<br>b. If yes, describe the alternative: |
| 13. Risks: Identify all potential risks/discomforts to subjects.<br>Survey will take some of their time. |
| 14. Benefits:<br>a. Are there direct benefits to participants? ~ Yes X No If yes, please list.<br>b. Are there potential benefits to society? X Yes ~ No If yes, please list. |

Advancing research in the area.

15. Study Procedures:

a. What will be the duration of the subjects' participation?

Depends on how long they take to answer the questions – estimated to be 20-30 minutes in length.

b. Will the subjects be followed after their participation ends? ~ Yes X No If yes, please describe.

c. Describe the number, duration and nature of visits/encounters.

Only one encounter, one survey, duration about 20 minutes to 30 minutes.

d. Is the study ~ Therapeutic? X Non-therapeutic?

e. List all procedures that will be performed to generate data for the research.

Email and flyer designed to recruit volunteers; survey questions about the paraphrases.

f. List all procedures/questionnaires done solely for the purpose of the research study.

Email and flyer designed to recruit volunteers; survey questions about the paraphrases.

g. List all procedures/questionnaires participants already do regardless of research.

None.

16. Informed Consent:

a. Are you requesting Waiver or Alteration of Informed Consent? X Yes ~ No If yes, please fill out the waiver of informed consent and attach it.

b. Briefly describe your process to obtain consent:

Identifying information is not necessary, so the survey should be able to use completion of survey as consent to participate.

17. Confidentiality:

a. Are the subject's social security number, BYU ID number or any identifier (other than study number and initials) being sent off site? ~ Yes X No If yes, describe and explain reasons.

b. Will any entity other than the investigative staff have access to medical, health or psychological information about the subject? ~ Yes X No If yes, please indicate who.

c. Briefly describe provisions made to maintain confidentiality of data, including who will have access to raw data, what will be done with the tapes, etc.

There is no need to have any identifier other than the study number on each of the surveys.

d. Will raw data be made available to anyone other than the PI and immediate study personnel? X Yes ~ No

If yes, describe the procedure for sharing data. Include with whom it will be shared, how and why.

Data will be available to the BYU-NLP research group, and upon request might be available to other researchers in the field.

INSTITUTIONAL REVIEW BOARD FOR
HUMAN SUBJECTS

March 2, 2006

Rebecca Madsen
1169 W. 500 N.
Provo, UT 84601

Dear Rebecca:

Thank you for your recent correspondence concerning your protocol entitled "Generating Paraphrases with Greater Variation Using Syntactic Phrases." The proposal has been assigned the following number: 06-0052. The research appears to pose minimal risk to human subjects and meets the Federal guidelines.

You are approved to begin your research. This approval is good until March 1, 2007 (a year from the date it was approved). A few months before this date we will send out a continuing review form. There will only be two reminders. Please fill this form out in a timely manner to ensure that there is not a lapse in your approval.

Please notify Nancy Davis, (801) 422-2970, A-285 ASB, of any changes made in the instruments, consent form, or research process before instigating the alterations, so that we can approve them before the change is implemented.

If you have any questions, please let us know. We wish you well with your research!

Sincerely,

Nancy A Davis

Dr. Renea L. Beckstrand, Chair /
Nancy A. Davis, CIM, Administrator
Institutional Review Board for Human Subjects
RLB/cfc

BRIGHAM YOUNG UNIVERSITY · A-285 ASB · PROVO, UTAH 84602
(801) 422-3841 / FAX: (801) 422-0620