PARADIGM: Paraphrase Diagnostics through Grammar Matching

Jonathan Weese and **Juri Ganitkevitch**
Johns Hopkins University

Chris Callison-Burch
University of Pennsylvania

Abstract

Paraphrase evaluation is typically done either manually or through indirect, task-based evaluation. We introduce an intrinsic evaluation PARADIGM which measures the goodness of paraphrase collections that are represented using synchronous grammars. We formulate two measures that evaluate these paraphrase grammars using gold standard sentential paraphrases drawn from a monolingual parallel corpus. The first measure calculates how often a paraphrase grammar is able to synchronously parse the sentence pairs in the corpus. The second measure enumerates paraphrase rules from the monolingual parallel corpus and calculates the overlap between this reference paraphrase collection and the paraphrase resource being evaluated. We demonstrate the use of these evaluation metrics on paraphrase collections derived from three different data types: multiple translations of classic French novels, comparable sentence pairs drawn from different newspapers, and bilingual parallel corpora. We show that PARADIGM correlates with human judgments more strongly than BLEU on a task-based evaluation of paraphrase quality.

1 Introduction

Paraphrases are useful in a wide range of natural language processing applications. A variety of data-driven approaches have been proposed to generate paraphrase resources (see Madnani and Dorr (2010) for a survey of these methods). Few objective metrics have been established to evaluate these resources. Instead, paraphrases are typically evaluated using subjective manual evaluation or through task-based evaluations.

Different researchers have used different criteria for manual evaluations. For example, Barzilay and McKeown (2001) evaluated their paraphrases by asking judges whether paraphrases were "approximately conceptually equivalent." Ibrahim et al. (2003) asked judges whether their paraphrases were "roughly interchangeable given the genre." Bannard and Callison-Burch (2005) replaced phrases with paraphrases in a number of sentences and asked judges whether the substitutions "preserved meaning and remained grammatical." The results of these subjective evaluations are not easily reusable.

Other researchers have evaluated their paraphrases through task-based evaluations. Lin and Pantel (2001) measured their potential impact on question-answering. Cohn and Lapata (2007) evaluate their applicability in the text-to-text generation task of sentence compression. Zhao et al. (2009) use them to perform sentence compression and simplification and to compute sentence similarity. Several researchers have demonstrated that paraphrases can improve machine translation evaluation (c.f. Kauchak and Barzilay (2006), Zhou et al. (2006), Madnani (2010) and Snover et al. (2010)).

We introduce an automatic evaluation metric called PARADIGM, PARAphrase DIagnostics through Grammar Matching. This metric evaluates paraphrase collections that are represented using synchronous grammars. Synchronous tree-adjoining grammars (STAGs), synchronous tree substitution grammars (STSGs), and synchronous context free grammars (SCFGs) are popular formalisms for representing paraphrase rules (Dras, 1997; Cohn and Lapata, 2007; Madnani, 2010; Ganitkevitch et al., 2011). We present two measures that evaluate these paraphrase grammars using gold standard sentential paraphrases drawn from a monolingual parallel corpus, which have been previously proposed as a good resource for paraphrase evaluation (Callison-Burch et al., 2008; Cohn et al., 2008).

The first of our two proposed metrics calculates how often a paraphrase grammar is able to synchronously parse the sentence pairs in a test set. The second measure enumerates paraphrase rules from a monolingual parallel corpus and calculates the overlap between this reference paraphrase collection, and the paraphrase resource being evaluated.

2 Related work and background

The most closely related work is ParaMetric (Callison-Burch et al., 2008), which is a set of objective measures for evaluating the quality of *phrase-based* paraphrases. ParaMetric extracts a set of gold-standard phrasal paraphrases from sentential paraphrases that have been manually word-aligned. The sentential paraphrases used in ParaMetric were drawn from a data set originally created to evaluate machine translation output using the BLEU metric. Cohn et al. (2008) argue that these sorts of monolingual parallel corpora are appropriate for evaluating paraphrase systems, because they are naturally occurring sources of paraphrases.

Callison-Burch et al. (2008) calculated three types of metrics in ParaMetric. The manual word alignments were used to calculate how well an automatic paraphrasing technique is able *to align* the paraphrases in a sentence pair. This measure is limited to a class of paraphrasing techniques that perform alignment (like MacCartney et al. (2008)). Most methods produce a list of paraphrases for a given input phrase. So Callison-Burch et al. (2008) calculate two more generally applicable measures by comparing the paraphrases in an automatically extracted resource to gold standard paraphrases extracted via the alignments. These allow a *lower-bound on precision* and *relative recall* to be calculated.

Liu et al. (2010) introduce the PEM metric as an alternative to BLEU, since BLEU prefers identical paraphrases. PEM uses a second language as a pivot to judge semantic equivalence. This requires use of some bilingual data. Chen and Dolan (2011) suggest using BLEU together with their metric PINC, which uses n-grams to measure lexical difference between paraphrases.

PARADIGM extends the ideas in ParaMetric from *lexical* and *phrasal* paraphrasing techniques to paraphrasing techniques that also generate *syntactic templates*, such as Zhao et al. (2008), Cohn and Lapata (2009), Madnani (2010) and Ganitkevitch et al. (2011). Instead of extracting gold standard paraphrases using techniques from phrase-based machine translation, we use grammar extraction techniques (Weese et al., 2011) to extract gold standard paraphrase grammar rules from ParaMetric's word-aligned sentential paraphrases. Using these rules, we calculate the overlap between a gold standard paraphrase grammar and an automatically generated paraphrase grammar.

Moreover, like ParaMetric, PARADIGM is able to do further analysis on a restricted class of paraphrasing models. In this case, PARADIGM evaluates how well certain models are able to produce synchronous parses of sentence pairs drawn from monolingual parallel corpora. PARADIGM's different metrics are explained in Section 4, but first we give background on synchronous parsing and synchronous grammars.

2.1 Synchronous parsing with SCFGs

Synchronous context-free grammars

An SCFG (Lewis and Stearns, 1968; Aho and Ullman, 1972) is similar to a context-free grammar, except that it generates *pairs* of strings in correspondence. Each production rule in an SCFG rewrites a non-terminal symbol as a pair of phrases, which may have contain a mix of words and non-terminals symbols. The grammar is *synchronous* because both phrases in the pair must have an identical set of non-terminals (though they can come in different orders), and corresponding non-terminals must be rewritten using the same rule.

Much recent work in MT (and, by extension, paraphrasing approaches that use MT machinery) has been focused on choosing an appropriate set of non-terminal symbols. The Hiero model (Chiang, 2007) used a single non-terminal symbol X. Other approaches have read symbols from constituent parses of the training data (Galley et al., 2004; Galley et al., 2006; Zollmann and Venugopal, 2006). Labels based combinatory categorial grammar (Steedman and Baldridge, 2011) have also been used (Almaghout et al., 2010; Weese et al., 2012).

Synchronous parsing

Wu (1997) introduced a parsing algorithm using a variant of CKY. Dyer recently showed (2010)

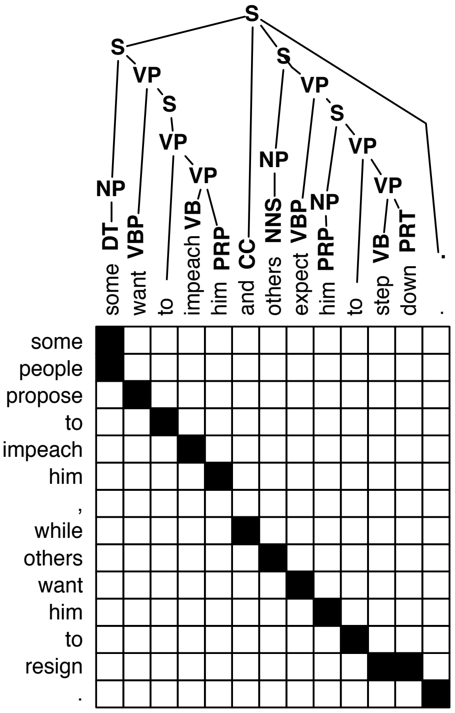

Figure 1: PARADIGM extracts lexical, phrasal and syntactic paraphrases from parsed, word-aligned sentence pairs.

that the average parse time can be significantly improved by using a two-pass algorithm.

The question of whether a source-reference pair is reachable under a model must be addressed in end-to-end discriminative training in MT (Liang et al., 2006a; Gimpel and Smith, 2012). Auli et al. (2009) showed that only approximately 30% of training pairs are reachable under a phrase-based model. This result is confirmed by our results in paraphrasing.

3 Paraphrase grammar extraction

Like ParaMetric, PARADIGM extracts gold standard paraphrases from word-aligned sentential paraphrases. PARADIGM goes further by parsing one of the two input sentences, and uses the parse tree to extract *syntactic* paraphrase rules, following recent advances in syntactic approaches to machine translation (like Galley et al. (2004), Zollmann and Venugopal (2006), and others). Figure 1 shows an example of a parsed sentence pair. From that pair it is possible to extract a wide variety of non-identical paraphrases, which include lexical paraphrases (single word synonyms), phrasal paraphrases, and syntactic paraphrases that include a mix of words and syntactic non-terminal symbols. Figure 2 shows a set of four examples for each type that can be extracted from Figure 1.

CC →	and	while
VBP →	want	propose
VBP →	expect	want
DT →	some	some people
S →	him to step down	him to resign
VP →	step down	resign
VP →	to step down	to resign
VP →	want to impeach him	propose to impeach him
VP →	want VP	propose VP
VP →	want to impeach PRP	propose to impeach PRP
VP →	VBP him to step down	VBP him to resign
S →	PRP to step down	PRP to resign

Figure 2: Four examples each of lexical, phrasal, and syntactic paraphrases that can be extracted from the sentence pair in Figure 1.

These rules are formulated as SCFG rules, with a syntactic left-hand nonterminal symbol and two English right-hand sides representing the paraphrase. The examples above include non-terminal symbols that represent whole syntactic constituents. It is also possible to create more complex non-terminal symbols that describe CCG-like non-constituent phrases. For example, we could extract a rule like

S/VP → <NNS want him to, NNS expect him to>

Using constituents only, we are able to extract 45 paraphrase rules from Figure 1. Adding CCG-style slashed constituents yields 66 additional rules.

4 PARADIGM: Evaluating paraphrase grammars

By considering a paraphrase model as a synchronous context-free grammar, we propose to measure the model's goodness using the following criteria:

1. What percentage of sentential paraphrases are reachable under the model? That is, given a collection of sentence pairs (a_i, b_i) and an SCFG G, where each pair of a and b are sentential paraphrases, how many of the pairs are in the language of G? We evaluate this by producing a synchronous parse for the pairs, as shown in Figure 3.

2. Given a collection of gold-standard paraphrase rules, how many of those paraphrases exist as rules in G? To calculate this, we look at the overlap of grammars (described in

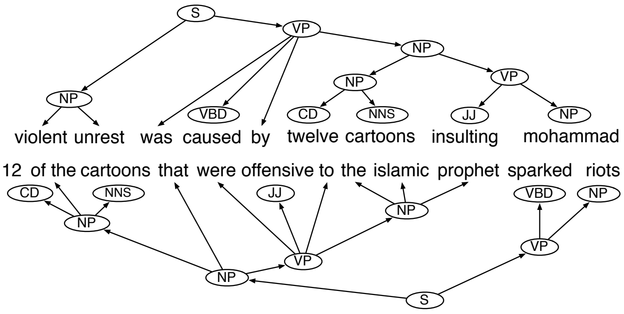
S
VP
NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
violent unrest was caused by twelve cartoons insulting mohammad
12 of the cartoons that were offensive to the islamic prophet sparked riots
CD
NNS
JJ
NP
VBD
NP
NP
VP
VP
NP
S

Figure 3: We measure the goodness of paraphrase grammars by determine how often they can be used to synchronously parse gold-standard sentential paraphrases. Note we do not require the synchronous derivation to match a gold-standard parse tree.

Section 4.2 below), examining different categories of rules and thresholding based on how frequently the rule was used in the gold standard data.

These criteria correspond to properties that we think are desirable in paraphrase models. They also have the advantage that they do not depend on human judgments and so can be calculated automatically.

4.1 Synchronous parse coverage

Paraphrase grammars should be able to explain sentential paraphrases. For example, Figure 3 shows a sentence pair that is synchronously parseable by one paraphrase grammar. In general, we say that the more such sentence pairs that a paraphrase grammar can synchronously parse, the better it is.

The synchronous derivation allows us to draw inferences about parts of the sentence pair that are in correspondence; for instance, in Figure 3, *violent unrest* corresponds to *riots* and *mohammad* corresponds to *the islamic prophet*.

4.2 Grammar overlap defined

We measure grammar overlap by comparing the sets of production rules for two different grammars. If the grammars contain rules that are equivalent, the equivalent rules are in the grammars' overlap.

We consider two types of overlapping, which we will call *strict* and *non-strict* overlap. For strict overlap, we say that two rules are equivalent if they are identical, that is, if they have the same left-hand side non-terminal symbol, their source sides are identical strings, and their target sides are identical strings. (This includes identical indexing on non-terminal symbols on the right hand sides of the rule.)

To calculate non-strict overlap, we ignore the identities of non-terminal symbols in the left-hand and right-hand sides of the rules. That is, two rules are considered equivalent if they are identical after all the non-terminal symbols have been replaced by one equivalent symbol.

For example, in non-strict overlap, the syntactic rule

$$NP \to \langle N_1 \text{ 's } N_2;\ \text{the } N_2 \text{ of } N_1 \rangle$$

would match the Hiero rule

$$X \to \langle X_1 \text{ 's } X_2;\ \text{the } X_2 \text{ of } X_1 \rangle$$

If we are considering two Hiero grammars, strict and non-strict intersection are the same operation since they only have on non-terminal X.

4.3 Precision lower bound and relative recall

Callison-Burch et al. (2008) use the notion of overlap between two paraphrase sets to define two metrics, *precision lower bound* and *relative recall*. These are calculated the same way as standard precision and recall. Relative recall is qualified as "relative" because it is calculated on a potentially incomplete set of gold standard paraphrases. There may exist valid paraphrases that do not occur in that set. Similarly, only a lower bound on precision can be calculated because the candidate set may contain valid paraphrases that do not occur in the gold standard set.

5 Experiments

5.1 Data

We extracted paraphrase grammars from a variety of different data sources, including four collections of sentential paraphrases. These included:

- **Multiple translation corpora** that were compiled by the Linguistics Data Consortium (LDC) for the purposes of evaluating machine translation quality with the BLEU metric. We collected eight LDC corpora that all have multiple English translations.[1]

[1] LDC Catalog numbers LDC2002T01, LDC2005T05, LDC2010T10, LDC2010T11, LDC2010T12, LDC2010T14, LDC2010T17, and LDC2010T23.

Corpus	sentence pairs	total words
LDC Multiple Translations	83,284	2,254,707
Classic French Literature	75,106	682,978
MSR Paraphrase Corpus	5,801	219,492
ParaMetric	970	21,944

Table 1: Amount of English–English parallel data. LDC data has 4 parallel translations per sentence. Literature data is from Barzilay and McKeown (2001). MSR data is from Quirk et al. (2004) and Dolan et al. (2004). ParaMertic data is from Callison-Burch et al. (2008).

Grammar	Rules
LDC Hiero	52,784,462
Lit. Hiero	3,288,546
MSR Hiero	2,456,513
ParaMetric Hiero	584,944
LDC Syntax	23,978,477
Lit. Syntax	715,154
MSR Syntax	406,115
ParaMetric Syntax	317,772
PPDB-v0.2-small	1,292,224
PPDB-v0.2-large	9,456,356
PPDB-v0.2-xl	46,592,161

Table 2: Size of various paraphrase grammars.

Grammar	freq. ≥ 1	freq. ≥ 2
ParaMetric Syntax	317,772	21,709
LDC Hiero	5,840 (1.8%)	416 (1.9%)
Lit. Hiero	6,152 (1.9%)	359 (1.7%)
MSR Hiero	10,012 (3.2%)	315 (1.5%)
LDC Syntax	**48,833 (15.3%)**	7,748 (35.6%)
Lit. Syntax	14,431 (4.5%)	1,960 (9.0%)
MSR Syntax	21,197 (6.7%)	2,053 (9.5%)
PPDB-v0.2-small	15,831 (5.0%)	5,673 (26.1%)
PPDB-v0.2-large	31,277 (9.8%)	8,245 (37.9%)
PPDB-v0.2-xl	47,720 (15.0%)	**10,049 (46.2%)**

Table 3: Size of strict overlap (number of rules and % of the gold standard) of each grammar with a syntactic grammar derived from ParaMetric. freq. $\geq$ 2 means we first removed all rules that appeared only once from the ParaMetric grammar. The number in parentheses shows the percentage of ParaMetric rules that are present in the overlap.

- **Classic French Literature** that were translated by different translators, and which were compiled by Barzilay and McKeown (2001).
- **The MSR Paraphrase corpus** which consists of sentence pairs drawn from comparable news articles drawn from different web sites in the same date rate. The sentence pairs were aligned heuristically aligned and then manually judged to be paraphrases.
- **The ParaMetric data** which consists of 900 manually word-aligned sentence pairs collected by Cohn et al. (2008). 300 sentence pairs were drawn from each of the 3 above sources. We use this to extract the gold standard paraphrase grammar.

The size of the data from each source is summarized in Table 1.

For each dataset, after tokenizing and normalizing, we parsed one sentence in each English pair using the Berkeley constituency parser (Liang et al., 2006b). We then obtained word-level alignments, either using GIZA++ (Och and Ney, 2000) or, in the case of ParaMetric, using human annotations.

We used the Thrax grammar extractor (Weese et al., 2011) to extract Hiero-style and syntactic SCFGs from the paraphrase data. In the syntactic setting we allowed labeling of rules with either constituent labels or CCG-style slashed categories. The size of the extracted grammars is shown in Table 2.

We also used version 0.2 of the SCFG-based paraphrase collection known as the ParaPhrase DataBase or PPDB (Ganitkevitch et al., 2013). The PPDB paraphrases were extracted using the pivoting technique (Bannard and Callison-Burch, 2005) on bilingual parallel corpora containing over 42 million sentence pairs.

The PPDB release includes a tool for pruning the grammar to a smaller size by retaining only high-precision paraphrases. We include PPDB grammars for several different pruning settings in our analysis.

5.2 Experimental setup

We calculated our two metrics for each of the grammars listed in Table 2.

To perform synchronous parsing, we used the Joshua decoder (Post et al., 2013), which includes an implementation of Dyer's two-pass parsing algorithm (2010). After splitting the LDC data into 10 equal pieces, we trained paraphrase models on nine-tenths of the data and parsed the other tenth.

Grammars trained from other sources (the MSR corpus, French literature domain, and PPDB) were also evaluated on the held-out tenth of LDC data.

Grammar	freq. ≥ 1	freq. ≥ 2
ParaMetric Syntax	200,385	20,699
LDC Hiero	41,346 (20.6%)	5,323 (25.8%)
Lit. Hiero	36,873 (18.4%)	4,606 (22.3%)
MSR Hiero	**58,970 (29.4%)**	**6,741 (32.6%)**
LDC Syntax	37,231 (11.7%)	5,055 (24.5%)
Lit. Syntax	19,530 (9.7%)	3,121 (15.1%)
MSR Syntax	28,016 (14.0%)	3,564 (17.2%)
PPDB-v0.2-small	13,003 (6.5%)	3,661 (17.7%)
PPDB-v0.2-large	22,431 (11.2%)	4,837 (23.4%)
PPDB-v0.2-xl	31,294 (15.6%)	5,590 (27.0%)

Table 4: Size of non-strict overlap of each grammar with the syntactic grammar derived from ParaMetric. The number in parentheses shows the percentage of ParaMetric rules that are present in the overlap.

Grammar	syntactic	phrasal	lexical
ParaMetric	238,646	73,320	5,806
LDC$_{Syn}$	36,375 (15%)	8,806 (12%)	3,652 (62%)
MSR$_{Syn}$	7,734 (3%)	11,254 (15%)	2,209 (38%)
PPDB-xl	40,822 (17%)	3,765 (5%)	3,142 (54%)

Table 5: Number of paraphrases of each type in each grammar's strict overlap with the syntactic ParaMetric grammar. Numbers in parentheses show the percentage of ParaMetric rules of each type.

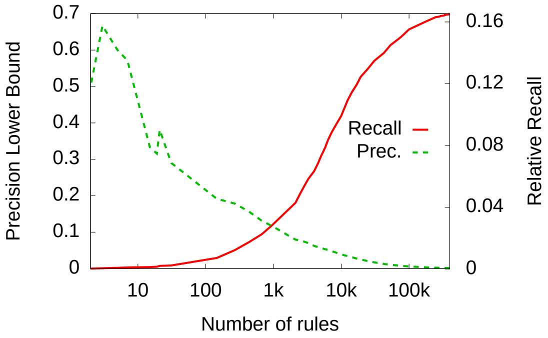

Figure 4: Precision lower bound and relative recall when overlapping different sizes of PPDB with the syntactic ParaMetric grammar.

Note that the LDC data contains 4 independent translations of each foreign sentence, giving 6 possible (unordered) paraphrase pairs. We evaluated coverage in two ways (corresponding to the two columns in Table 6): first, considering all possible sentence pairs from the test data, how many were able to be parsed?

Secondly, if we consider all the English sentences that correspond to one foreign sentence, how many foreign sentences had at least one pair of English translations that could be parsed synchronously?

For grammar overlap, we perform both strict and non-strict calculations (see Section 4.2) against a syntactic grammar derived from hand-aligned ParaMetric data.

5.3 Grammar overlap results

In Table 5 we see a breakdown of the types of paraphrases in the overlap for three of the models. Although the PPDB-xl overlap is much larger than the other two, about 80% of its rules are syntactic transformations. The LDC and MSR models have a much larger proportion of phrasal and lexical rules.

Next we will look at the grammar overlap numbers presented in Table 3 and Table 4.

Note the non-intuitive result that for some grammars (notably PPDB), the non-strict overlap is smaller than the strict overlap. This is because rules with different non-terminals only count once in the non-strict overlap; for example, in PPDB-small,

$$\text{NN} \rightarrow \langle \text{ answer ; reply } \rangle$$
$$\text{VB} \rightarrow \langle \text{ answer ; reply } \rangle$$

count as separate entries when calculating strictly, but when ignoring non-terminals, they count as only one type of rule.

The fact that the non-strict overlaps are smaller means that there must be many rules in PPDB that are identical except for non-terminal labels.

5.4 Precision and recall results

Figure 4 shows relative recall and precision lower bound calculated for various sizes of PPDB relative to the ParaMetric grammar. The x-axis represents the size of the grammar as we vary from keeping only the most probable rules to including less probable ones. Restricting to high probability rules makes the grammar much smaller, resulting in higher precision.

5.5 Synchronous parsing results

Table 6 shows the percentage of sentence pairs that were reachable in a held-out portion of the LDC multiple-translation data.

We find that a grammar trained on LDC data vastly outperforms data from any other domain. This is not surprising — we shouldn't expect a model trained on French literature to be able to

Grammar	% (all)	% (any)
LDC Hiero	9.5	33.0
Lit. Hiero	1.8	9.6
MSR Hiero	1.7	9.2
LDC Syntax	9.1	30.2
Lit. Syntax	2.0	10.7
MSR Syntax	1.9	10.4
PM Syntax	1.7	9.8
PPDB-v0.2-small	1.8	3.3
PPDB-v0.2-large	2.5	4.5
PPDB-v0.2-xl	3.5	6.2

Table 6: Parse coverage on held-out LDC data. The *all* column considers every possible sentential paraphrase in the test set. The *any* column considers a sentence parsed if any of its paraphrases was able to parsed.

handle some of the vocabulary found in news stories that were originally in Arabic or Chinese.

The PPDB data outperforms both French literature and MSR models if we look all possible sentence pairs from test data (the column labeled "all" in the table). However, when we consider whether *any* pair from a set of 4 translations can be translated, the PPDB models do not do as well. This implies that PPDB tends to be able to reach many pairs from the same set of translations, but there are many translations that it cannot handle at all. By contrast, the literature- and MSR-trained models can reach at least one pair from 10% of the test examples, even though the absolute number of pairs they can reach is lower.

5.6 Effects of grammar size and choice of syntactic labels

Table 2 shows that the PPDB-derived grammars are much larger than the syntactic models derived from other domains. It may seem surprising that they should perform worse, but adding more rules to the grammar just by varying non-terminal labels isn't likely to help overall parse coverage. This suggests a new pruning method: keep only the top k label variations for each rule type.

If we compare the syntactic models to the Hiero models trained from the same data, we see that their overall reachability performance is not very different. This implies that paraphrases can be annotated with linguistic information without necessarily hurting their ability to explain particular sentence pairs. Contrast this result, with, for example, those of Koehn et al. (2003), showing that restricting translation models to only syntactic phrases hurts overall translation performance. The comparable performance between Hiero and syntactic models seems to hold regardless of domain.

6 Correlation with human judgments

To validate PARADIGM, we calculated its correlation with human judgments of paraphrase quality on the sentence compression text-to-text generation task, which has been used to evaluate paraphrase grammars in previous research (Cohn and Lapata, 2007; Zhao et al., 2009; Ganitkevitch et al., 2011; Napoles et al., 2011). We created sentence compression systems for five of the paraphrase grammars described in Section 5.1. We followed the methodology outlined by Ganitkevitch et al. (2011) and did the following:

- Each paraphrase grammar was augmented with an appropriate set of rule-level features that capture information pertinent to the task. In this case, the paraphrase rules were given two additional features that shows how the number of words and characters changed after applying the rule.

- Similarly to how the weights of the models are set using minimum error rate training in statistical machine translation, the weights for each of the paraphrase grammars using the PRO tuning method (Hopkins and May, 2011).

- Instead of optimizing to the BLEU metric, as is done in machine translation, we optimized to PRÉCIS, a metric developed for sentence compression that adapts BLEU so that it includes a "verbosity penalty" (Ganitkevitch et al., 2011) to encourage the compression systems to produce shorter output.

- We created a development set with sentence compressions by selecting 1000 pairs of sentences from the multiple translation corpus where two English translations of the same foreign sentences differed in each other by a length ratio of 0.67–0.75.

- We decoded a test set of 1000 sentences using each of the grammars and its optimized weights with the Joshua decoder (Ganitkevitch et al., 2012). The selected in the same fashion as the dev sentences, so each one had a human-created reference compression.

We conducted a human evaluation to judge the meaning and grammaticality of the sentence compressions derived from each paraphrase grammar. We presented workers on Mechanical Turk with the input sentence to the compression sentence (the long sentence), along with 5 shortened outputs from our compression systems. To ensure that workers were producing reliable judgments we also presented them with a positive control (a reference compression written by a person) and a negative controls (a compressed output that was generated by randomly deleted words). We excluded judgments from workers who did not perform well on the positive and negative controls.

Meaning and grammaticality were scored on 5-point scales where 5 is best. These human scores were averaged over 2000 judgments (1000 sentences x 2 annotators) for each system. The systems' outputs were then scored with BLEU, PRÉCIS, and their paraphrase grammars were scored PARADIGM's relative recall and precision lower-bound estimates. For each grammar, we also calculated the average length of parseable sentences.

We calculated the correlation between the human judgements and the automatic scores, using Spearman's rank correlation coefficient ρ. This is methodology is the same that is used to quantify the goodness of automatic evaluation metrics in the machine translation literature (Przybocki et al., 2008; Callison-Burch et al., 2010). The possible values of ρ range between 1 (where all systems are ranked in the same order) and -1 (where the systems are ranked in the reverse order). Thus an automatic evaluation metric with a higher absolute value for ρ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower absolute ρ.

Table 7 shows that our PARADIGM scores correlate more highly with human judgments than either BLEU or PRÉCIS for the 5 systems in our evaluation. This suggests that it may be a better predictor of the goodness of paraphrase grammars than MT metrics, when the paraphrase grammars are used for text-to-text generation tasks.

	MEANING	GRAMMAR
BLEU	-0.7	-0.1
PRÉCIS	-0.6	+0.2
PINC	+0.1	**+0.4**
PARADIGM$_{precision}$	**+0.6**	+0.1
PARADIGM$_{recall}$	+0.1	**+0.4**
PARADIGM$_{avg-len}$	-0.3	**+0.4**

Table 7: The correlation (Spearman's ρ) of different automatic evaluation metrics with human judgments of paraphrase quality for the text-to-text generation task of sentence compression.

7 Summary

We have introduced two new metrics for evaluating paraphrase grammars, and looked at several models from a variety of domains. Using these metrics we can perform a variety of analyses about SCFG-based paraphrase models:

- Automatically-extracted grammars can parse a small fraction of held-out data ($\leq$30%). This is comparable to results in MT (Auli et al., 2009).

- In-domain training data is necessary in order to parse held-out data. A model trained on newswire data parsed 30% of held-out newswire sentence pairs, versus to $<$10% for literature or parliamentary data.

- SCFGs with syntactic labels perform just as well as simpler models with a single non-terminal label.

- Automatically-extracted syntactic grammars tend to have a reasonable overlap with grammars derived from human-aligned data, including more 45% of the gold-standard grammar's paraphrase rules that occurred at least twice.

- We showed that PARADIGM more strongly correlates with human judgments of the meaning and grammaticality of paraphrases produced by sentence compression systems than standard automatic evaluation measures like BLEU.

PARADIGM will help researchers developing paraphrase resources to perform similar diagnostics on their models, and quickly evaluate their systems.

Acknowledgements

This material is based on research sponsored by the NSF under grant IIS-1249516 and DARPA under agreement number FA8750-13-2-0017 (the DEFT program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling*. Prentice Hall.

Hala Almaghout, Jie Jiang, and Andy Way. 2010. CCG augmented hierarchical phrase-based machine translation. In *Proc. of IWSLT*.

Michael Auli, Adam Lopez, Hieu Hoang, and Philipp Koehn. 2009. A systematic analysis of translation model search spaces. In *Proc. WMT*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of ACL*.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. ParaMetric: An automatic evaluation metric for paraphrasing. In *Proc. of COLING*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT10)*.

David L. Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proc. of ACL*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Trevor Cohn and Mirella Lapata. 2007. Large margin synchronous generation and its application to sentence compression. In *Proceedings of EMNLP-CoLing*.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research (JAIR)*, 34:637–674.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4).

William Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrases corpora: Exploiting massively parallel news sources. In *Proc. of COLING*.

Mark Dras. 1997. Representing paraphrases using synchronous tree adjoining grammars. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 516–518, Madrid, Spain, July. Association for Computational Linguistics.

Chris Dyer. 2010. Two monolingual parses are better than one (synchronous parse). In *Proceedings of HLT/NAACL*, pages 263–266. Association for Computational Linguistics.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL 2004: Main Proceedings*, pages 273–280.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve Deneefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of ACL*, pages 961–968.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of EMNLP*.

Juri Ganitkevitch, Yuan Cao, Jonathan Weese, Matt Post, and Chris Callison-Burch. 2012. Joshua 4.0: Packing, pro, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 283–291, Montréal, Canada, June. Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proc. NAACL*.

Kevin Gimpel and Noah A. Smith. 2012. Structured ramp loss minimization for machine translation. In *Proc. of NAACL*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proc. of the Second International Workshop on Paraphrasing*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of EMNLP*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philip M. Lewis and Richard E. Stearns. 1968. Syntax-directed transduction. *Journal of the ACM*, 15(3):465–488.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006a. An end-to-end discriminative approach to machine translation. In *Proc. of ACL*.

Percy Liang, Ben Taskar, and Dan Klein. 2006b. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules from text. *Natural Language Engineering*, 7(3):343–360.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. PEM: a paraphrase evaluation metric exploiting parallel texts. In *Proc. of EMNLP*.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811, Honolulu, Hawaii, October. Association for Computational Linguistics.

Nitin Madnani and Bonnie Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–388.

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, Department of Computer Science, University of Maryland College Park.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. 2011. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97, Portland, Oregon, June. Association for Computational Linguistics.

Franz Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, China, October.

Matt Post, Juri Ganitkevitch, Luke Orland, Jonathan Weese, Yuan Cao, and Chris Callison-Burch. 2013. Joshua 5.0: Sparser, better, faster, server. In *Proc. of WMT*.

Mark Przybocki, Kay Peterson, and Sebastian Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" challenge (MetricsMATR08). In *AMTA-2008 workshop on Metrics for Machine Translation*.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monlingual machine translation for paraphrase generation. In *Proc. of EMNLP*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2010. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. In Robert Borsley and Kersti Börjars, editors, *Non-Transformational Syntax*. Wiley-Blackwell.

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proc. of ACL*.

Jonathan Weese, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the thrax grammar extractor. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 478–484, Edinburgh, Scotland, July. Association for Computational Linguistics.

Jonathan Weese, Chris Callison-Burch, and Adam Lopez. 2012. Using categorial grammar to label translation rules. In *Proc. of WMT*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL/HLT*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of ACL*.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of HLT/NAACL*.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June. Association for Computational Linguistics.