Syntactic Constraints on Paraphrases Extracted from Parallel Corpora

Chris Callison-Burch
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland
ccb@cs.jhu.edu

Abstract

We improve the quality of paraphrases extracted from parallel corpora by requiring that phrases and their paraphrases be the same syntactic type. This is achieved by parsing the English side of a parallel corpus and altering the phrase extraction algorithm to extract phrase labels alongside bilingual phrase pairs. In order to retain broad coverage of non-constituent phrases, complex syntactic labels are introduced. A manual evaluation indicates a 19% absolute improvement in paraphrase quality over the baseline method.

1 Introduction

Paraphrases are alternative ways of expressing the same information. Being able to identify or generate paraphrases automatically is useful in a wide range of natural language applications. Recent work has shown how paraphrases can improve question answering through query expansion (Riezler et al., 2007), automatic evaluation of translation and summarization by modeling alternative lexicalization (Kauchak and Barzilay, 2006; Zhou et al., 2006; Owczarzak et al., 2006), and machine translation both by dealing with out of vocabulary words and phrases (Callison-Burch et al., 2006) and by expanding the set of reference translations for minimum error rate training (Madnani et al., 2007). While all applications require the preservation of meaning when a phrase is replaced by its paraphrase, some additionally require the resulting sentence to be grammatical.

In this paper we examine the effectiveness of placing syntactic constraints on a commonly used paraphrasing technique that extracts paraphrases from parallel corpora (Bannard and Callison-Burch, 2005). The paraphrasing technique employs various aspects of phrase-based statistical machine translation including phrase extraction heuristics to obtain bilingual phrase pairs from word alignments. English phrases are considered to be potential paraphrases of each other if they share a common foreign language phrase among their translations. Multiple paraphrases are frequently extracted for each phrase and can be ranked using a paraphrase probability based on phrase translation probabilities.

We find that the quality of the paraphrases that are generated in this fashion improves significantly when they are required to be the same syntactic type as the phrase that they are paraphrasing. This constraint:

- Eliminates a trivial but pervasive error that arises from the interaction of unaligned words with phrase extraction heuristics.
- Refines the results for phrases that can take on different syntactic labels.
- Applies both to phrases which are linguistically coherent and to arbitrary sequences of words.
- Results in much more grammatical output when phrases are replaced with their paraphrases.

A thorough manual evaluation of the refined paraphrasing technique finds a 19% absolute improve-

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 196–205,
Honolulu, October 2008.

ment in the number of paraphrases that are judged to be correct.

This paper is structured as follows: Section 2 describes related work in syntactic constraints on phrase-based SMT and work utilizing syntax in paraphrase discovery. Section 3 details the problems with extracting paraphrases from parallel corpora and our improvements to the technique. Section 4 describes our experimental design and evaluation methodology. Section 5 gives the results of our experiments, and Section 6 discusses their implications.

2 Related work

A number of research efforts have focused on employing syntactic constraints in statistical machine translation. Wu (1997) introduced the inversion transduction grammar formalism which treats translation as a process of parallel parsing of the source and target language via a synchronized grammar. The synchronized grammar places constraints on which words can be aligned across bilingual sentence pairs. To achieve computational efficiency, the original proposal used only a single non-terminal label rather than a linguistic grammar.

Subsequent work used more articulated parses to improve alignment quality by applying cohesion constraints (Fox, 2002; Lin and Cherry, 2002). If two English phrases are in disjoint subtrees in the parse, then the phrasal cohesion constraint prevents them from being aligned to overlapping sequences in the foreign sentence. Other recent work has incorporated constituent and dependency subtrees into the translation rules used by phrase-based systems (Galley et al., 2004; Quirk et al., 2005). Phrase-based rules have also been replaced with synchronous context free grammars (Chiang, 2005) and with tree fragments (Huang and Knight, 2006).

A number of techniques for generating paraphrases have employed syntactic information, either in the process of extracting paraphrases from monolingual texts or in the extracted patterns themselves. Lin and Pantel (2001) derived paraphrases based on the distributional similarity of paths in dependency trees. Barzilay and McKeown (2001) incorporated part-of-speech information and other morphosyntactic clues into their co-training algorithm. They extracted paraphrase patterns that incorporate this information. Ibrahim et al. (2003) generated structural paraphrases capable of capturing long-distance dependencies. Pang et al. (2003) employed a syntax-based algorithm to align equivalent English sentences by merging corresponding nodes in parse trees and compressing them down into a word lattice.

Perhaps the most closely related work is a recent extension to Bannard and Callison-Burch's paraphrasing method. Zhao et al. (2008b) extended the method so that it is capable of generating richer paraphrase patterns that include part-of-speech slots, rather than simple lexical and phrasal paraphrases. For example, they extracted patterns such as *consider NN → take NN into consideration*. To accomplish this, Zhao el al. used dependency parses on the English side of the parallel corpus. Their work differs from the work presented in this paper because their syntactic constraints applied to slots within paraphrase patters, and our constraints apply to the paraphrases themselves.

3 Paraphrasing with parallel corpora

Bannard and Callison-Burch (2005) extract paraphrases from bilingual parallel corpora. They give a probabilistic formation of paraphrasing which naturally falls out of the fact that they use techniques from phrase-based statistical machine translation:

$$\hat{e_2} = \arg\max_{e_2 : e_2 \neq e_1} p(e_2|e_1) \quad (1)$$

where

$$p(e_2|e_1) = \sum_f p(f|e_1)p(e_2|f,e_1) \quad (2)$$

$$\approx \sum_f p(f|e_1)p(e_2|f) \quad (3)$$

Phrase translation probabilities $p(f|e_1)$ and $p(e_2|f)$ are commonly calculated using maximum likelihood estimation (Koehn et al., 2003):

$$p(f|e) = \frac{count(e,f)}{\sum_f count(e,f)} \quad (4)$$

where the counts are collected by enumerating all bilingual phrase pairs that are consistent with the

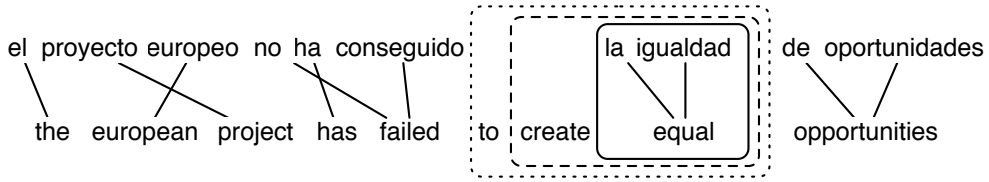

Figure 1: The interaction of the phrase extraction heuristic with unaligned English words means that the Spanish phrase *la igualdad* aligns with *equal*, *create equal*, and *to create equal*.

word alignments for sentence pairs in a bilingual parallel corpus. Various phrase extraction heuristics are possible. Och and Ney (2004) defined consistent bilingual phrase pairs as follows:

$$BP(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n}) : \forall (i', j') \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n \wedge \exists (i', j') \in A : j \leq j' \leq j+m \wedge \leftrightarrow i \leq i' \leq i+n\}$$

where f_1^J is a foreign sentence, e_1^I is an English sentence and A is a set of word alignment points.

The heuristic allows unaligned words to be included at the boundaries of the source or target language phrases. For example, when enumerating the consistent phrase pairs for the sentence pair given in Figure 1, *la igualdad* would align not only to *equal*, but also to *create equal*, and *to create equal*. In SMT these alternative translations are ranked by the translation probabilities and other feature functions during decoding.

The interaction between the phrase extraction heuristic and unaligned words results in an undesirable effect for paraphrasing. By Bannard and Callison-Burch's definition, *equal*, *create equal*, and *to create equal* would be considered paraphrases because they are aligned to the same foreign phrase. Tables 1 and 2 show how sub- and super-phrases can creep into the paraphrases: *equal* can be paraphrased as *equal rights* and *create equal* can be paraphrased as *equal*. Obviously when e_2 is substituted for e_1 the resulting sentence will generally be ungrammatical. The first case could result in *equal equal rights*, and the second would drop the verb.

This problem is pervasive. To test its extent we attempted to generate paraphrases for 900,000 phrases using Bannard and Callison-Burch's method trained on the Europarl corpora (as described in Section 4). It generated a total of 3.7 million paraphrases for

equal			
equal	.35	equally	.02
same	.07	the	.02
equality	.03	fair	.01
equals	.02	equal rights	.01

Table 1: The baseline method's paraphrases of *equal* and their probabilities (excluding items with $p < .01$).

create equal			
create equal	.42	same	.03
equal	.06	created	.02
to create a	.05	conditions	.02
create	.04	playing	.02
to create equality	.03	creating	.01

Table 2: The baseline's paraphrases of *create equal*. Most are clearly bad, and the most probable $e_2 \neq e_1$ is a substring of e_1.

400,000 phrases in the list.[1] We observed that 34% of the paraphrases (excluding the phrase itself) were super- or sub-strings of the original phrase. The most probable paraphrase was a super- or sub-string of the phrase 73% of the time.

There are a number of strategies that might be adopted to alleviate this problem:

- Bannard and Callison-Burch (2005) rank their paraphrases with a language model when the paraphrases are substituted into a sentence.

- Bannard and Callison-Burch (2005) sum over multiple parallel corpora C to reduce the problems associated with systematic errors in the

[1] The remaining 500,000 phrases could not be paraphrased either because $e_2 \neq e_1$ or because they were not consistently aligned to any foreign phrases.

word alignments in one language pair:

$$\hat{e_2} = \arg\max_{e_2} \sum_{c \in C} \sum_{f} p(f|e_1)p(e_2|f) \quad (5)$$

- We could change the phrase extraction heuristic's treatment of unaligned words, or we could attempt to ensure that we have fewer unaligned items in our word alignments.
- The paraphrase criterion could be changed from being $e_2 \neq e_1$ to specifying that $e2$ is not sub- or super-string of $e1$.

In this paper we adopt a different strategy. The essence of our strategy is to constrain paraphrases to be the same syntactic type as the phrases that they are paraphrasing. Syntactic constraints can apply in two places: during phrase extraction and when substituting paraphrases into sentences. These are described in sections 3.1 and 3.2.

3.1 Syntactic constraints on phrase extraction

When we apply syntactic constraints to the phrase extraction heuristic, we change how bilingual phrase pairs are enumerated and how the component probabilities of the paraphrase probability are calculated.

We use the syntactic type s of e_1 in a refined version of the paraphrase probability:

$$\hat{e_2} = \arg\max_{e_2 : e_2 \neq e_1 \wedge s(e_2) = s(e_1)} p(e_2|e_1, s(e_1)) \quad (6)$$

where $p(e_2|e_1, s(e_1))$ can be approximated as:

$$\sum_{c \in C} \frac{\sum_f p(f|e_1, s(e_1))p(e_2|f, s(e_1))}{|C|} \quad (7)$$

We define a new phrase extraction algorithm that operates on an English parse tree P along with foreign sentence f_1^J, English sentence e_1^I, and word alignment A. We dub this SBP for *syntactic bilingual phrases*:

$$\begin{aligned} SBP(f_1^J, e_1^I, A, P) = \{(f_j^{j+m}, e_i^{i+n}, s(e_i^{i+n})) : \\ \forall (i', j') \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n \\ \wedge \exists (i', j') \in A : j \leq j' \leq j+m \wedge \leftrightarrow i \leq i' \leq i+n \\ \wedge \exists \text{ subtree } \in P \text{ with label } s \text{ spanning words } (i, i+n)\} \end{aligned}$$

equal				
JJ	equal	.60	similar	.02
	same	.14	equivalent	.01
	fair	.02		
ADJP	equal	.79	the same	.01
	necessary	.02	equal in law	.01
	similar	.02	equivalent	.01
	identical	.02		

Table 3: Syntactically constrained paraphrases for *equal* when it is labeled as an adjective or adjectival phrase.

The SBP phrase extraction algorithm produces tuples containing a foreign phrase, an English phrase and a syntactic label (f, e, s). After enumerating these for all phrase pairs in a parallel corpus, we can calculate $p(f|e_1, s(e_1))$ and $p(e_2|f, s(e_1))$ as:

$$\begin{aligned} p(f|e_1, s(e_1)) &= \frac{count(f, e_1, s(e_1))}{\sum_f count(f, e_1, s(e_1))} \\ p(e_2|f, s(e_1)) &= \frac{count(f, e_2, s(e_1))}{\sum_{e_2} count(f, e_2, s(e_1))} \end{aligned}$$

By redefining the probabilities in this way we partition the space of possible paraphrases by their syntactic categories.

In order to enumerate all phrase pairs with their syntactic labels we need to parse the English side of the parallel corpus (but not the foreign side). This limits the potential applicability of our refined paraphrasing method to languages which have parsers.

Table 3 gives an example of the refined paraphrases for *equal* when it occurs as an adjective or adjectival phrase. Note that most of the paraphrases that were possible under the baseline model (Table 1) are now excluded. We no longer get the noun *equality*, the verb *equals*, the adverb *equally*, the determier *the* or the NP *equal rights*. The paraphrases seem to be higher quality, especially if one considers their fidelity when they replace the original phrase in the context of some sentence.

We tested the rate of paraphrases that were sub- and super-strings when we constrain paraphrases based on non-terminal nodes in parse trees. The percent of the best paraphrases being substrings dropped from 73% to 24%, and the overall percent of paraphrases subsuming or being subsumed by the original phrase dropped from 34% to 12%. However, the number of phrases for which we were able

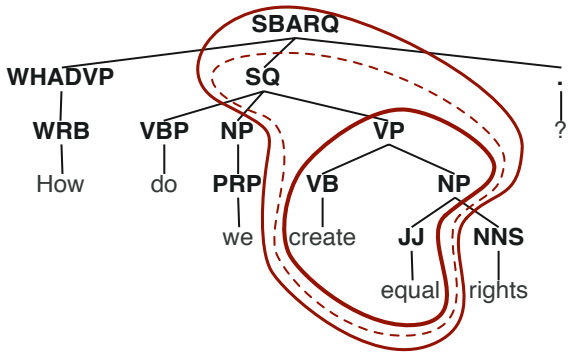

Figure 2: In addition to extracting phrases that are dominated by a node in the parse tree, we also generate labels for non-syntactic constituents. Three labels are possible for *create equal*.

to generated paraphrases dropped from 400,000 to 90,000, since we limited ourselves to phrases that were valid syntactic constituents. The number of unique paraphrases dropped from several million to 800,000.

The fact that we are able to produce paraphrases for a much smaller set of phrases is a downside to using syntactic constraints as we have initially proposed. It means that we would not be able to generate paraphrases for phrases such as *create equal*. Many NLP tasks, such as SMT, which could benefit from paraphrases require broad coverage and may need to paraphrases for phrases which are not syntactic constituents.

Complex syntactic labels

To generate paraphrases for a wider set of phrases, we change our phrase extraction heuristic again so that it produces phrase pairs for arbitrary spans in the sentence, including spans that aren't syntactic constituents. We assign every span in a sentence a syntactic label using CCG-style notation (Steedman, 1999), which gives a syntactic role with elements missing on the left and/or right hand sides.

$$SBP(f_1^J, e_1^I, A, P) = \{(f_j^{j+m}, e_i^{i+n}, s) :$$
$$\forall (i', j') \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n$$
$$\wedge \exists (i', j') \in A : j \leq j' \leq j+m \wedge \leftrightarrow i \leq i' \leq i+n$$
$$\wedge \exists s \in CCG\text{-}labels(e_i^{i+n}, P)\}$$

The function $CCG\text{-}labels$ describes the set of CCG-labels for the phrase spanning positions i to $i+n$ in a parse tree P. It generates three complex syntactic labels for the non-syntactic constituent phrase *create equal* in the parse tree given in Figure 2:

1. VP/(NP/NNS) – This label corresponds to the innermost circle. It indicates that *create equal* is a verb phrase missing a noun phrase to its right. That noun phrase in turn missing a plural noun (NNS) to its right.

2. SQ\VBP NP/(VP/(NP/NNS)) – This label corresponds to the middle circle. It indicates that *create equal* is an SQ missing a VBP and a NP to its left, and the complex VP to its right.

3. SBARQ\WHADVP (SQ\VBP NP/(VP/(NP/NNS)))/. – This label corresponds to the outermost circle. It indicates that *create equal* is an SBARQ missing a WHADVP and the complex SQ to its left, and a punctuation mark to its right.

We can use these complex labels instead of atomic non-terminal symbols to handle non-constituent phrases. For example, Table 4 shows the paraphrases and syntactic labels that are generated for the non-constituent phrase *create equal*. The paraphrases are significantly better than the paraphrases generated for the phrase by the baseline method (refer back to Table 2).

The labels shown in the figure are a fraction of those that can be derived for the phrase in the parallel corpus. Each of these corresponds to a different

create equal		
VP/(NP/NNS)	create equal	.92
	creating equal	.08
VP/(NP/NNS) PP	create equal	.96
	promote equal	.03
	establish fair	.01
VP/(NP/NNS) PP PP	create equal	.80
	creating equal	.10
	provide equal	.06
	create genuinely fair	.04
VP/(NP/(NP/NN) PP)	create equal	.83
	create a level playing	.17
VP/(NP/(NP/NNS) PP)	create equal	.83
	creating equal	.17

Table 4: Paraphrases and syntactic labels for the non-constituent phrase *create equal*.

syntactic context, and each has its own set of associated paraphrases.

We increase the number of phrases that are paraphrasable from the 90,000 in our initial definition of SBP to 250,000 when we use complex CCG labels. The number of unique paraphrases increases from 800,000 to 3.5 million, which is nearly as many paraphrases that were produced by the baseline method for the sample.

3.2 Syntactic constraints when substituting paraphrases into a test sentence

In addition to applying syntactic constraints to our phrase extraction algorithm, we can also apply them when we substitute a paraphrase into a sentence. To do so, we limit the paraphrases to be the same syntactic type as the phrase that it is replacing, based on the syntactic labels that are derived from the phrase tree for a test sentence. Since each phrase normally has a set of different CCG labels (instead of a single non-termal symbol) we need a way of choosing which label to use when applying the constraint.

There are several different possibilities for choosing among labels. We could simultaneously choose the best paraphrase and the best label for the phrase in the parse tree of the test sentence:

$$\hat{e_2} = \arg\max_{e_2 : e_2 \neq e_1} \arg\max_{s \in CCG\text{-}labels(e_1, P)} p(e_2|e_1, s) \quad (8)$$

Alternately, we could average over all of the labels that are generated for the phrase in the parse tree:

$$\hat{e_2} = \arg\max_{e_2 : e_2 \neq e_1} \sum_{s \in CCG\text{-}labels(e_1, P)} p(e_2|e_1, s) \quad (9)$$

The potential drawback of using Equations 8 and 9 is that the CCG labels for a particular sentence significantly reduces the paraphrases that can be used. For instance, VP/(NP/NNS) is the only label for the paraphrases in Table 4 that is compatible with the parse tree given in Figure 2.

Because the CCG labels for a given sentence are so specific, many times there are no matches. Therefore we also investigated a looser constraint. We choose the highest probability paraphrase with any label (i.e. the set of labels extracted from all parse trees in our parallel corpus):

$$\hat{e_2} = \arg\max_{e_2 : e_2 \neq e_1} \arg\max_{s \in \cap_{\forall T \text{ in } C} CCG\text{-}labels(e_1, T)} p(e_2|e_1, s) \quad (10)$$

Equation 10 only applies syntactic constraints during phrase extraction and ignores them during substitution.

In our experiments, we evaluate the quality of the paraphrases that are generated using Equations 8, 9 and 10. We compare their quality against the Bannard and Callison-Burch (2005) baseline.

4 Experimental design

We conducted a manual evaluation to evaluate paraphrase quality. We evaluated whether paraphrases retained the meaning of their original phrases and whether they remained grammatical when they replaced the original phrase in a sentence.

4.1 Training materials

Our paraphrase model was trained using the Europarl corpus (Koehn, 2005). We used ten parallel corpora between English and (each of) Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish, with approximately 30 million words per language for a total of 315 million English words. Automatic word alignments were created for these using Giza++ (Och and Ney, 2003). The English side of each parallel corpus was parsed using the Bikel parser (Bikel, 2002). A total of 1.6 million unique sentences were parsed. A trigram language model was trained on these English sentences using the SRI language modeling toolkit (Stolcke, 2002).

The paraphrase model and language model for the Bannard and Callison-Burch (2005) baseline were trained on the same data to ensure a fair comparison.

4.2 Test phrases

The test set was the English portion of test sets used in the shared translation task of the ACL-2007 Workshop on Statistical Machine Translation (Callison-Burch et al., 2007). The test sentences were also parsed with the Bikel parser.

The phrases to be evaluated were selected such that there was an even balance of phrase lengths (from one word long up to five words long), with half of the phrases being valid syntactic constituents and half being arbitrary sequences of words. 410 phrases were selected at random for evaluation. 30 items were excluded from our results subsequent to evaluation on the grounds that they consisted

solely of punctuation and stop words like determiners, prepositions and pronouns. This left a total of 380 unique phrases.

4.3 Experimental conditions

We produced paraphrases under the following eight conditions:

1. **Baseline** – The paraphrase probability defined by Bannard and Callison-Burch (2005). Calculated over multiple parallel corpora as given in Equation 5. Note that under this condition the best paraphrase is the same for each occurrence of the phrase irrespective of which sentence it occurs in.

2. **Baseline + LM** – The paraphrase probability (as above) combined with the language model probability calculated for the sentence with the phrase replaced with the paraphrase.

3. **Extraction Constraints** – This condition selected the best paraphrase according to Equation 10. It chooses the single best paraphrase over all labels. Conditions 3 and 5 only apply the syntactic constraints at the phrase extraction stage, and do not require that the paraphrase have the same syntactic label as the phrase in the sentence that it is being subtituted into.

4. **Extraction Constraints + LM** – As above, but the paraphrases are also ranked with a language model probability.

5. **Substitution Constraints** – This condition corresponds to Equation 8, which selects the highest probability paraphrase which matches at least one of the syntactic labels of the phrase in the test sentence. Conditions 5–8 apply the syntactic constraints both and the phrase extraction and at the substitution stages.

6. **Syntactic Constraints + LM** – As above, but including a language model probability as well.

7. **Averaged Substitution Constraints** – This condition corresponds to Equation 9, which averages over all of the syntactic labels for the phrase in the sentence, instead of choosing the single one which maximizes the probability.

	MEANING
5	All of the meaning of the original phrase is retained, and nothing is added
4	The meaning of the original phrase is retained, although some additional information may be added but does not transform the meaning
3	The meaning of the original phrase is retained, although some information may be deleted without too great a loss in the meaning
2	Substantial amount of the meaning is different
1	The paraphrase doesn't mean anything close to the original phrase
	GRAMMAR
5	The sentence with the paraphrase inserted is perfectly grammatical
4	The sentence is grammatical, but might sound slightly awkward
3	The sentence has an agreement error (such as between its subject and verb, or between a plural noun and singular determiner)
2	The sentence has multiple errors or omits words that would be required to make it grammatical
1	The sentence is totally ungrammatical

Table 5: Annotators rated paraphrases along two 5-point scales.

8. **Averaged Substitution Constraints + LM** – As above, but including a language model probability.

4.4 Manual evaluation

We evaluated the paraphrase quality through a substitution test. We retrieved a number of sentences which contained each test phrase and substituted the phrase with automatically-generated paraphrases. Annotators judged whether the paraphrases had the same meaning as the original and whether the resulting sentences were grammatical. They assigned two values to each sentence using the 5-point scales given in Table 5. We considered an item to have the same meaning if it was assigned a score of 3 or greater, and to be grammatical if it was assigned a score of 4 or 5.

We evaluated several instances of a phrase when it occurred multiple times in the test corpus, since paraphrase quality can vary based on context (Szpektor et al., 2007). There were an average of 3.1 instances for each phrase, with a maximum of 6. There were a total of 1,195 sentences that para-

phrases were substituted into, with a total of 8,422 judgements collected. Note that 7 different paraphrases were judged on average for every instance. This is because annotators judged paraphrases for eight conditions, and because we collected judgments for the 5-best paraphrases for many of the conditions.

We measured inter-annotator agreement with the Kappa statistic (Carletta, 1996) using the 1,391 items that two annotators scored in common. The two annotators assigned the same absolute score 47% of the time. If we consider chance agreement to be 20% for 5-point scales, then $K = 0.33$, which is commonly interpreted as "fair" (Landis and Koch, 1977). If we instead measure agreement in terms of how often the annotators both judged an item to be above or below the thresholds that we set, then their rate of agreement was 80%. In this case chance agreement would be 50%, so $K = 0.61$, which is "substantial".

4.5 Data and code

In order to allow other researchers to recreate our results or extend our work, we have prepared the following materials for download[2]:

- The complete set of paraphrases generated for the test set. This includes the 3.7 million paraphrases generated by the baseline method and the 3.5 million paraphrases generated with syntactic constraints.
- The code that we used to produce these paraphrases and the complete data sets (including all 10 word-aligned parallel corpora along with their English parses), so that researchers can extract paraphrases for new sets of phrases.
- The manual judgments about paraphrase quality. These may be useful as development material for setting the weights of a log-linear formulation of paraphrasing, as suggested in Zhao et al. (2008a).

5 Results

Table 6 summarizes the results of the manual evaluation. We can observe a strong trend in the syntactically constrained approaches performing better than the baseline. They retain the correct meaning more often (ranging from 4% to up to 15%). They are judged to be grammatical far more frequently (up to 26% more often without the language model, and 24% with the language model) . They perform nearly 20% better when both meaning and grammaticality are used as criteria.[3]

	correct meaning	correct grammar	both correct
Baseline	.56	.35	.30
Baseline+LM	.46	.44	.36
Extraction Constraints	**.62**	.57	.46
Extraction Const+LM	.60	.65	.50
Substitution Constraints	.60	.60	.50
Substitution Const+LM	.61	**.68**	.54
Avg Substitution Const	**.62**	.61	.51
Avg Substit Const+LM	.61	**.68**	**.55**

Table 6: The results of the manual evaluation for each of the eight conditions. Correct meaning is the percent of time that a condition was assigned a 3, 4, or 5, and correct grammar is the percent of time that it was given a 4 or 5, using the scales from Table 5.

Another trend that can be observed is that incorporating a language model probability tends to result in more grammatical output (a 7–9% increase), but meaning suffers as a result in some cases. When the LM is applied there is a drop of 12% in correct meaning for the baseline, but only a slight dip of 1-2% for the syntactically-constrained phrases.

Note that for the conditions where the paraphrases were required to have the same syntactic type as the phrase in the parse tree, there was a reduction in the number of paraphrases that could be applied. For the first two conditions, paraphrases were posited for 1194 sentences, conditions 3 and 4 could be applied to 1142 of those sentences, but conditions 5–8 could only be applied to 876 sentences. The substitution constraints reduce coverage to 73% of the test sentences. Given that the extraction constraints have better coverage and nearly identical performance on

[2]Available from `http://cs.jhu.edu/~ccb/`.

[3]Our results show a significantly lower score for the baseline than reported in Bannard and Callison-Burch (2005). This is potentially due to the facts that in this work we evaluated on out-of-domain news commentary data, and we randomly selected phrases. In the pervious work the test phrases were drawn from WordNet, and they were evaluated solely on in-domain European parliament data.

the *meaning* criterion, they might be more suitable in some circumstances.

6 Conclusion

In this paper we have presented a novel refinement to paraphrasing with bilingual parallel corpora. We illustrated that a significantly higher performance can be achieved by constraining paraphrases to have the same syntactic type as the original phrase. A thorough manual evaluation found an absolute improvement in quality of 19% using strict criteria about paraphrase accuracy when comparing against a strong baseline. The syntactically enhanced paraphrases are judged to be grammatically correct over two thirds of the time, as opposed to the baseline method which was grammatically correct under half of the time.

This paper proposed constraints on paraphrases at two stages: when deriving them from parsed parallel corpora and when substituting them into parsed test sentences. These constraints produce paraphrases that are better than the baseline and which are less commonly affected by problems due to unaligned words. Furthermore, by introducing complex syntactic labels instead of solely relying on non-terminal symbols in the parse trees, we are able to keep the broad coverage of the baseline method.

Syntactic constraints significantly improve the quality of this paraphrasing method, and their use opens the question about whether analogous constraints can be usefully applied to paraphrases generated from purely monolingual corpora. Our improvements to the extraction of paraphrases from parallel corpora suggests that it may be usefully applied to other NLP applications, such as generation, which require grammatical output.

Acknowledgments

Thanks go to Sally Blatz, Emily Hinchcliff and Michelle Bland for conducting the manual evaluation and to Michelle Bland and Omar Zaidan for proofreading and commenting on a draft of this paper.

This work was supported by the National Science Foundation under Grant No. 0713448. The views and findings are the author's alone.

References

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

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

Dan Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of HLT*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT/NAACL*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT/NAACL*.

Bryant Huang and Kevin Knight. 2006. Relabeling syntax trees to improve syntax-based machine translation quality. In *Proceedings of HLT/NAACL*.

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

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 *Proceedings of HLT/NAACL*.

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*, Phuket, Thailand.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Dekang Lin and Colin Cherry. 2002. Word alignment with cohesion constraint. In *Proceedings of HLT/NAACL*.

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

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Contextual bitext-derived paraphrases in automatic MT evaluation. In *Proceedings of the SMT Workshop at HLT-NAACL*.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of ACL*.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of ACL*.

Mark Steedman. 1999. Alternative quantier scope in ccg. In *Proceedings of ACL*.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, September.

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

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

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008a. Combining multiple resources to improve SMT-based paraphrasing model. In *Proceedings of ACL/HLT*.

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

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In *Proceedings of EMNLP*.