# Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity

**Tsz Ping Chan, Chris Callison-Burch** and **Benjamin Van Durme**
Center for Language and Speech Processing, and HLTCOE
Johns Hopkins University

## Abstract

This paper improves an existing bilingual paraphrase extraction technique using monolingual distributional similarity to rerank candidate paraphrases. Raw monolingual data provides a complementary and orthogonal source of information that lessens the commonly observed errors in bilingual pivot-based methods. Our experiments reveal that monolingual scoring of bilingually extracted paraphrases has a significantly stronger correlation with human judgment for grammaticality than the probabilities assigned by the bilingual pivoting method does. The results also show that monolingual distribution similarity can serve as a threshold for high precision paraphrase selection.

## 1 Introduction

Paraphrasing is the rewording of a phrase such that meaning is preserved. Data-driven paraphrase acquisition techniques can be categorized by the type of data that they use (Madnani and Dorr, 2010). Monolingual paraphrasing techniques cluster phrases through statistical characteristics such as dependency path similarities or distributional co-occurrence information (Lin and Pantel, 2001; Pasca and Dienes, 2005). Bilingual paraphrasing techniques use parallel corpora to extract potential paraphrases by grouping English phrases that share the same foreign translations (Bannard and Callison-Burch, 2005). Other efforts blur the lines between the two, applying techniques from statistical machine translation to monolingual data or extracting paraphrases from multiple English translations of the same foreign text (Barzilay and McKeown, 2001; Pang et al., 2003; Quirk et al., 2004).

We exploit both methodologies, applying a monolingually-derived similarity metric to the output of a pivot-based bilingual paraphrase model. In this paper we investigate the strengths and weaknesses of scoring paraphrases using monolingual distributional similarity versus the bilingually calculated paraphrase probability. We show that monolingual cosine similarity calculated on large volumes of text ranks bilingually-extracted paraphrases better than the paraphrase probability originally defined by Bannard and Callison-Burch (2005). While our current implementation shows improvement mainly in grammaticality, other contextual features are expected to enhance the meaning preservation of paraphrases. We also show that monolingual scores can provide a reasonable threshold for picking out high precision paraphrases.

## 2 Related Work

### 2.1 Paraphrase Extraction from Bitexts

Bannard and Callison-Burch (2005) proposed identifying paraphrases by pivoting through phrases in a bilingual parallel corpora. Figure 1 illustrates their paraphrase extraction process. The *target* phrase, e.g. *thrown into jail*, is found in a German-English parallel corpus. The corresponding foreign phrase (*festgenommen*) is identified using word alignment and phrase extraction techniques from phrase-based statistical machine translation (Koehn et al., 2003). Other occurrences of the foreign phrase in the parallel corpus may align to a distinct English phrase, such as *jailed*. As the original phrase occurs several times and aligns with many different foreign phrases, each of these may align to a variety of other English paraphrases. Thus, *thrown into jail* not only paraphrases as *jailed*, but also as *arrested*, *detained*, *imprisoned*, *incarcerated*, *locked up*, and so on. Bad paraphrases, such as *maltreated*, *thrown*, *cases*, *custody*, *arrest*, and *protection*, may also arise due to poor word alignment quality and other factors.



*Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, pages 33–42,
Edinburgh, Scotland, UK, July 31, 2011. 

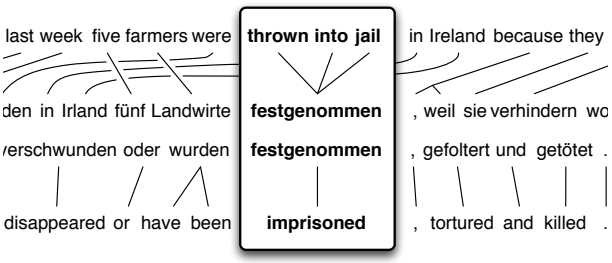


Figure 1: Using a bilingual parallel corpus to extract paraphrases.

Bannard and Callison-Burch (2005) defined a paraphrase probability to rank these paraphrase candidates, as follows:

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

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

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

$$\approx \sum_f p(e_2|f)p(f|e_1) \quad (4)$$

where $p(e_2|e_1)$ is the paraphrase probability, and $p(e|f)$ and $p(f|e)$ are translation probabilities from a statistical translation model.

Anecdotally, this paraphrase probability sometimes seems unable to discriminate between good and bad paraphrases, so some researchers disregard it and treat the extracted paraphrases as an unsorted set (Snover et al., 2010). Callison-Burch (2008) attempts to improve the ranking by limiting paraphrases to be the same syntactic type.

We attempt to rerank the paraphrases using other information. This is similar to the efforts of Zhao et al. (2008), who made use of multiple resources to derive feature functions and extract paraphrase tables. The paraphrase that maximizes a log-linear combination of various feature functions is then selected as the optimal paraphrase. Feature weights in the model are optimized by minimizing a *phrase substitution error rate*, a measure proposed by the authors, on a development set.

### 2.2 Monolingual Distributional Similarity

Prior work has explored the acquisition of paraphrases using distributional similarity computed from monolingual resources, such as in the DIRT results of Lin and Pantel (2001). In these models, phrases are judged to be similar based on the cosine distance of their associated context vectors. In some cases, such as by Lin and Pantel, or the seminal work of Church and Hanks (1991), distributional context is defined using frequencies of words appearing in various syntactic relations with other lexical items. For example, the nouns *apple* and *orange* are contextually similar partly because they both often appear as the object of the verb *eat*. While syntactic contexts provide strong evidence of distributional preferences, it is computationally expensive to parse very large corpora, so it is also common to represent context vectors with simpler representations like adjacent words and n-grams (Lapata and Keller, 2005; Bhagat and Ravichandran, 2008; Lin et al., 2010; Van Durme and Lall, 2010). In these models, *apple* and *orange* might be judged similar because both tend to be one word to the right of *some*, and one to the left of *juice*.

Here we calculate distributional similarity using a web-scale n-gram corpus (Brants and Franz, 2006; Lin et al., 2010). Given both the size of the collection, and that the n-grams are sub-sentential (the n-grams are no longer than 5 tokens by design), it was not feasible to parse, which led to the use of n-gram contexts. Here we use adjacent unigrams. For each phrase $x$ we wished to paraphrase, we extracted the context vector of $x$ from the n-gram collection as such: every (n-gram, frequency) pair of the form: ($ax$, $f$), or ($xb$, $f$), gave rise to the (feature, value) pair: ($w_{i-1}$=$a$, $f$), or ($w_{i+1}$=$b$, $f$), respectively. In order to scale to this size of a collection, we relied on Locality Sensitive Hashing (LSH), as was done previously by Ravichandran et al. (2005) and Bhagat and Ravichandran (2008). To avoid computing feature vectors explicitly, which can be a memory intensive bottleneck, we employed the online LSH variant described by Van Durme and Lall (2010).

This variant, based on the earlier work of Indyk and Motwani (1998) and Charikar (2002), approximates the cosine similarity between two feature vectors based on the Hamming distance in a reduced bitwise representation. In brief, for the feature vectors $\vec{u}$, $\vec{v}$, each of dimension $d$, then the cosine similarity is defined as: $\frac{\vec{u}\cdot\vec{v}}{|\vec{u}||\vec{v}|}$. If we *project* $\vec{u}$ and $\vec{v}$ through a $d$ by $b$ random matrix populated with draws from

| *huge amount of* | | |
|---|---|---|
| **BiP** | **SyntBiP** | **BiP-MonoDS** |
| *large number of*, .33 | *large number of*, .38 | *huge amount of*, 1.0 |
| *in large numbers*, .11 | *great number of*, .09 | *large quantity of*, .98 |
| *great number of*, .08 | *huge amount of*, .06 | *large number of*, .98 |
| *large numbers of*, .06 | *vast number of*, .06 | *great number of*, .97 |
| *vast number of*, .06 | | *vast number of*, .94 |
| *huge amount of*, .06 | | *in large numbers*, .10 |
| *large quantity of*, .03 | | *large numbers of*, .08 |

Table 1: Paraphrases for *huge amount of* according to the bilingual pivoting (BiP), syntactic-constrainted bilingual pivoting (SyntBiP) translation score and the monolingual similarity score via LSH (MonoDS), ranked by corresponding scores listed next to each paraphrase. Syntactic type of the phrase is [JJ+NN+IN].

$N(0,1)$, then we convert our feature vectors to *bit signatures* of length $b$, by setting each bit of the signature conditioned on whether or not the respective projected value is greater than or equal to 0. Given the bit signatures $h(\vec{u})$ and $h(\vec{v})$, we approximate cosine with the formula: $\cos(\frac{D(h(\vec{u}),h(\vec{v}))}{b}\pi)$, where $D()$ is Hamming distance.

## 3 Ranking Paraphrases

We use several different methods to rank candidate sets of paraphrases that are extracted from bilingual parallel corpora. Our three scoring methods are:

- **MonoDS** – monolingual distributional similarity calculated over the Google n-gram corpus via LSH, as described in Section 2.2.
- **BiP** – bilingual pivoting is calculated as in Equation 4 following Bannard and Callison-Burch (2005). The translation model probabilities are estimated from a French-English parallel corpus.
- **SyntBiP** – syntactically-constrained bilingual pivoting. This refinement to BiP, proposed in Callison-Burch (2008), constrains paraphrases to be the same syntactic type as the original phrase in the pivoting step of the paraphrase table construction.

When we use MonoDS to re-score a candidate set, we indicate which bilingual paraphrase extraction method was used to extract the candidates as prefix, as in **BiP-MonoDS** or **SyntBiP-MonoDS**.

| *reluctant* | |
|---|---|
| **MonoDS**$_{hand-selected}$ | **BiP** |
| **willing*, .99 | *not*, .56 |
| *loath*, .98 | *unwilling*, .04 |
| **eager*, .98 | *reluctance*, .03 |
| *somewhat reluctant*, .98 | *reticent*, .03 |
| *unable*, .98 | *hesitant*, .02 |
| *denied access*, .98 | *reticent about*, .01 |
| *disinclined*, .98 | *reservations*, .01 |
| *very unwilling*, .97 | *reticence*, .01 |
| *conducive*, .97 | *hesitate*, .01 |
| *linked*, .97 | *are reluctant*, .01 |

Table 2: Ordered reranked paraphrase candidates for the phrase *reluctant* according to monolingual distributional similarity (MonoDS$_{hand-selected}$) and bilingual pivoting paraphrase (BiP) method. Two hand-selected phrases are labeled with asterisks.

### 3.1 Example Paraphrase Scores

Table 1 shows the paraphrase candidates for the phrase *huge amount of* along with the values for each of our three scoring methods. Although MonoDS does not explicitly impose syntactic restrictions, the syntactic structure of the paraphrase *in large numbers* contributes to the large difference in the left and right context of the paraphrase and of the original phrase. Hence, the paraphrase was assigned a low score of 0.098 as compared to other paraphrase candidates with the correct syntactic type. Note that the SyntBiP produced significantly fewer paraphrase candidates, since its paraphrase candidates must be the same syntactic type as the original phrase. Identity paraphrases are excluded for the rest of the discussion in this paper.

### 3.2 Susceptibility to Antonyms

Monolingual distributional similarity is widely known to conflate words with opposite meaning and has motivated a large body of prior work on antonym detection (Lin and Zhao, 2003; Lin and Pantel, 2001; Mohammad et al., 2008a; Mohammad et al., 2008b; Marneffe et al., 2008; Voorhees, 2008). In contrast, the antonyms of a phrase are rarely produced during pivoting of the BiP methods because they tend not to share the same foreign translations. Since the reranking framework proposed here begins with paraphrases acquired by the BiP methodol-

ogy, MonoDS can considerably enhance the quality of ranking while sidestepping the antonym problem that arises from using MonoDS alone.

To support this intuition, an example of a paraphrase list with inserted hand-selected phrases ranked by each reranking methods is shown in Table 2[1]. Hand-selected antonyms of *reluctant* are inserted into the paraphrase candidates extracted by BiP before they are reranked by MonoDS. This is analogous to the case without pre-filtering of paraphrases by BiP and all phrases are treated equally by MonoDS alone. BiP cannot rank these hand-selected paraphrases since, by construction, they do not share any foreign translation and hence their paraphrase scores are not defined. As expected from the drawbacks of monolingual-based statistics, *willing* and *eager* are assigned top scores by MonoDS, although good paraphrases such as *somewhat reluctant* and *disinclined* are also ranked highly. This illustrates how BiP complements the monolingual reranking technique by providing orthogonal information to address the issue of antonyms for MonoDS.

### 3.3 Implementation Details

For BiP and SyntBiP, the French-English parallel text from the Europarl corpus (Koehn, 2005) was used to train the paraphrase model. The parallel corpus was extracted from proceedings of the European parliament with a total of about 1.3 million sentences and close to 97 million words in the English text. Word alignments were generated with the Berkeley aligner. For SyntBiP, the English side of the parallel corpus was parsed using the Stanford parser (Klein and Manning, 2003). The translation models were trained with Thrax, a grammar extractor for machine translation (Weese et al., 2011). Thrax extracts phrase pairs that are labeled with complex syntactic labels following Zollmann and Venugopal (2006).

For MonoDS, the web-scale n-gram collection of Lin et al. (2010) was used to compute the monolingual distributional similarity features, using 512 bits per signature in the resultant LSH projection. Following Van Durme and Lall (2010), we implicitly represented the projection matrix with a *pool* of size 10,000. In order to expand the coverage of the candidates scored by the monolingual method, the LSH signatures are obtained only for the phrases in the union set of the phrase-level outputs from the original and from the syntactically constrained paraphrase models. Since the n-gram corpus consists of at most 5-gram and each distributional similarity feature requires a single neighboring token, the LSH signatures are generated only for phrases that are 4-gram or less. Phrases that didn't appear in the n-grams with at least one feature were discarded.

[1]Generating a paraphrase list by MonoDS alone requires building features for all phrases in the corpus, which is computationally impractical and hence, was not considered here.

## 4 Human Evaluation

The different paraphrase scoring methods were compared through a manual evaluation conducted on Amazon Mechanical Turk. A set of 100 test phrases were selected and for each test phrase, five distinct sentences were randomly sampled to capture the fact that paraphrases are valid in some contexts but not others (Szpektor et al., 2007). Judges evaluated the paraphrase quality through a substitution test: For each sampled sentence, the test phrase is substituted with automatically-generated paraphrases. The sentences and the phrases are drawn from the English side of the Europarl corpus. Judges indicated the amount of the original **meaning** preserved by the paraphrases and the **grammaticality** of the resulting sentences. They assigned two values to each sentence using the **5-point scales** defined in Callison-Burch (2008).

The 100 test phrases consisted of 25 unigrams, 25 bigrams, 25 trigrams and 25 4-grams. These 25 phrases were randomly sampled from the paraphrase table generated by the bilingual pivoting method, with the following restrictions:

- The phrase must have occurred at least 5 times in the parallel corpus and must have appeared in the web-scale n-grams.
- The size of the union of paraphrase candidates from BiP and SyntBiP must be 10 or more.

### 4.1 Calculating Correlation

In addition to their average scores on the 5-point scales, the different paraphrase ranking methods were quantitatively evaluated by calculating their correlation with human judgments. Their correlation is calculated using **Kendall's tau coefficient**, a

| Reranking Method | Meaning | Grammar |
|---|---|---|
| BiP | 0.14 | 0.04 |
| BiP-MonoDS | 0.14 | **0.24**‡ |
| SyntBiP | **0.19** | 0.08 |
| SyntBip-MonoDS | 0.15 | 0.22‡ |
| SyntBiP$_{matched}$ | 0.20 | 0.15 |
| SyntBiP$_{matched}$-MonoDS | 0.17 | 0.16 |
| SyntBiP* | **0.21** | 0.09 |
| SyntBiP-MonoDS* | 0.16 | **0.22**† |

Table 3: Kendall's Tau rank correlation coefficients between human judgment of meaning and grammaticality for the different paraphrase scoring methods. Bottom panel: SyntBiP$_{matched}$ is the same as SyntBiP except paraphrases must match with the original phrase in syntactic type. SyntBiP* and MonoDS* are the same as before except they share the same phrase support with SyntBiP$_{matched}$. (‡: MonoDS outperforms the corresponding BiP reranking at $p$-value $\leq$0.01, and † at $\leq$0.05)

common measure of correlation between two ranked lists. Kendall's tau coefficient ranges between -1 and 1, where 1 indicates a perfect agreement between a pair of ranked lists.

Since tied rankings occur in the human judgments and reranking methods, Kendall's tau b, which ignores pairs with ties, is used in our analysis. An overall Kendall's tau coefficient presented in the results section is calculated by averaging all Kendall's tau coefficients of a particular reranking method over all phrase-sentence combinations.

## 5 Experimental Results

### 5.1 Correlation

The Kendall's tau coefficients for the three paraphrase ranking methods are reported Table 3. A total of 100 phrases and 5 sentence per phrase are selected for the experiment, resulting in a maximum support size of 500 for Kendall's tau coefficient calculation. The overall sizes of support are 500, 335, and 304 for BiP, SyntBiP and SyntBiP$_{matched}$, respectively. The positive values of Kendall's tau confirm both monolingual and bilingual approaches for paraphrase reranking are positively correlated with human judgments overall. **For grammaticality, monolingual distributional similarity reranking correlates stronger with human judgments than bilingual pivoting methods.** For example, in the top panel, given a paraphrase table generated through bilingual pivoting, Kendall's tau for monolingual distributional similarity (BiP-MonoDS) achieves 0.24 while that of the bilingual pivoting ranking (BiP) is only 0.04. Similarly, reranking of the paraphrases extracted with syntactically-constrained bilingual pivoting shows a stronger correlation between SyntBiP-MonoDS and grammar judgments (0.22) than the SyntBiP (0.08). *This result further supports the intuition of distributional similarity being suitable for paraphrase reranking in terms of grammaticality.*

In terms of meaning preservation, **the Kendall's tau coefficient for MonoDS is often lower than the bilingual approaches**, suggesting that paraphrase probability from the bilingual approach correlates better with phrasal meaning than the monolingual metric. For instance, SyntBiP reaches a Kendall's tau of 0.19, which is a slightly stronger correlation than that of SyntBiP-MonoDS. Although paraphrase candidates were generated by bilingual pivoting, distributional similarity depends only on contextual similarity and does not guarantee paraphrases that match with the original meaning; whereas Bilingual pivoting methods are derived based on shared foreign translations which associate meaning.

In the bottom panel of Table 3, only paraphrases of the same syntactic type as the source phrase are included in the ranked list for Kendall's tau calculation. The phrases associated with these paraphrases are used for calculating Kendall's tau for the original reranking methods (labeled as SyntBiP* and SyntBiP-MonoDS*). Comparing only the bilingual methods across panels, syntactic matching increases the correlation of bilingual pivoting metrics with human judgments in grammaticality (e.g., 0.15 for SyntBiP$_{matched}$ and 0.08 for SyntBiP) but with only minimal effects on meaning. The maximum values in the bottom panel for both categories are roughly the same as that in the corresponding category in the upper panel ($\{$0.21,0.19$\}$ in meaning and $\{$0.22, 0.24$\}$ in grammar for lower and upper panels, respectively.) This suggests that syntactic type matching offers similar improvement in grammaticality as MonoDS, although syntactically-constrained approaches have more confined paraphrase coverage.

We performed a one-tailed sign test on the Kendall's Tau values across phrases to examine

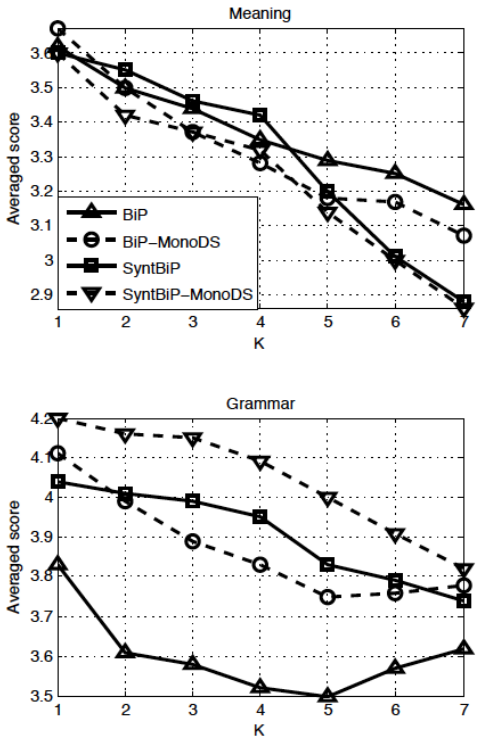


Figure 2: Averaged scores in the top K paraphrase candidates as a function of K for different reranking metrics. All methods performs similarly in meaning preservation, but SyntBiP-MonoDS outperforms other scoring methods in grammaticality, as shown in the bottom graph.

the statistical significance of the performance gain due to MonoDS. For grammaticality, except for the case of syntactic type matching (SyntBiP$_{matched}$), $p$-values are less than 0.05, confirming the hypothesis that MonoDS outperforms BiP. The $p$-value for comparing MonoDS and SyntBiP$_{matched}$ exceeds 0.05, agreeing with our conclusion from Table 3 that the two methods perform similarly.

### 5.2 Thresholding Using MonoDS Scores

One possible use for the paraphrase scores would be as a cutoff threshold where any paraphrases exceeding that value would be selected. Ideally, this would retain only high precision paraphrases.

To verify whether scores from each method correspond to human judgments for paraphrases extracted by BiP, human evaluation scores are averaged for meaning and grammar within each range of paraphrase score for BiP and approximate cosine distance for MonoDS, as shown in Table 4. The BiP paraphrase score bin sizes are linear in log scale.

| BiP Paraphrase Score | | | MonoDS LSH Score | | |
|---|---|---|---|---|---|
| **Region** | **M** | **G** | **Region** | **M** | **G** |
| 1.00 $\geq$ x $>$ 0.37 | 3.6 | 3.7 | 1 $\geq$ x $>$ 0.95 | 4.0 | 4.4 |
| 0.37 $\geq$ x $>$ 0.14 | 3.6 | 3.7 | 0.95 $\geq$ x $>$ 0.9 | 3.2 | 4.0 |
| 0.14 $\geq$ x $>$ 0.05 | 3.4 | 3.6 | 0.9 $\geq$ x $>$ 0.85 | 3.3 | 4.0 |
| 0.05 $\geq$ x $>$ 1.8e-2 | 3.4 | 3.6 | 0.85 $\geq$ x $>$ 0.8 | 3.3 | 4.0 |
| 1.8e-2 $\geq$ x $>$ 6.7e-3 | 3.4 | 3.6 | 0.8 $\geq$ x $>$ 0.7 | 3.2 | 3.9 |
| 6.7e-3 $\geq$ x $>$ 2.5e-3 | 3.2 | 3.7 | 0.7 $\geq$ x $>$ 0.6 | 3.3 | 3.8 |
| 2.5e-3 $\geq$ x $>$ 9.1e-4 | 3.0 | 3.6 | 0.6 $\geq$ x $>$ 0.5 | 3.1 | 3.7 |
| 9.1e-4 $\geq$ x $>$ 3.4e-4 | 3.0 | 3.8 | 0.5 $\geq$ x $>$ 0.4 | 3.1 | 3.6 |
| 3.4e-4 $\geq$ x $>$ 1.2e-4 | 2.6 | 3.6 | 0.4 $\geq$ x $>$ 0.3 | 3.1 | 3.5 |
| 1.2e-4 $\geq$ x $>$ 4.5e-5 | 2.7 | 3.6 | 0.3 $\geq$ x $>$ 0.2 | 2.9 | 3.4 |
| x $\leq$ 4.5e-5 | 2.5 | 3.7 | 0.2 $\geq$ x $>$ 0.1 | 3.0 | 3.3 |
| | | | 0.1 $\geq$ x $>$ 0 | 2.9 | 3.2 |

Table 4: Averaged human judgment scores as a function of binned paraphrase scores and binned LSH scores. MonoDS serves as much better thresholding score for extracting high precision paraphrases.

| MonoDS LSH | BiP Paraphrase Threshold | | |
|---|---|---|---|
| **Threshold** | $\geq$ 0.05 | $\geq$ 0.01 | $\geq$ 6.7e-3 |
| $\geq$ 0.9 | 4.2 / 4.4 | 4.1 / 4.4 | 4.0 / 4.4 |
| $\geq$ 0.8 | 4.0 / 4.3 | 3.9 / 4.3 | 3.9 / 4.2 |
| $\geq$ 0.7 | 3.9 / 4.1 | 3.8 / 4.2 | 3.8 / 4.1 |

Table 5: Thresholding using both the MonoDS and BiP scores further improves the average human judgment of Meaning / Grammar.

Observe that for the BiP paraphrase scores on the left panel, no trend on the averaged grammar scores across all score bins is present. While a mild correlation exists between the averaged meaning scores and the paraphrase scores, the top score region (1$>$ x $\geq$ 0) corresponds to merely an averaged value of 3.6 on a 5-point scale. Therefore, thresholding on BiP scores among a set of candidates would not guarantee accurate paraphrases in grammar or meaning.

On the right panel, MonoDS LSH scores on paraphrase candidates produced by BiP are uniformly higher in grammar than meaning across all score bins, similar to the correlation results in Table 3. The averaged grammar scores decreases monotonically and proportionally to the change in LSH values. With regard to meaning scores, the averaged values roughly correspond to the decrease of LSH values, implying distributional similarity correlates weakly with human judgment in the meaning preser-

vation of paraphrase. Note that the drop in averaged scores is the largest from the top bin ($1 \geq x > 0.95$) to the second bin ($0.95 \geq x > 0.9$) is the largest within both meaning and grammar. **This suggests that thresholding on top tiered MonoDS scores can be a good filter for extracting high precision paraphrases.** BiP scores, by comparison, are not as useful for thresholding grammaticality.

Additional performance gain attained by combining the two thresholding are illustrated in Table 5, where averaged meaning and grammar scores are listed for each combination of thresholding. At a threshold of 0.9 for MonoDS LSH score and 0.05 for BiP paraphrase score, the averaged meaning exceeds the highest value reported in Table 4, whereas the grammar scores reaches the value in the top bin in Table 4. General trends of improvement from utilizing the two reranking methods are observed by comparing Tables 4 and 5.

### 5.3 Top K Analysis

Figure 2 shows the mean human assigned score within the top K candidates averaged across all phrases. Compared across the two categories, meaning scores have lower range of score and a more uniform trend of decreasing values as K grows. In grammaticality, BiP clearly underperforms whereas the SyntBiP-MonoDS maintains the best score among all methods over all values of K. In addition, a slow drop-off up until K = 4 in the curve for SyntBiP-MonoDS implies that the quality of paraphrases remains relatively high going from top 1 to top 4 candidates.

In applications such as question answering or search, the order of answers presented is important because the lower an answer is ranked, the less likely it would be looked at by a user. Based on this intuition, the paraphrase ranking methods are evaluated using the maximum human judgment score among the top K candidates obtained by each method. As shown in Table 6, when only the top candidate is considered, the averaged score corresponding to the monolingual reranking methods are roughly the same as that to the bilingual methods in meaning, but as K grows, the bilingual methods outperforms the monolingual methods. In terms of grammaticality, scores associated with monolingual reranking methods are consistently higher than the bilingual methods but the difference tapers off as K increases. This suggests that when only limited top paraphrase candidates can be evaluated, MonoDS is likely to provide better quality of results.

| | | Reranking Method | | | |
|---|---|---|---|---|---|
| | **K** | BiP | BiP-MonoDS | SyntBiP | SyntBiP-MonoDS |
| **M** | 1 | 3.62 | 3.67 | 3.58 | 3.58 |
| | 3 | 4.13 | 4.07 | 4.13 | 4.01 |
| | 5 | 4.26 | 4.19 | 4.20 | 4.09 |
| | 10 | 4.39 | 4.30 | 4.25 | 4.23 |
| **G** | 1 | 3.83 | 4.11 | 4.04 | 4.23 |
| | 3 | 4.22 | 4.45 | 4.47 | 4.54 |
| | 5 | 4.38 | 4.54 | 4.55 | 4.62 |
| | 10 | 4.52 | 4.62 | 4.63 | 4.67 |

Table 6: Average of the *maximum* human evaluation score from top K candidates for each reranking method. Support sizes for BiP- and SyntBiP-based metrics are 500 and 335, respectively. (M = Meaning, G = Grammar)

## 6 Detailed Examples

### 6.1 MonoDS Filters Bad BiP Paraphrases

The examples in the top panel of Table 7 illustrates a few disadvantages of the bilingual paraphrase scores and how monolingual reranking complements the bilingual methods. Translation models based on bilingual corpora are known to suffer from misalignment of the parallel text (Bannard and Callison-Burch, 2005), producing incorrect translations that propagate through in the paraphrase model. This issue is exemplified in the phrase pairs {*considerable changes, caused quite*}, {*always declared, always been*}, and {*significantly affected, known*} listed Table 7. The paraphrases are clearly unrelated to the corresponding phrases as evident from the low rankings from human judges. Nonetheless, they were included as candidates likely due to misalignment and were ranked relatively high by BiP metric. For example, *considerable changes* was aligned to *modifie considérablement* correctly. However, due to a combination of loose translations and difficulty in aligning multiple words that are spread out in a sentence, the French phrase was inaccurately matched with *caused quite* by the aligner, inducing a bad paraphrase. Note that in these cases LSH produces the results that agrees with the human rankings.

| | | Ranking | | | | |
|---|---|---|---|---|---|---|
| **Phrase** | **Paraphrase** | **Size$_{pool}$** | **Meaning** | **Grammar** | **BiP** | **BiP-MonoDS** |
| *significantly affected* | *known* | 20 | 19 | 18.5 | 1 | 17 |
| *considerable changes* | *caused quite* | 23 | 23 | 23 | 2.5 | 23 |
| *always declared* | *always been* | 20 | 20 | 20 | 2 | 13 |
| *hauled* | *delivered* | 23 | 7 | 5.5 | 21.5 | 5.0 |
| *fiscal burden*† | *taxes* | 18 | 13.5 | 18 | 6 | 16 |
| *fiscal burden*† | *taxes* | 18 | 2 | 8 | 6 | 16 |
| *legalise* | *legalize* | 23 | 1 | 1 | 10 | 1 |
| *to deal properly with* | *address* | 35 | 4.5 | 5.5 | 4 | 29.5 |
| *you have just stated* | *you have just suggested* | 31 | 13.5 | 8.5 | 4 | 30 |

Table 7: Examples of phrase pair rankings by different reranking methods and human judgments in terms of meaning and grammar. Higher rank (smaller numbers) corresponds to more favorable paraphrases by the associated metric. (†: Phrases are listed twice to show the ranking variation when substitutions are evaluated in different sentences.)

### 6.2 Context Matters

Occasionally, paraphrases are context-dependent, meaning the relevance of the paraphrase depends on the context in a sentence. Bilingual methods can capture limited context through syntactic constraints if the POS tags of the paraphrases and the sentence are available, while the distributional similarity metric, in its current implementation, is purely based on the pattern of co-occurrence with neighboring context n-grams. As a result, LSH scores should be slightly better at gauging the paraphrases defined by context, as suggested by some examples in Table 7. The phrase pair {*hauled, delivered*} differ slightly in how they describe the manner that an object is moved. However, in the context of the following sentence, they roughly correspond to the same idea:

*countries which do not comply with community legislation should be **hauled** before the court of justice and i think mrs palacio will do so .*

As a result, out of 23 candidates, human judges ranked *delivered* 7 and 5.5 for meaning and grammar, respectively. The monolingual-based metric also assigns a higher rank to the paraphrase while BiP puts it near the lowest rank.

Another example of context-dependency is the phrase pair {*fiscal burden, taxes*}, which could have some foreign translations in common. The original phrase appears in the following sentence:

*... the member states can reduce the **fiscal burden** consisting of taxes and social contributions .*

The paraphrase candidate *taxes* is no longer appropriate with the consideration of the context surrounding the original phrase. As such, *taxes* received rankings of 13.5, 18 and 16 out of 18 for meaning, grammar, and MonoDS, respectively, whereas BiP assigns a 6 to the paraphrase. The same phrase pair but a different sentence, the context induces opposite effects on the paraphrase judgments, where the paraphrase received 2 and 8 in the two categories as shown in Table 7:

*the economic data for our eu as regards employment and economic growth are not particularly good , and , in addition , the **fiscal burden** in europe , which is to be borne by the citizen , has reached an all-time high of 46 % .*

Hence, distributional similarity offers additional advantages over BiP only when the paraphrase appears in a context that also defines most of the non-zero dimensions of the LSH signature vector.

The phrase pair {*legalise*, *legalize*} exemplifies the effect of using different corpora to train 2 paraphrase reranking models as shown in Table 7. Meaning, grammar and MonoDS all received top rank out of all paraphrases, whereas BiP ranks the paraphrase 10 out of 23. Since the BiP method was trained with Europarl data, which is dominated by British English, BiP fails to acknowledge the American spelling of the same word. On the other hand, distributional similarity feature vectors were extracted from the n-gram corpus with different variations of English, which was informative for paraphrase ranking. This property can be exploited for adaptation of specific domain of paraphrases selection.

### 6.3 Limitations of MonoDS Implementation

While the monolingual distributional similarity shows promise as a paraphrase ranking method, there are a number of additional drawbacks associated with the implementation.

The method is currently limited to phrases with up to 4 contiguous words that are present in the n-gram corpus for LSH feature vector extraction. Since cosine similarity is a function of the angle between 2 vectors irrespective of the vector magnitudes, thresholding on low occurrences of higher n-grams in the corpus construction causes larger n-grams to suffer from feature sparsity and be susceptible to noise. A few examples from the experiment demonstrate such scenario. For a phrase *to deal properly with*, a paraphrase candidate *address* receives rankings of 4.5, 5.5 and 4 out of 35 for meaning, grammar and BiP, respectively, it is ranked 29.5 by BiP-MonoDS. The two phrases are expected to have similar neighboring context in regular English usage, but it might be misrepresented by the LSH feature vector due to the lack of occurrences of the 4-gram in the corpus.

Another example of how sparsity affects LSH feature vectors is the phrase *you have just stated*. An acceptable paraphrase *you have just suggested* was ranked 13.5, 8.5 and 6.5 out of a total of 31 candidates by meaning, grammar and BiP, respectively, but MonoDS only ranks it at 30. The cosine similarity between the phrases are 0.05, which is very low. However, the only tokens that differentiate the 4-gram phrases, i.e. {*stated*,*suggested*}, have a similarity score of 0.91. This suggests that even though the additional words in the phrase don't alter the meaning significantly, the feature vectors are misrepresented due to the sparsity of the 4-gram. This highlights a weakness of the current implementation of distributional similarity, namely that context within a phrase is not considered for larger n-grams.

## 7 Conclusions and Future Work

We have presented a novel paraphrase ranking metric that assigns a score to paraphrase candidates according to their monolingual distributional similarity to the original phrase. While bilingual pivoting-based paraphrase models provide wide coverage of paraphrase candidates and syntactic constraints on the model confines the structural match, additional contextual similarity information provided by monolingual semantic statistics increases the accuracy of paraphrase ranking within the target language. Through a manual evaluation, it was shown that monolingual distributional scores strongly correlate with human assessment of paraphrase quality in terms of grammaticality, yet have minimal effects on meaning preservation of paraphrases.

While we speculated that MonoDS would improve both meaning and grammar scoring for paraphrases, we found in the results that only grammaticality was improved from the monolingual approach. This is likely due to the choice of how context is represented, which in this case is only single neighboring words. A consideration for future work to enhance paraphrasal meaning preservation would be to explore other contextual representations, such as syntactic dependency parsing (Lin, 1997), mutual information between co-occurences of phrases Church and Hanks (1991), or increasing the number of neighboring words used in n-gram based representations.

In future work we will make use of other complementary bilingual and monolingual knowledge sources by combining other features such as n-gram length, language model scores, etc. One approach would be to perform minimum error rate training similar to Zhao et al. (2008) in which linear weights of a feature function for a set of paraphrases candidate are trained iteratively to minimize the phrasal-substitution-based error rate. Instead of phrasal substitution in Zhao's method, quantitative measure of correlation with human judgment can be used as the objective function to be optimized during training. Other techniques such as SVM-rank (Joachims, 2002) may also be investigated for aggregating results from multiple ranked lists.

## 8 Acknowledgements

Thanks to Courtney Napoles for advice regarding a pilot version of this work. Thanks to Jonathan Weese, Matt Post and Juri Ganitkevitch for their assistance with Thrax. This research was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme), and by the NSF under grant IIS-0713448. Opinions, interpretations, and conclusions are the authors' 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*.

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL-HLT*.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*.

Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of STOC*.

Kenneth Church and Patrick Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 6(1):22–29.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of STOC*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. *Advances in NIPS*, 15:3–10.

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*.

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1).

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Dekang Lin and Shaojun Zhao. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI-03*, pages 1492–1493.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of LREC*.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL*.

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

Marie-Catherine De Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL 2008*.

Saif Mohammad, Bonnie Dorr, and Graeme Hirst. 2008a. Computing word-pair antonymy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 982–991. Association for Computational Linguistics.

Saif Mohammad, Bonnie J. Dorr, Melissa Egan, Nitin Madnani, David Zajic, and Jimmy Lin. 2008b. Multiple alternative sentence compressions and word-pair antonymy for automatic text summarization and recognizing textual entailment.

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*.

Marius Pasca and Peter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In *Proceedings of IJCNLP*, pages 119–130.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, pages 142–149.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized Algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering. In *Proceedings of ACL*.

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.

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

Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *Proceedings of ACL, Short Papers*.

Ellen M. Voorhees. 2008. Contradictions and justifications: Extensions to the textual entailment task.

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. EMNLP 2011 - Workshop on statistical machine translation.

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

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of WMT06*.