The Multilingual Paraphrase Database

Juri Ganitkevitch[1] **Chris Callison-Burch**[2]

[1] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD

[2] Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA

Abstract

We release a massive expansion of the paraphrase database (PPDB) that now includes a collection of paraphrases in 23 different languages. The resource is derived from large volumes of bilingual parallel data. Our collection is extracted and ranked using state of the art methods. The multilingual PPDB has over a billion paraphrase pairs in total, covering the following languages: Arabic, Bulgarian, Chinese, Czech, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portugese, Romanian, Russian, Slovak, Slovenian, and Swedish.

Keywords: paraphrasing, multilingual resources, lexical resources

1. Introduction

Paraphrases are differing ways of expressing the same meaning within a single language (Bhagat and Hovy, 2013; Vila et al., 2014). They have proven useful for a wide variety of natural language processing applications. For instance, in multi-document summarization, paraphrase detection is used to collapse redundancies (Barzilay et al., 1999). Paraphrase generation can be used for query expansion in information retrieval and question answering systems (McKeown, 1979; Ravichandran and Hovy, 2002; Riezler et al., 2007). Paraphrases also allow for more flexible matching of system output against human references for tasks like evaluating machine translation or evaluating automatic summarization (Zhou et al., 2006; Madnani et al., 2007; Snover et al., 2010). Paraphrases have been used to perform text normalization on the irregular writing style used on Twitter (Xu et al., 2013; Ling et al., 2013). Furthermore, paraphrases are helpful for natural language understanding, including tasks like recognizing textual entailment (Bosma and Callison-Burch, 2007; Androutsopoulos and Malakasiotis, 2010) and semantic parsing (Berant and Liang, 2014).

Different algorithms have been developed for extracting paraphrases from different types of data (Madnani and Dorr, 2010). Various types of data have been used, including regular monolingual texts (Lin and Pantel, 2001; Bhagat and Ravichandran, 2008), comparable corpora (Barzilay and Lee, 2003; Dolan et al., 2004; Chen and Dolan, 2011), and bilingual parallel data (Quirk et al., 2004; Bannard and Callison-Burch, 2005; Madnani et al., 2007; Ganitkevitch et al., 2011; Ganitkevitch et al., 2013). Several of these research efforts have made paraphrase collections available as a resource, in addition to describing their algorithms. Paraphrase resources include the DIRT database which contains 12 million paraphrase rules (Lin and Pantel, 2001), the MSR paraphrase phrase table which has 13 million rules (Dolan et al., 2004), and the paraphrase database (PPDB) which has 170 million paraphrase rules (Ganitkevitch et al., 2013). Typically, large paraphrase resources are only available for English. However, the initial version of PPDB was released for English and one other language (Spanish).

There have been several efforts to extract non-English paraphrases for use in natural language processing applications. For example, paraphrase tables across five different languages were extracted as a part of METEOR-NEXT, a multilingual extension of the METEOR metric for machine translation evaluation (Denkowski and Lavie, 2010). Similarly, automatically extracted paraphrases in Arabic and Chinese have been used to improve English-Arabic (Denkowski et al., 2010) and Chinese-Japanese (Zhang and Yamamoto, 2002; Zhang and Yamamoto, 2005) machine translation systems. Other individualized efforts have sought to create paraphrase resources for single languages, like Mizukami et al. (2014)'s efforts to create a Japanese version of PPDB. While achieving good results, many of the paraphrase collections used in these efforts have remained unavailable to the community.

Here we release a massively expanded version of PPDB that includes collections of paraphrases for 21 additional languages: Arabic, Bulgarian, Chinese, Czech, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portugese, Romanian, Russian, Slovak, Slovenian, and Swedish. We extract paraphrases for each of these languages by pivoting through bitexts. Our bitext collection encompasses over 100 million sentence pairs between English and the foreign languages. The multilingual paraphrase database is freely available from `paraphrase.org`. We expect our non-English paraphrases to be useful for a variety of multilingual natural language processing applications.

2. Paraphrase Extraction

We extract paraphrases from bilingual parallel corpora (bitexts). Although bitexts are more commonly used as training data for statistical machine translation, Bannard and Callison-Burch (2005) showed how they could be leveraged to find meaning-equivalent phrases in a single language by "pivoting" over a shared translation in another language. Although Bannard and Callison-Burch extracted equivalent English expressions by pivoting over shared foreign phrases, it is simple to see how the method can be used to find equivalent expressions in other languages:

Two expressions in language F, f_1 and f_2, that translate to a shared expression e in another language E can be assumed to have the same meaning. We can thus find

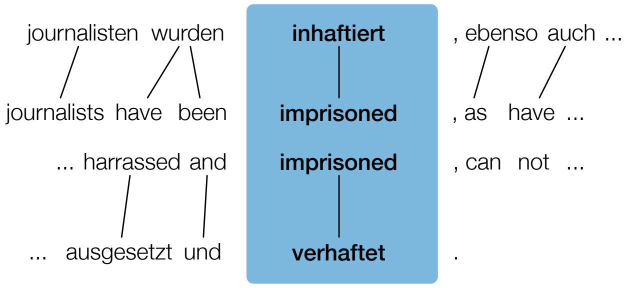

Figure 1: German paraphrases are extracted by pivoting over a shared English translation.

paraphrases of a German phrase like *inhaftiert* by pivoting over a shared English translation like *imprisoned* and extract German paraphrase pair ⟨*inhaftiert, verhaftet*⟩, as illustrated in Figure 1.

Since *inhaftiert* can have many possible translations, and since each of those can map back to many possible German phrases, we extract not only *verhaftet* as a paraphrase, but also *eingesperrt, festgenommen, eingekerkert, festgehalten, festnahmen, festnahme, statt, stattfinden, gefangenen, gefangengenommen, haft, innerhalb,* and others. Bannard and Callison-Burch (2005) propose a paraphrase probability for sorting this set of paraphrases, using the translation probabilities derived from the bitext:

$$p(f_2|f_1) \approx \sum_e p(f_2|e)p(e|f_1). \tag{1}$$

Paraphrases need not be extracted from a single pivot language. They can be obtained from multiple bitexts where the language of interest is contained on one side of the parallel corpus. Thus, instead of extracting German paraphrases just by pivoting over English, we could extract additional paraphrases from a German-French or a German-Spanish bitext. Although it is easy to construct parallel corpora for all pairs of languages in the European Union using existing resources like the Europarl parallel corpus (Koehn, 2005) or the JRC corpus (Steinberger et al., 2006), we only pivot over English for this release of the multilingual PPDB.

The reason that we limit ourselves to pivoting over English, is that we extend the bilingual pivoting method to incorporate syntactic information. Abundant NLP resources, such as statistical parsers, are available for English. By using annotations from the English side of the bitext, we are able to create syntactic paraphrases for languages for which we do not have syntactic parsers.

Syntactic information can be incorporated into the paraphrase process in a variety of ways. Callison-Burch (2008) showed that constraining paraphrases to be the same syntactic type as the original phrase significantly improved their quality. Ganitkevitch et al. (2011) showed how paraphrases could be represented using synchronous context free grammars (SCFGs).

We project the English syntax onto the foreign sentence via the automatic word alignments. The notion of projecting syntax across aligned bitexts has been explored for bootstrapping parsers (Hwa et al., 2005). The method that we use to find the syntactic labels for the foreign phrases is described in Zollmann and Venugopal (2006) and Weese et al. (2011). Only the English side of each parallel corpus needs to be parsed, which we do with the Berkeley Parser (Petrov et al., 2006).

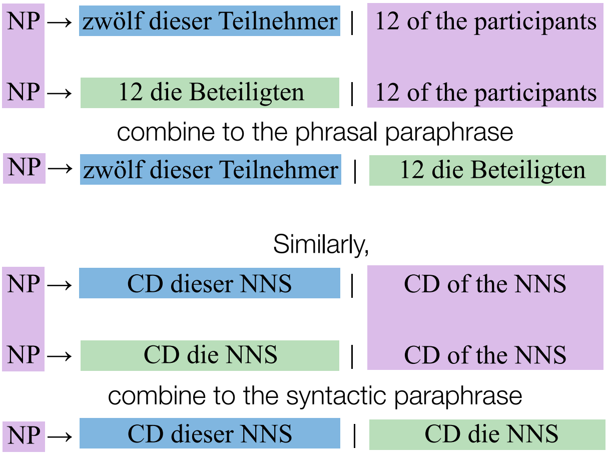
NP → zwölf dieser Teilnehmer | 12 of the participants
NP → 12 die Beteiligten | 12 of the participants
combine to the phrasal paraphrase
NP → zwölf dieser Teilnehmer | 12 die Beteiligten
Similarly,
NP → CD dieser NNS | CD of the NNS
NP → CD die NNS | CD of the NNS
combine to the syntactic paraphrase
NP → CD dieser NNS | CD die NNS

Figure 2: In addition to extracting lexical and phrasal paraphrases, we also extract syntactic paraphrases. These have nonterminal symbols that act as slots that can be filled by other paraphrases that match that syntactic type. The syntactic labels are drawn from parse trees of the English sentences in our bitexts.

With this addition, each of the paraphrase databases in our multilingual set have three types of paraphrases:

- **Lexical paraphrases** – single word paraphrases or synonyms.
- **Phrasal paraphrases** – multi-word paraphrases, including cases where a single word maps onto a multi-word paraphrase and many-to-many paraphrases.
- **Syntactic paraphrases** – paraphrase rules that contain a placeholder symbol. These allow any paraphrase that matches that syntactic types of placeholder symbol to be substituted into that site.

Figure 2 shows how a phrasal paraphrase can be generalized into a syntactic paraphrase by replacing words and phrases that are themselves paraphrases with appropriate nonterminal symbols.

The syntactic paraphrases can be used in conjunction with our Joshua decoder (Post et al., 2013) for monolingual text-to-text (T2T) generation applications, like sentence compression (Ganitkevitch et al., 2011; Napoles et al., 2011). This opens up the possibilities of developing new natural language generation (NLG) applications for the languages in our new PPDB release.

3. Paraphrase Scores

Each of the paraphrase entries in PPDB has a set of associated feature functions. These may be useful for ranking the quality of the paraphrases themselves. For instance, Zhao et al. (2008) proposed a log-linear model for scoring paraphrases instead of Bannard and Callison-Burch's

[VBZ] ||| *arbeitet* ||| *agiert* ||| Abstract=0 Adjacent=0 CharCountDiff=-2 CharLogCR=-0.28768 ContainsX=0 GlueRule=0 Identity=0 Lex(e|f)=5.37270 Lex(f|e)=7.24933 Lexical=1 LogCount=0.69315 Monotonic=1 PhrasePenalty=1 RarityPenalty=0.00248 SourceTerminalsButNoTarget=0 SourceWords=1 TargetTerminalsButNoSource=0 TargetWords=1 UnalignedSource=0 UnalignedTarget=0 WordCountDiff=0 WordLenDiff=-2.00000 WordLogCR=0 p(LHS|e)=2.03377 p(LHS|f)=1.54623 p(e|LHS)=11.70324 p(e|f)=5.90426 p(e|f,LHS)=6.02217 p(f|LHS)=9.26713 p(f|e)=3.95569 p(f|e,LHS)=3.58605

Table 1: An example paraphrase rule for German. The four fields are the left hand size nonterminal, the phrase, the paraphrase and the features associated with the rule.

paraphrase probability. Malakasiotis and Androutsopoulos (2011) re-ranked paraphrases using an maximum entropy classifier and a support vector regression ranker to set weights for features associated with a set of paraphrases, optimizing to a development set that was manually labeled with quality scores. Ganitkevitch et al. (2011) used a variety of paraphrase features and optimized their weights through minimum error rate training (Och, 2003) on a T2T generation task.

Each of the language-specific paraphrase grammars is a collection of paraphrase rules. Formally, these rules comprise a synchronous context free grammar (SCFG). Table 1 gives an example paraphrase rule for German. The entry contains 4 fields separated by |||. The first field is the left-hand side (LHS) nonterminal symbol that dominates the of the SCFG rule. The second field is the original phrase (which can be a mix of words and nonterminal symbols). The third field is the paraphrase. If the paraphrase is a syntactic rule it will have an identical set of nonterminal symbols as the original phrase, but they can appear in different orders. The mapping between nonterminal symbols is given with indices like $[NP, 1]$ and $[NP, 2]$. The fourth field is a collection of features associated with the rule.

The features we estimate for each paraphrase rule are related to features typically used in machine translation systems. As such, we follow traditional SMT notation in designating the input phrase as f and its paraphrase as e. To estimate the count- and probability-based features, we rely on Equation 1. Following the log-linear feature model, the resulting (un-normalized) probablity estimates, like $p(e|f)$, are stored as their negative logarithm $-\log p(e|f)$. In detail, the 31 features we compute for a PPDB rule are:

- Abstract – a binary feature that indicates whether the rule is composed exclusively of nonterminal symbols.

- Adjacent – a binary feature that indicates whether rule contains adjacent nonterminal symbols.

- CharCountDiff – a feature that calculates the difference in the number of characters between the phrase and the paraphrase. This feature is used for our sentence compression experiments (Napoles et al., 2011).

- CharLogCR – the log-compression ratio in characters, $\log \frac{chars(f_2)}{chars(f_1)}$, another feature used in sentence compression.

- ContainsX – a binary feature that indicates whether the nonterminal symbol X is used in this rule. X is the symbol used in Hiero grammars (Chiang, 2007), and is sometimes used by our syntactic SCFGs when we are unable to assign a linguistically motivated nonterminal.

- GlueRule – a binary feature that indicates whether this is a glue rule. Glue rules are treated specially by the Joshua decoder (Post et al., 2013). They are used when the decoder cannot produce a complete parse using the other grammar rules.

- Identity – a binary feature that indicates whether the phrase is identical to the paraphrase.

- Lex(e|f) – the "lexical translation" probability of the paraphrase given the original phrase. This feature is estimated as defined by Koehn et al. (2003).

- Lex(f|e) – the lexical translation probability of phrase given the paraphrase.

- Lexical – a binary feature that says whether this is a single word paraphrase.

- LogCount – the log of the frequency estimate for this paraphrase pair.

- Monotonic – a binary feature that indicates whether multiple nonterminal symbols occur in the same order (are monotonic) or if they are re-ordered.

- PhrasePenalty – this feature is used by the decoder to count how many rules it uses in a derivation. Turning helps it to learn to prefer fewer longer phrases, or more shorter phrases. The value of this feature is always 1.

- RarityPenalty – this feature marks rules that have only been seen a handful of times. It is calculated as $\exp(1 - c(e, f))$, where $c(e, f)$ is the estimate of the frequency of this paraphrase pair.

- SourceTerminalsButNoTarget – a binary feature that fires when the phrase contains terminal symbols, but the paraphrase contains no terminal symbols.

- SourceWords – the number of words in the original phrase.

- TargetTerminalsButNoSource – a binary feature that fires when the paraphrase contains terminal symbols but the original phrase only contains nonterminal symbols.

- TargetWords – the number of words in the paraphrase.

- UnalignedSource – a binary feature that fires if there are any words in the original phrase that are not aligned to any words in the paraphrase.

- UnalignedTarget – a binary feature that fires if there are any words in the paraphrase that are not aligned to any words in the original phrase.

- WordCountDiff – the difference in the number of words in the original phrase and the paraphrase. This feature is used for our sentence compression experiments.

- WordLenDiff – the difference in average word length between the original phrase and the paraphrase. This feature is useful for text compression and simplification experiments.

- WordLogCR – the log-compression ratio in words, estimated as $\log \frac{words(e)}{words(f)}$. This feature is used for our sentence compression experiments.

- p(LHS|e) – the (negative log) probability of the lefthand side nonterminal symbol given the paraphrase.

- p(LHS|f) – the (negative log) probability of the lefthand side nonterminal symbol given the original phrase.

- p(e|LHS) – the (negative log) probability of the paraphrase given the lefthand side nonterminal symbol (this is typically a very low probability).

- p(e|f) – the paraphrase probability of the paraphrase given the original phrase, as defined in Equation 1. This is given as a negative log value.

- p(e|f,LHS) – the (negative log) probability of paraphrase given the the lefthand side nonterminal symbol and the original phrase.

- p(f|LHS) – the (negative log) probability of original phrase given the the lefthand side nonterminal (this is typically a very low probability).

- p(f|e) – the paraphrase probability of the original phrase given the paraphrase, as defined in Equation 1. This is given as a negative log value.

- p(f|e,LHS) – the (negative log) probability of original phrase given the the lefthand side nonterminal symbol and the paraphrase.

To sort each language version of PPDB, we combine a subset of the features as follows: SCORE = p(e|f) + p(f|e) + p(e|f,lhs) + p(f|e,lhs) + 100·RarityPenalty + 0.3·p(lhs|e) + 0.3·p(lhs|f). The selection of features and the values for their weights are chosen in an ad hoc fashion, based on our intuitions about which features seem to be useful for sorting higher quality paraphrases from lower quality paraphrases. A more principled approach would be to collect a set of judgments about the quality of a random sample of the paraphrases, and then use logistic regression to fit the weights to the human judgments, for instance, in a similar fashion to (Malakasiotis and Androutsopoulos, 2011). We leave that task to users of our resource. We provide the full feature set so that users can re-sort the resource to fit native-speaker judgments or to fit the needs of a specific NLP task.

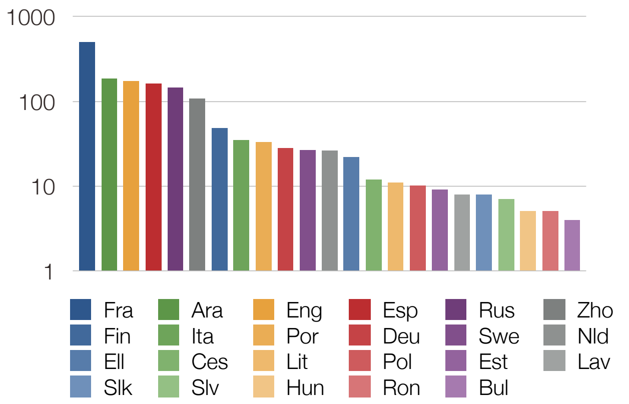

Figure 3: An overview of paraphrase collection size per language, measured in millions of paraphrase pairs.

4. Analysis

4.1. Resource Size

We extract significantly different numbers of paraphrases for each the languages. The number of paraphrases is roughly proportional to the size of the bitext that was used to extract the paraphrases for that language. Figure 3 sorts the languages in order of how many paraphrases we extract for them. Unsurprisingly, we observe a large difference in size between the French, Arabic, and Chinese paraphrase sets, and the others. This is due to the comparatively large bilingual corpora that we used for the three languages, versus the smaller bitexts that we used for the other languages (see Table 2). Table 3 gives a detailed breakdown of the number of each kind of paraphrases (lexical, phrasal, syntactic) that we have extracted for each language.

In the future, we hope to incorporate larger parallel corpora, taking advantage of improved bitext mining techniques (Smith et al., 2013), and we hope to incorporate other pivot languages in addition to English. We expect that these expansions will further improve the coverage and quality of our paraphrase sets for the lower-resource languages.

4.2. Resource Partitioning

We recognize that our paraphrase collection can feel unwieldy large, especially for Arabic, French, Chinese, Spanish, and Russian. We therefore divide the sets into different sizes. The are named by size: S (small), M (medium), L (large), XL (extra large), XXL (double extra large), and XXXL (Royal with cheese). Each step up in size is designed to roughly double the number of paraphrases across each type: lexical, phrasal, and syntactic. The larger sizes subsume the smaller sizes.

Before partitioning, we sort the paraphrases according to the score given at the end of Section 3. This helps to ensure that the higher quality paraphrases are included in the smaller sized sets. The the larger sized sets include these high precision paraphrases, but also include paraphrases that are not as high quality (but which do offer better coverage or higher recall). The choice of which size to use will depend on the needs of a particular application.

Language	Sentence Pairs	Foreign Words	English Words	Corpora
Arabic	9,542,054	205,508,319	204,862,233	GALE
Bulgarian	406,934	9,306,037	9,886,401	Europarl-v7
Chinese	11,097,351	229,364,807	244,690,254	GALE
Czech	596,189	12,285,430	14,277,300	Europarl-v7
Dutch	1,997,775	49,533,217	50,661,711	Europarl-v7
Estonian	651,746	11,214,489	15,685,939	Europarl-v7
Finnish	1,924,942	32,330,289	47,526,505	Europarl-v7
French	52,004,519	932,475,412	821,546,279	Europarl-v7, 10^9 word parallel corpus, JRC, OpenSubtitles, UN
German	1,720,573	39,301,114	41,212,173	Europarl-v7
Greek	1,235,976	32,031,068	31,939,677	Europarl-v7
Hungarian	624,934	12,422,462	15,096,547	Europarl-v7
Italian	1,909,115	48,011,261	49,732,033	Europarl-v7
Latvian	637,599	11,957,078	15,412,186	Europarl-v7
Lithuanian	635,146	11,394,858	15,342,163	Europarl-v7
Polish	632,565	12,815,795	15,269,016	Europarl-v7
Portugese	1,960,407	49,961,396	49,283,373	Europarl-v7
Romanian	399,375	9,628,356	9,710,439	Europarl-v7
Russian	2,376,138	40,765,979	43,273,593	CommonCrawl, Yandex 1M corpus, News Commentary
Slovak	640,715	15,442,442	12,942,700	Europarl-v7
Slovenian	623,490	12,525,860	15,021,689	Europarl-v7
Swedish	1,862,234	45,767,032	41,602,279	Europarl-v7

Table 2: The sizes of the bilingual training data used to extract each language-specific version of PPDB.

Language	Code	Number of Paraphrases			
		Lexical	Phrasal	Syntactic	Total
Arabic	Ara	119.7M	45.1M	20.1M	185.7M
Bulgarian	Bul	1.3M	1.4M	1.2M	3.9M
Czech	Ces	7.3M	2.7M	2.6	12.1M
German	Deu	7.9M	15.4M	4.9M	28.3M
Greek	Ell	5.4M	9.4M	7.4M	22.3M
Estonian	Est	7.9M	1.0M	0.4M	9.2M
Finnish	Fin	41.4M	4.9M	2.3M	48.6M
French	Fra	78.8M	254.2M	170.5M	503.5M
Hungarian	Hun	3.8M	1.3M	0.2M	5.3M
Italian	Ita	8.2M	17.9M	9.7M	35.8M
Lithuanian	Lit	8.7M	1.5M	0.8M	11.0M
Latvian	Lav	5.5M	1.4M	1.0M	7.9M
Dutch	Nld	6.1M	15.3M	4.5M	25.9M
Polish	Pol	6.5M	2.2M	1.4M	10.1M
Portuguese	Por	7.0M	17.0M	9.0M	33.0M
Romanian	Ron	1.5M	1.8M	1.1M	4.5M
Russian	Rus	81M	46M	16M	144.4M
Slovak	Slk	4.8M	1.8M	1.7M	8.2M
Slovenian	Slv	3.6M	1.6M	1.4M	6.7M
Swedish	Swe	6.2M	10.3M	10.3M	26.8M
Chinese	Zho	52.5M	46.0M	8.9M	107.4M

Table 3: An overview over the sizes of the multilingual PPDB. The number of extracted paraphrases varies by language, depending on the amount of data available as well as the languages morphological richness. The language names are coded following ISO 639-2, using the terminology (“T”) code where applicable.

4.3. Morphological Variants as Paraphrases

Many of the languages covered by our resource are more morphologically complex than English. Since we are using English pivot phrases and English syntactic labels, the pivoting approach tends to group a variety of morphological variants of a foreign word into the same paraphrase cluster. For example, French adjectives inflect for gender and number, but English adjectives do not. Therefore, the French words *grand*, *grande*, *grands* and *grandes* would all share the English translation *tall*, and would therefore all be

Tag	Phrase	Paraphrases
VB	vais	va, vas, irai, vont, allons, ira, allez, irons
	vas	va, vont, allez, vais, allons, aller
	vont	vas, va, allons, allez, vais, aller
VBD	allais	allait, alliez, allaient, allions
VB	denke	denken, denkt

Table 4: Top paraphrases extracted for forms of the French *aller* and the German *denken*. The English part-of-speech label used preserves the unifying morphological characteristic quite well: present tense forms of *aller* dominate the ranking for the VB (which best corresponds with present tense usage in Englich). Similarly, imperfect forms are reliably captured for the past tense VBD tag.

grouped together as paraphrases of each other. It is unclear whether this grouping is desirable or not, and the answer may depend on the downstream task. It is clear that there are distinctions that are made in the French language that our paraphrasing method currently does not make.

This is also observable in verbs. Other languages often have more inflectional variation than English does. Whereas English verbs only distinguish between past versus present tense and 3rd person singular versus non-3rd person singular, other languages exhibit more forms. For instance, the English verb *go*, aligns to a variety of present forms of the French *aller*. The high-ranking paraphrases of *vais*, the first person singular form of *aller*, are all other forms of the verb. These are shown in Table 4. Similar effects can be observed across other verb paraphrases, both in French and other languages. The minimal distinction in the Penn Treebank tags between past tense verbs (VBD), base form verbs (VB) and present tense verbs (VBN/VBP), partitions the foreign verbs to some extent. But clearly there is a semantic distinction between verb forms that are marked for person and number, which our method is not currently making.

The interaction between out bilingual pivoting method and English's impoverished morphologic system, open up avenues for improving the quality of the multilingual paraphrases. Our method makes distinctions between paraphrases when they have different syntactic labels. This does a good job of separating out things that make a sense distinction based on part of speech (like *squash* which paraphrases as *racquetball* as a noun and *crush* as a verb). It also limits different paraphrases based on which form the original phrase takes. For instance, *divide* can paraphrase as *fracture* or *split* in both noun and verb forms, but it can only paraphrase as *gap* when the original phrase is a noun. Currently we use Penn Treebank tags, which are rather English-centric. This tag set could be replaced or refined to make finer-grained distinctions that are present in the foreign language. Refined, language-specific tag sets would do a better job at partitioning paraphrase sets that should be distinct.

5. Future work

We have previously shown significant improvements to the quality of English paraphrases when we re-score the bilingually extracted paraphrases with monolingually-derived similarity measures (Chan et al., 2011; Ganitkevitch et al., 2012). Distributional similarity measures can be computed from large monolingual corpora by constructing vectors that represent the contexts that a word or phrase appears in. The similarity of different words can be measured by comparing their vector space representations (Turney and Pantel, 2010). Previous work on paraphrasing, like DIRT (Lin and Pantel, 2001), has used monolingual distributional similarity directly. This sometimes results in distributionally similar antonyms (like *rise* and *fall* in English) or terms that are related but mutually exclusive (like *boys* and *girls*) being incorrectly grouped as paraphrases. We use monolingual distributional similarity to re-rank our bilingually derived paraphrases. Since it is less common to group antonyms with our bilingual pivoting method, the quality can be higher than DIRT-like methods. The vector space models provide an orthogonal signal to improve the ranking of our paraphrases.

Since large amounts of monolingual data are readily available, we expect a significant improvement in paraphrase quality by re-ranking our non-English paraphrases, especially for language for which we only have small amounts of bitexts, such as Bulgarian or Romanian.

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.

6. References

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research (JAIR)*, 38:135–187.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Barzilay, R. and Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of ACL*.

Berant, J. and Liang, P. (2014). Semantic parsing via paraphrasing. In *Proceedings of ACL*.

Bhagat, R. and Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472, March.

Bhagat, R. and Ravichandran, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL/HLT*.

Bosma, W. and Callison-Burch, C. (2007). Paraphrase substitution for recognizing textual entailment. In Peters, C., Clough, P., Gey, F., Karlgren, J., Magnini, B., Oard, D., Rijke, M., and Stempfhuber, M., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, Lecture Notes in Computer Science, pages 502–509. Springer.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*.

Chan, T. P., Callison-Burch, C., and Van Durme, B. (2011). Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–42, Edinburgh, UK, July. Association for Computational Linguistics.

Chen, D. and Dolan, B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*.

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

Denkowski, M. and Lavie, A. (2010). METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 339–342.

Denkowski, M., Al-Haj, H., and Lavie, A. (2010). Turker-assisted paraphrasing for english-arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 66–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the COLING*.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of EMNLP*.

Ganitkevitch, J., Durme, B. V., and Callison-Burch, C. (2012). Monolingual distributional similarity for text-to-text generation. In *Proceedings of *SEM*. Association for Computational Linguistics.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–325.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5.

Lin, D. and Pantel, P. (2001). Discovery of inference rules from text. *Natural Language Engineering*.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2013). Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84, Seattle, Washington, USA, October. Association for Computational Linguistics.

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

Madnani, N., Ayan, N. F., Resnik, P., and Dorr, B. (2007). Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of WMT07*.

Malakasiotis, P. and Androutsopoulos, I. (2011). A generate and rank approach to sentence paraphrasing. In *Proceedings of EMNLP*, pages 96–106.

McKeown, K. R. (1979). Paraphrasing using given and new information in a question-answer system. In *Proceedings of ACL*.

Mizukami, M., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). Creation of a Japanese paraphrase database and its application to linguistic individuality transformation. In *Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing (NLP2014)*, pages 773–776, Hokkaido, Japan, March. This paper is written in Japanese.

Napoles, C., Callison-Burch, C., Ganitkevitch, J., and Van Durme, B. (2011). Paraphrastic sentence compression with a character-based metric: Tightening without deletion. *Workshop on Monolingual Text-To-Text Generation*.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., and Callison-Burch, C. (2013). Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 206–212, Sofia, Bulgaria, August.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*.

Ravichandran, D. and Hovy, E. (2002). Learning sufrace text patterns for a question answering system. In *Proceedings of ACL*.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V., and Liu, Y. (2007). Statistical machine translation for query expansion in answer retrieval. In *Proceedings of ACL*.

Smith, J., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, July. Association for Computational Linguistics.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2010). Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genoa, Italy.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *CoRR*, abs/1003.1141.

Vila, M., Martí, M. A., and Rodríguez, H. (2014). Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Weese, J., Ganitkevitch, J., Callison-Burch, C., Post, M., and Lopez, A. (2011). Joshua 3.0: Syntax-based machine translation with the Thrax grammar extractor. In *Proceedings of WMT11*.

Xu, W., Ritter, A., and Grishman, R. (2013). Gathering and generating paraphrases from twitter with application to normalization. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 121–128, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zhang, Y. and Yamamoto, K. (2002). Paraphrasing of chinese utterances. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhang, Y. and Yamamoto, K. (2005). Paraphrasing spoken chinese using a paraphrase corpus. *Nat. Lang. Eng.*, 11(4):417–434, December.

Zhao, S., Niu, C., Zhou, M., Liu, T., and Li, S. (2008). Combining multiple resources to improve SMT-based paraphrasing model. In *Proceedings of ACL/HLT*.

Zhou, L., Lin, C.-Y., Munteanu, D. S., and Hovy, E. (2006). ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of HLT/NAACL*.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings of WMT06*.