PPDB: The Paraphrase Database

Juri Ganitkevitch[1] **Benjamin Van Durme**[1,2] **Chris Callison-Burch**[2,3]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Human Language Technology Center of Excellence, Johns Hopkins University
[3]Computer and Information Science Department, University of Pennsylvania

Abstract

We present the 1.0 release of our paraphrase database, PPDB. Its English portion, PPDB:Eng, contains over 220 million paraphrase pairs, consisting of 73 million phrasal and 8 million lexical paraphrases, as well as 140 million paraphrase patterns, which capture many meaning-preserving syntactic transformations. The paraphrases are extracted from bilingual parallel corpora totaling over 100 million sentence pairs and over 2 billion English words. We also release PPDB:Spa, a collection of 196 million Spanish paraphrases. Each paraphrase pair in PPDB contains a set of associated scores, including paraphrase probabilities derived from the bitext data and a variety of monolingual distributional similarity scores computed from the Google n-grams and the Annotated Gigaword corpus. Our release includes pruning tools that allow users to determine their own precision/recall tradeoff.

1 Introduction

Paraphrases, i.e. differing textual realizations of the same meaning, have proven useful for a wide variety of natural language processing applications. Past paraphrase collections include automatically derived resources like DIRT (Lin and Pantel, 2001), the MSR paraphrase corpus and phrase table (Dolan et al., 2004; Quirk et al., 2004), among others. Although several groups have independently extracted paraphrases using Bannard and Callison-Burch (2005)'s bilingual pivoting technique (see Zhou et al. (2006), Riezler et al. (2007), Snover et al. (2010), among others), there has never been an official release of this resource.

In this work, we release version 1.0 of the *Para-Phrase DataBase* PPDB,[1] a collection of ranked English and Spanish paraphrases derived by:

- Extracting lexical, phrasal, and syntactic paraphrases from large bilingual parallel corpora (with associated paraphrase probabilities).
- Computing distributional similarity scores for each of the paraphrases using the Google n-grams and the Annotated Gigaword corpus.

In addition to the paraphrase collection itself, we provide tools to filter PPDB to only retain high precision paraphrases, scripts to limit the collection to phrasal or lexical paraphrases (synonyms), and software that enables users to extract paraphrases for languages other than English.

2 Extracting Paraphrases from Bitexts

To extract paraphrases we follow Bannard and Callison-Burch (2005)'s bilingual pivoting method. The intuition is that two English strings e_1 and e_2 that translate to the same foreign string f can be assumed to have the same meaning. We can thus *pivot* over f and extract $\langle e_1, e_2 \rangle$ as a pair of paraphrases, as illustrated in Figure 1. The method extracts a diverse set of paraphrases. For *thrown into jail*, it extracts *arrested, detained, imprisoned, incarcerated, jailed, locked up, taken into custody*, and *thrown into prison*, along with a set of incorrect/noisy paraphrases that have different syntactic types or that are due to misalignments.

For PPDB, we formulate our paraphrase collection as a weighted *synchronous context-free grammar* (SCFG) (Aho and Ullman, 1972; Chiang, 2005)

[1]Freely available at `http://paraphrase.org`.

Proceedings of NAACL-HLT 2013, pages 758–764,
Atlanta, Georgia, 9–14 June 2013.

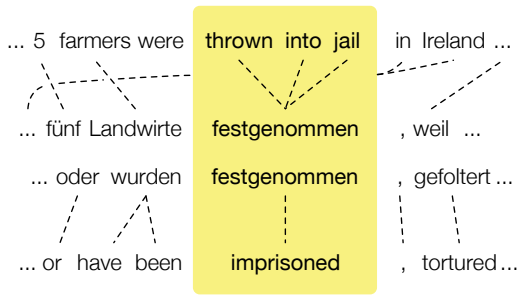


Figure 1: Phrasal paraphrases are extracted via bilingual pivoting.

with syntactic nonterminal labels, similar to Cohn and Lapata (2008) and Ganitkevitch et al. (2011). An SCFG rule has the form:

$$\mathbf{r} \stackrel{\text{def}}{=} C \rightarrow \langle f, e, \sim, \vec{\varphi} \rangle,$$

where the left-hand side of the rule, C, is a nonterminal and the right-hand sides f and e are strings of terminal and nonterminal symbols. There is a one-to-one correspondence, $\sim$, between the nonterminals in f and e: each nonterminal symbol in f has to also appear in e. Following Zhao et al. (2008), each rule $\mathbf{r}$ is annotated with a vector of feature functions $\vec{\varphi} = \{\varphi_1 ... \varphi_N\}$ which are combined in a log-linear model (with weights $\vec{\lambda}$) to compute the *cost* of applying $\mathbf{r}$:

$$cost(\mathbf{r}) = -\sum_{i=1}^{N} \lambda_i \log \varphi_i. \quad (1)$$

To create a syntactic paraphrase grammar we first extract a foreign-to-English translation grammar from a bilingual parallel corpus, using techniques from syntactic machine translation (Koehn, 2010). Then, for each pair of translation rules where the left-hand side C and foreign string f match:

$$\mathbf{r}_1 \stackrel{\text{def}}{=} C \rightarrow \langle f, e_1, \sim_1, \vec{\varphi}_1 \rangle$$
$$\mathbf{r}_2 \stackrel{\text{def}}{=} C \rightarrow \langle f, e_2, \sim_2, \vec{\varphi}_2 \rangle,$$

we *pivot* over f to create a paraphrase rule $\mathbf{r}_p$:

$$\mathbf{r}_p \stackrel{\text{def}}{=} C \rightarrow \langle e_1, e_2, \sim_p, \vec{\varphi}_p \rangle,$$

with a combined nonterminal correspondency function $\sim_p$. Note that the common source side f implies that e_1 and e_2 share the same set of nonterminal symbols.

The paraphrase rules obtained using this method are capable of making well-formed generalizations of meaning-preserving rewrites in English. For instance, we extract the following example paraphrase, capturing the English possessive rule:

$$NP \rightarrow \text{ the } NP_1 \text{ of } NNS_2 \mid \text{the } NNS_2 \text{ 's } NP_1.$$

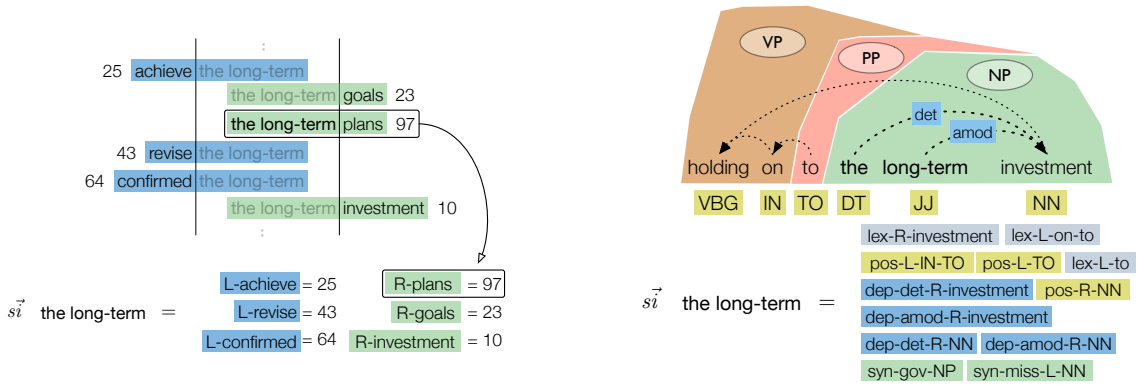
The paraphrase feature vector $\vec{\varphi}_p$ is computed from the translation feature vectors $\vec{\varphi}_1$ and $\vec{\varphi}_2$ by following the pivoting idea. For instance, we estimate the conditional paraphrase probability $p(e_2|e_1)$ by marginalizing over all shared foreign-language translations f:

$$p(e_2|e_1) \approx \sum_f p(e_2|f) p(f|e_1). \quad (2)$$

3 Scoring Paraphrases Using Monolingual Distributional Similarity

The bilingual pivoting approach anchors paraphrases that share an interpretation because of a shared foreign phrase. Paraphrasing methods based on monolingual text corpora, like DIRT (Lin and Pantel, 2001), measure the similarity of phrases based on distributional similarity. This results in a range of different types of phrases, including paraphrases, inference rules and antonyms. For instance, for *thrown into prison* DIRT extracts good paraphrases like *arrested, detained,* and *jailed*. However, it also extracts phrases that are temporarily or causally related like *began the trial of, cracked down on, interrogated, prosecuted* and *ordered the execution of*, because they have similar distributional properties. Since bilingual pivoting rarely extracts these non-paraphrases, we can use monolingual distributional similarity to re-rank paraphrases extracted from bitexts (following Chan et al. (2011)) or incorporate a set of distributional similarity scores as features in our log-linear model.

Each similarity score relies on precomputed distributional signatures that describe the contexts that a phrase occurs in. To describe a phrase e, we gather counts for a set of contextual features for each occurrence of e in a corpus. Writing the context vector for the i-th occurrence of e as $\vec{s}_{e,i}$, we can aggregate over all occurrences of e, resulting in a *distributional* signature for e, $\vec{s}_e = \sum_i \vec{s}_{e,i}$. Following the intuition that phrases with similar meanings occur in



(a) The n-gram corpus records *the long-term* as preceded by *revise* (43 times), and followed by *plans* (97 times). We add corresponding features to the phrase's distributional signature retaining the counts of the original n-grams.

(b) Here, position-aware lexical and part-of-speech n-gram features, labeled dependency links , and features reflecting the phrase's CCG-style label NP/NN are included in the context vector.

Figure 2: Features extracted for the phrase *the long term* from the n-gram corpus (2a) and Annotated Gigaword (2b).

similar contexts, we can then quantify the goodness of e' as a paraphrase of e by computing the cosine similarity between their distributional signatures:

$$sim(e, e') = \frac{\vec{s}_e \cdot \vec{s}_{e'}}{|\vec{s}_e||\vec{s}_{e'}|}.$$

A wide variety of features have been used to describe the distributional context of a phrase. Rich, linguistically informed feature-sets that rely on dependency and constituency parses, part-of-speech tags, or lemmatization have been proposed in work such as by Church and Hanks (1991) and Lin and Pantel (2001). For instance, a phrase is described by the various syntactic relations such as: "what verbs have this phrase as the subject?", or "what adjectives modify this phrase?". Other work has used simpler n-gram features, e.g. "what words or bigrams have we seen to the left of this phrase?". A substantial body of work has focussed on using this type of feature-set for a variety of purposes in NLP (Lapata and Keller, 2005; Bhagat and Ravichandran, 2008; Lin et al., 2010; Van Durme and Lall, 2010).

For PPDB, we compute n-gram-based context signatures for the 200 million most frequent phrases in the Google n-gram corpus (Brants and Franz, 2006; Lin et al., 2010), and richer linguistic signatures for 175 million phrases in the Annotated Gigaword corpus (Napoles et al., 2012). Our features extend beyond those previously used in the work by Ganitkevitch et al. (2012). They are:

- n-gram based features for words seen to the left and right of a phrase.
- Position-aware lexical, lemma-based, part-of-speech, and named entity class unigram and bigram features, drawn from a three-word window to the right and left of the phrase.
- Incoming and outgoing (wrt. the phrase) dependency link features, labeled with the corresponding lexical item, lemmata and POS.
- Syntactic features for any constituents governing the phrase, as well as for CCG-style slashed constituent labels for the phrase.

Figure 2 illustrates the feature extraction for an example phrase.

4 English Paraphrases – PPDB:Eng

We combine several English-to-foreign bitext corpora to extract PPDB:Eng: Europarl v7 (Koehn, 2005), consisting of bitexts for the 19 European languages, the 10^9 French-English corpus (Callison-Burch et al., 2009), the Czech, German, Spanish and French portions of the News Commentary data (Koehn and Schroeder, 2007), the United Nations French- and Spanish-English parallel corpora (Eisele and Chen, 2010), the JRC Acquis corpus (Steinberger et al., 2006), Chinese and Arabic

	Identity	Paraphrases	Total
Lexical	0.6M	7.6M	8.1M
Phrasal	4.9M	68.4M	73.2M
Syntactic	46.5M	93.6M	140.1M
All	52.0M	169.6M	221.4M

Table 1: A breakdown of PPDB:Eng size by paraphrase type. We distinguish lexical (i.e. one-word) paraphrases, phrasal paraphrases and syntactically labeled paraphrase patterns.

	Identity	Paraphrases	Total
Lexical	1.0M	33.1M	34.1M
Phrasal	4.3M	73.2M	77.5M
Syntactic	29.4M	55.3M	84.7M
All	34.7M	161.6M	196.3M

Table 2: An overview of PPDB:Spa. Again, we partition the resource into lexical (i.e. one-word) paraphrases, phrasal paraphrases and syntactically labeled paraphrase patterns.

newswire corpora used for the GALE machine translation campaign,[2] parallel Urdu-English data from the NIST translation task,[3] the French portion of the OpenSubtitles corpus (Tiedemann, 2009), and a collection of Spanish-English translation memories provided by TAUS.[4]

The resulting composite parallel corpus has more than 106 million sentence pairs, over 2 billion English words, and spans 22 pivot languages. To apply the pivoting technique to this multilingual data, we treat the various pivot languages as a joint *Non-English* language. This simplifying assumption allows us to share statistics across the different languages and apply Equation 2 unaltered.

Table 1 presents a breakdown of PPDB:Eng by paraphrase type. We distinguish *lexical* (a single word), *phrasal* (a continuous string of words), and *syntactic* paraphrases (expressions that may contain both words and nonterminals), and separate out identity paraphrases. While we list lexical and phrasal paraphrases separately, it is possible that a single word paraphrases as a multi-word phrase and vice versa – so long they share the same syntactic label.

5 Spanish Paraphrases – PPDB:Spa

We also release a collection of Spanish paraphrases: PPDB:Spa is extracted analogously to its English counterpart and leverages the Spanish portions of the bitext data available to us, totaling almost 355 million Spanish words, in nearly 15 million sentence pairs. The paraphrase pairs in PPDB:Spa are annotated with distributional similarity scores based on lexical features collected from the Spanish portion of the multilingual release of the Google n-gram corpus (Brants and Franz, 2009), and the Spanish Gigaword corpus (Mendonca et al., 2009). Table 2 gives a breakdown of PPDB:Spa.

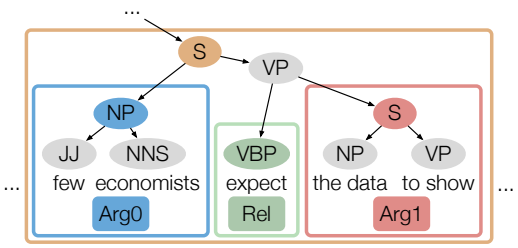


Figure 3: To inspect our coverage, we use the Penn Treebank's parses to map from Propbank annotations to PPDB's syntactic patterns. For the above annotation predicate, we extract $VBP \rightarrow$ expect, which is matched by paraphrase rules like $VBP \rightarrow$ expect | anticipate and $VBP \rightarrow$ expect | hypothesize. To search for the entire relation, we replace the argument spans with syntactic nonterminals. Here, we obtain $S \rightarrow NP$ expect S, for which PPDB has matching rules like $S \rightarrow NP$ expect S | NP would hope S, and $S \rightarrow NP$ expect S | NP trust S. This allows us to apply sophisticated paraphrases to the predicate while capturing its arguments in a generalized fashion.

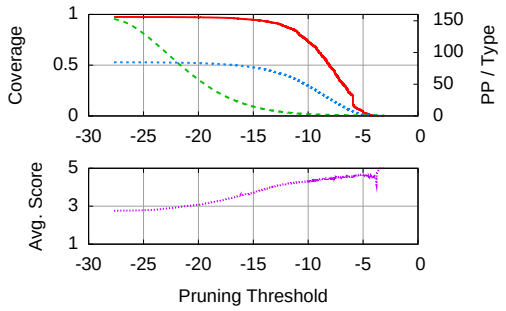
6 Analysis

To estimate the usefulness of PPDB as a resource for tasks like semantic role labeling or parsing, we analyze its coverage of Propbank predicates and predicate-argument tuples (Kingsbury and Palmer, 2002). We use the Penn Treebank (Marcus et al., 1993) to map Propbank annotations to patterns which allow us to search PPDB:Eng for paraphrases that match the annotated predicate. Figure 3 illus-

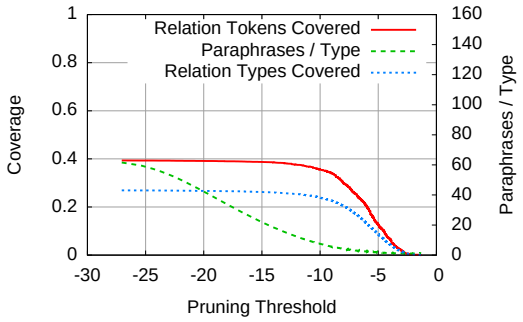
[2] http://projects.ldc.upenn.edu/gale/data/Catalog.html

[3] LDC Catalog No. LDC2010T23

[4] http://www.translationautomation.com/



(a) PPDB:Eng coverage of Propbank predicates (top), and average human judgment score (bottom) for varying pruning thresholds.



(b) PPDB:Eng's coverage of Propbank predicates with up to two arguments. Here we consider rules that paraphrase the full predicate-argument expression.

Figure 4: An illustration of PPDB's coverage of the manually annotated Propbank predicate phrases (4a) and binary relations with argument non-terminals (4b). The curves indicate the coverage on tokens (solid) and types (dotted), as well as the average number of paraphrases per covered type (dashed) at the given pruning level.

trates this mapping.

In order to quantify PPDB's precision-recall tradeoff in this context, we perform a sweep over our collection, beginning with the full set of paraphrase pairs and incrementally discarding the lowest-scoring ones. We choose a simple estimate for each paraphrase pair's score by uniformly combining its paraphrase probability features in Eq. 1.

The top graph in Figure 4a shows PPDB's coverage of predicates (e.g. $VBP \rightarrow$ expect) at the type level (i.e. counting distinct predicates), as well as the token level (i.e. counting predicate occurrences in the corpus). We also keep track of average number of paraphrases per covered predicate type for varying pruning levels. We find that PPDB has a predicate type recall of up to 52% (accounting for 97.5% of tokens). Extending the experiment to full predicate-argument relations with up to two arguments (e.g. $S \rightarrow NNS$ expect S), we obtain a 27% type coverage rate that accounts for 40% of tokens (Figure 4b). Both rates hold even as we prune the database down to only contain high precision paraphrases. Our pruning method here is based on a simple uniform combination of paraphrase probabilities and similarity scores.

To gauge the quality of our paraphrases, the authors judged 1900 randomly sampled predicate paraphrases on a scale of 1 to 5, 5 being the best. The bottom graph in Figure 4a plots the resulting human score average against the sweep used in the coverage experiment. It is clear that even with a simple weighing approach, the PPDB scores show a clear correlation with human judgements. Therefore they can be used to bias the collection towards greater recall or higher precision.

7 Conclusion and Future Work

We present the 1.0 release of PPDB:Eng and PPDB:Spa, two large-scale collections of paraphrases in English and Spanish. We illustrate the resource's utility with an analysis of its coverage of Propbank predicates. Our results suggest that PPDB will be useful in a variety of NLP applications.

Future releases of PPDB will focus on expanding the paraphrase collection's coverage with regard to both data size and languages supported. Furthermore, we intend to improve paraphrase scoring by incorporating additional sources of information, as well as by better utilizing information present in the data, like domain or topic. We will also address points of refinement such as handling of phrase ambiguity, and effects specific to individual pivot languages. Our aim is for PPDB to be a continuously updated and improving resource.

Finally, we will explore extensions to PPDB to include aspects of related large-scale resources such as lexical-semantic hierarchies (Snow et al., 2006), textual inference rules (Berant et al., 2011), relational patterns (Nakashole et al., 2012), and (lexical) conceptual networks (Navigli and Ponzetto, 2012).

Acknowledgements

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

References

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

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

Jonathan Berant, Jacob Goldberger, and Ido Dagan. 2011. Global learning of typed entailment rules. 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.

Thorsten Brants and Alex Franz. 2009. Web 1T 5-gram, 10 european languages version 1. *Linguistic Data Consortium, Philadelphia*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of WMT*, pages 1–28, Athens, Greece, March.

Tsz Ping Chan, Chris Callison-Burch, and Benjamin Van Durme. 2011. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *EMNLP Workshop on GEMS*.

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

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

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the COLING*.

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

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of LREC*, Valletta, Malta.

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

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2012. Monolingual distributional similarity for text-to-text generation. In *Proceedings of *SEM*. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of LREC*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of WMT*, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

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 from text. *Natural Language Engineering*.

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*.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2).

Angelo Mendonca, David Andrew Graff, and Denise DiPersio. 2009. *Spanish Gigaword Second Edition*. Linguistic Data Consortium.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: a taxonomy of relational patterns with semantic types. In *Proceedings of EMNLP*.

Courtney Napoles, Matt Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of AKBC-WEKEX 2012*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193.

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

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the 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.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the ACL/Coling*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genoa, Italy.

Jörg Tiedemann. 2009. News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5.

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

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*.

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