Large-Scale Paraphrasing for Natural Language Understanding

Juri Ganitkevitch
Center for Language and Speech Processing
Johns Hopkins University
juri@cs.jhu.edu

Abstract

We examine the application of data-driven paraphrasing to natural language understanding. We leverage bilingual parallel corpora to extract a large collection of syntactic paraphrase pairs, and introduce an adaptation scheme that allows us to tackle a variety of text transformation tasks via paraphrasing. We evaluate our system on the sentence compression task. Further, we use distributional similarity measures based on context vectors derived from large monolingual corpora to annotate our paraphrases with an orthogonal source of information. This yields significant improvements in our compression system's output quality, achieving state-of-the-art performance. Finally, we propose a refinement of our paraphrases by classifying them into natural logic entailment relations. By extending the synchronous parsing paradigm towards these entailment relations, we will enable our system to perform recognition of textual entailment.

1 Introduction

In this work, we propose an extension of current paraphrasing methods to tackle natural language understanding problems. We create a large set of paraphrase pairs in a data-driven fashion, rank them based on a variety of similarity metrics, and attach an entailment relation to each pair, facilitating natural logic inference. The resulting resource has potential applications to a variety of NLP applications, including summarization, query expansion, question answering, and recognizing textual entailment.

Specifically, we build on Callison-Burch (2007)'s pivot-based paraphrase extraction method, which uses bilingual parallel data to learn English *phrase pairs* that share the same meaning. Our approach extends the pivot method to learn meaning-preserving *syntactic transformations* in English. We represent these using synchronous context-free grammars (SCFGs). This representation allows us to re-use a lot of machine translation machinery to perform monolingual text-to-text generation. We demonstrate the method on a sentence compression task (Ganitkevitch et al., 2011).

To improve the system, we then incorporate features based on monolingual distributional similarity. This orthogonal source of signal allows us to re-scores the bilingually-extracted paraphrases using information drawn from large monolingual corpora. We show that the monolingual distributional scores yield significant improvements over a baseline that scores paraphrases only with bilingually-extracted features (Ganitkevitch et al., 2012).

Further, we propose a semantics for paraphrasing by classifying each paraphrase pair with one of the entailment relation types defined by natural logic (MacCartney, 2009). Natural logic is used to perform inference over pairs of natural language phrases, like our paraphrase pairs. It defines a set of relations including, equivalence ($\equiv$), forward- and backward-entailments ($\sqsubset$, $\sqsupset$), antonyms ($\wedge$), and others. We will build a classifier for our paraphrases that uses features extracted from annotated resources like WordNet and distributional information gathered over large text corpora to assign one or more entailment relations to each paraphrase pair. We will evaluate the entailment assignments by applying this enhanced paraphrasing system to the task of recognizing textual entailment (RTE).

2 Extraction of Syntactic Paraphrases from Bitexts

A variety of different types of corpora have been used to automatically induce paraphrase collections for English (see Madnani and Dorr (2010) for a sur-

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 62–68,
Atlanta, Georgia, 13 June 2013.

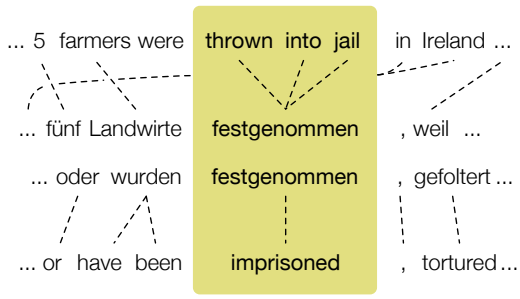


Figure 1: An example of pivot-based phrasal paraphrase extraction – we assume English phrases that translate to a common German phrase to be paraphrases. Thus we extract "imprisoned" as a paraphrase of "thrown into jail."

Possessive rule		
$NP \rightarrow$	the NN of the NNP	the NNP's NN
$NP \rightarrow$	the NP made by NN	the NN's NP
Dative shift		
$VP \rightarrow$	give NN to NP	give NP the NN
$VP \rightarrow$	provide NP_1 to NP_2	give NP_2 NP_1
Partitive constructions		
$NP \rightarrow$	CD of the NN	CD NN
$NP \rightarrow$	all NN	all of the NN
Reduced relative clause		
$SBAR/S \rightarrow$	although PRP VBP that	although PRP VBP
$ADJP \rightarrow$	very JJ that S	JJ S

Table 1: A selection of example paraphrase patterns extracted by our system. These rules demonstrate that, using the pivot approach from Figure 1, our system is capable of learning meaning-preserving syntactic transformations in English.

vey of these methods). Bannard and Callison-Burch (2005) extracted phrasal paraphrases from bitext by using foreign language phrases as a *pivot*: if two English phrases e_1 and e_2 both translate to a foreign phrase f, they assume that e_1 and e_2 are paraphrases of one another. Figure 1 gives an example of a phrasal paraphrase extracted by Bannard and Callison-Burch (2005).

Since "thrown into jail" is aligned to multiple German phrases, and since each of those German phrases align back to a variety of English phrases, the method extracts a wide range of possible paraphrases including good paraphrase like: *imprisoned* and *thrown into prison*. It also produces less good paraphrases like: *in jail* and *put in prison for*, and bad paraphrases, such as *maltreated* and *protection*, because of noisy/inaccurate word alignments and other problems. To rank these, Bannard and Callison-Burch (2005) derive a paraphrase probability $p(e_1|e_2)$:

$$p(e_2|e_1) \approx \sum_f p(e_2|f)p(f|e_1), \qquad (1)$$

where the $p(e_i|f)$ and $p(f|e_i)$ are translation probabilities estimated from the bitext (Brown et al., 1990; Koehn et al., 2003).

We extend this method to extract *syntactic paraphrases* (Ganitkevitch et al., 2011). Table 1 shows example paraphrases produced by our system. While phrasal systems memorize phrase pairs without any further generalization, a syntactic paraphrasing system can learn more generic patterns. These can be better applied to unseen data. The paraphrases implementing the *possessive rule* and the *dative shift* shown in Table 1 are good examples of this: the two noun-phrase arguments to the expressions are abstracted to nonterminals while each rule's lexicalization provides an appropriate frame of evidence for the transform.

2.1 Formal Representation

In this proposal we focus on a paraphrase model based on *synchronous context-free grammar* (SCFG). The SCFG formalism (Aho and Ullman, 1972) was repopularized for statistical machine translation by (Chiang, 2005). An *probabilistic* SCFG $\mathcal{G}$ contains rules $\mathbf{r}$ of the form $\mathbf{r} = C \rightarrow \langle \gamma, \alpha, \sim, w \rangle$. A rule $\mathbf{r}$'s left-hand side C is a nonterminal, while its right-hands sides γ and α can be mixed strings of words and nonterminal symbols. There is a one-to-one correspondency between the nonterminals in γ and α. Each rule is assigned a cost $w_\mathbf{r} \geq 0$, reflecting its likelihood.

To compute the cost $w_\mathbf{r}$ of the application of a rule $\mathbf{r}$, we define a set of feature functions $\vec{\varphi} = \{\varphi_1 ... \varphi_N\}$ that are combined in a log-linear model. The model weights are set to maximize a task-dependent objective function.

2.2 Syntactic Paraphrase Rules via Bilingual Pivoting

Our paraphrase acquisition method is based on the extraction of syntactic translation rules in statistical machine translation (SMT). In SMT, SCFG rules are extracted from English-foreign sentence pairs that are automatically parsed and word-aligned. For a

	CR	Meaning	Grammar
Reference	0.80	4.80	4.54
ILP	0.74	3.44	**3.41**
PP	0.78	3.53	2.98
PP + n-gram	0.80	3.65	3.16
PP + syntax	0.79	**3.70**	3.26
Random Deletions	0.78	2.91	2.53

Table 2: Results of the human evaluation on longer compressions: pairwise compression ratios (CR), meaning and grammaticality scores. Bold indicates a statistically significant best result at $p < 0.05$. The scores range from 1 to 5, 5 being perfect.

foreign phrase the corresponding English phrase is found via the word alignments. This phrase pair is turned into an SCFG rule by assigning a left-hand side nonterminal symbol, corresponding to the syntactic constituent that dominates the English phrase. To introduce nonterminals into the right-hand sides of the rule, we can replace corresponding sub-phrases in the English and foreign phrases with nonterminal symbols. Doing this for all sentence pairs in a bilingual parallel corpus results in a *translation grammar* that serves as the basis for syntactic machine translation.

To create a *paraphrase grammar* from a translation grammar, we extend the syntactically informed pivot approach of (Callison-Burch, 2008) to the SCFG model: for each pair of translation rules $\mathbf{r}_1$ and $\mathbf{r}_2$ with matching left-hand side nonterminal C and foreign language right-hand side γ: $\mathbf{r}_1 = C \rightarrow \langle \gamma, \alpha_1, \sim_1, \vec{\varphi}_1 \rangle$ and $\mathbf{r}_2 = C \rightarrow \langle \gamma, \alpha_2, \sim_2, \vec{\varphi}_2 \rangle$, we pivot over γ and create a paraphrase rule $\mathbf{r}_p$: $\mathbf{r}_p = C \rightarrow \langle \alpha_1, \alpha_2, \sim, \vec{\varphi} \rangle$. We estimate the cost for $\mathbf{r}_p$ following Equation 1.

2.3 Task-Based Evaluation

Sharing its SCFG formalism permits us to re-use much of SMT's machinery for paraphrasing applications, including decoding and minimum error rate training. This allows us to easily tackle a variety of monolingual text-to-text generation tasks, which can be cast as sentential paraphrasing with task-specific constraints or goals.

For our evaluation, we apply our paraphrase system to sentence compression. However, to successfully use paraphrases for sentence compression, we need to adapt the system to suit the task. We introduce a four-point adaptation scheme for text-to-text

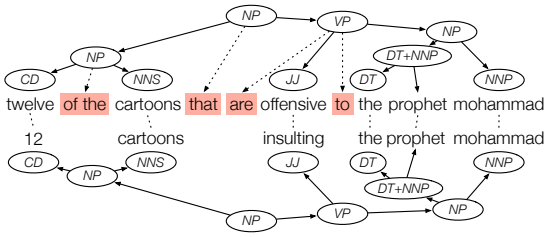


Figure 2: An example of a synchronous paraphrastic derivation in sentence compression.

generation via paraphrases, suggesting:

- The use *task-targeted features* that capture information pertinent to the text transformation. For sentence compression the features include word count and length-difference features.
- An *objective function* that takes into account the contraints imposed by the task. We use PRÉCIS, an augmentation of the BLEU metric, which introduces a verbosity penalty.
- *Development data* that represents the precise transformations we seek to model. We use a set of human-made example compressions mined from translation references.
- Optionally, *grammar augmentations* that allow for the incorporation of effects that the learned paraphrase grammar cannot capture. We experimented with automatically generated deletion rules.

Applying the above adaptations to our generic paraphraser (PP), quickly yields a sentence compression system that performs on par with a state-of-the-art integer linear programming-based (ILP) compression system (Clarke and Lapata, 2008). As Table 2 shows, human evaluation results suggest that our system outperforms the contrast system in meaning retention. However, it suffers losses in grammaticality. Figure 2 shows an example derivation produced as a result of applying our paraphrase rules in the decoding process.

3 Integrating Monolingual Distributional Similarity into Bilingually Extracted Paraphrases

Distributional similarity-based methods (Lin and Pantel, 2001; Bhagat and Ravichandran, 2008) rely

on the assumption that similar expressions appear in similar contexts – a signal that is orthogonal to bilingual pivot information we have considered thus far. However, the monolingual distributional signal is noisy: it suffers from problems such as mistaking cousin expressions or antonyms (such as $\langle rise, fall \rangle$ or $\langle boy, girl \rangle$) for paraphrases. We circumvent this issue by starting with a paraphrase grammar extracted from bilingual data and *reranking* it with information based on distributional similarity (Ganitkevitch et al., 2012).

3.1 Distributional Similarity

In order to compute the similarity of two expressions e_1 and e_2, their respective occurrences across a corpus are aggregated in context vectors $\vec{c}_1$ and $\vec{c}_2$. The $\vec{c}_i$ are typically vectors in a high-dimensional feature space with features like counts for words seen within a window of an e_i. For parsed data more sophisticated features based on syntax and dependency structure around an occurrence are possible. The comparison of e_1 and e_2 is then made by computing the cosine similarity between $\vec{c}_1$ and $\vec{c}_2$.

Over large corpora the context vectors for even moderately frequent e_i can grow unmanageably large. Locality sensitive hashing provides a way of dealing with this problem: instead of retaining the explicit sparse high-dimensional $\vec{c}_i$, we use a random projection $h(\cdot)$ to convert them into compact bit signatures in a dense b-dimensional boolean space in which approximate similarity calculation is possible.

3.2 Integrating Similarity with Syntactic Paraphrases

In order to incorporate distributional similarity information into the paraphrasing system, we need to calculate similarity scores for the paraphrastic SCFG rules in our grammar. For rules with purely lexical right-hand sides e_1 and e_2 this is a simple task, and the similarity score $sim(e_1, e_2)$ can be directly included in the rule's feature vector $\vec{\varphi}$. However, if e_1 and e_2 are long, their occurrences become sparse and their similarity can no longer be reliably estimated. In our case, the right-hand sides of our rules also contain non-terminal symbols and re-ordered phrases, so computing a similarity score is not straightforward.

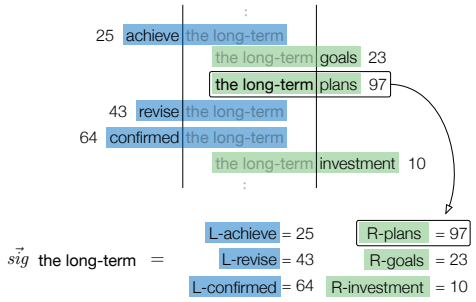


Figure 3: An example of the n-gram feature extraction on an n-gram corpus. Here, "the long-term" is seen preceded by "revise" (43 times) and followed by "plans" (97 times).

Our solution is to decompose the discontinuous patterns that make up the right-hand sides of a rule $\mathbf{r}$ into pairs of contiguous phrases, for which we then look up distributional signatures and compute similarity scores. To avoid comparing unrelated pairs, we require the phrase pairs to be consistent with a token alignment $\mathbf{a}$, defined and computed analogously to word alignments in machine translation.

3.3 Data Sets and Types of Distributional Signatures

We investigate the impact of the data and feature set used to construct distributional signatures. In particular we contrast two approaches: a large collection of distributional signatures with a relatively simple feature set, and a much smaller set of signatures with a rich, syntactically informed feature set.

The larger *n-gram model* is drawn from a web-scale n-gram corpus (Brants and Franz, 2006; Lin et al., 2010). Figure 3 illustrates this feature extraction approach. The resulting collection comprises distributional signatures for the 200 million most frequent 1-to-4-grams in the n-gram corpus.

For the syntactically informed model, we use the constituency and dependency parses provided in the Annotated Gigaword corpus (Napoles et al., 2012). Figure 4 illustrates this model's feature extraction for an example phrase occurrence. Using this method we extract distributional signatures for over 12 million 1-to-4-gram phrases.

3.4 Evaluation

For evaluation, we follow the task-based approach taken in Section 2 and apply the similarity-scored

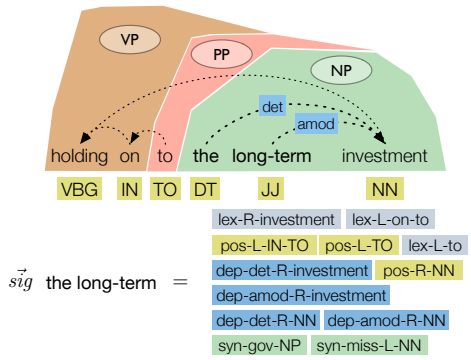


Figure 4: An example of the syntactic feature-set. The phrase "the long-term" is annotated with position-aware lexical and part-of-speech n-gram features, labeled dependency links, and features derived from the phrase's CCG label (NP/NN).

paraphrases to sentence compression. The distributional similarity scores are incorporated into the paraphrasing system as additional rule features into the log-linear model. The task-targeted parameter tuning thus results in a reranking of the rules that takes into consideration, the distributional information, bilingual alignment-based paraphrase probabilities, and compression-centric features.

Table 2 shows comparison of the bilingual baseline paraphrase grammar (PP), the reranked grammars based on signatures extracted from the Google n-grams (n-gram), the richer signatures drawn from Annotated Gigaword (Syntax), and Clarke and Lapata (2008)'s compression system (ILP). In both cases, the inclusion of distributional similarity information results in significantly better output grammaticality and meaning retention. Despite its lower coverage (12 versus 200 million phrases), the syntactic distributional similarity outperforms the simpler Google n-gram signatures.

3.5 PPDB

To facilitate a more widespread use of paraphrases, we release a collection of ranked paraphrases obtained by the methods outlined in Sections 2 and 3 to the public (Ganitkevitch et al., 2013).

4 Paraphrasing with Natural Logic

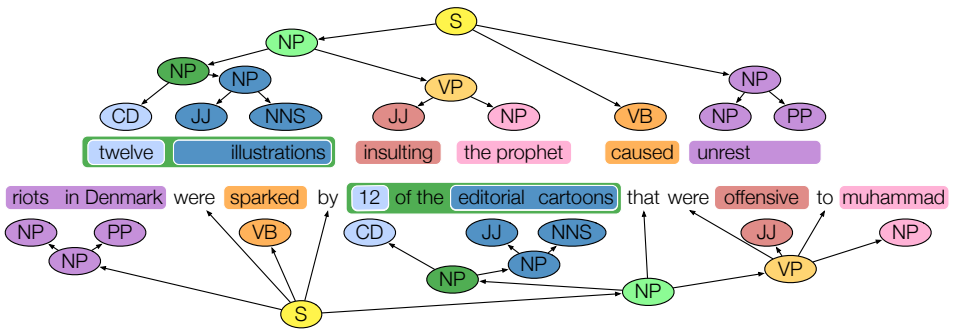
In the previously derived paraphrase grammar it is assumed that all rules imply the semantic equivalence of two textual expressions. The varying degrees of confidence our system has in this relationship are evidenced by the paraphrase probabilities and similarity scores. However, the grammar can also contain rules that in fact represent a range of semantic relationships, including hypernym- hyponym relationships, such as *India – this country*.

To better model such cases we propose an annotation of each paraphrase rule with *explicit relation labels* based on natural logic. Natural logic (MacCartney, 2009) defines a set of pairwise relations between textual expressions, such as equivalence ($\equiv$), forward ($\sqsubset$) and backward ($\sqsupset$) entailment, negation $\wedge$) and others. These relations can be used to not only detect semantic equivalence, but also infer entailment. Our resulting system will be able to tackle tasks like RTE, where the more a fine-grained resolution of semantic relationships is crucial to performance.

We favor a classification-based approach to this problem: for each pair of paraphrases in the grammar, we extract a feature vector that aims to capture information about the semantic relationship in the rule. Using a manually annotated development set of paraphrases with relation labels, we train a classifier to discriminate between the different natural logic relations.

We propose to leverage both labeled and unlabeled data resources to extract useful features for the classification. Annotated resources like WordNet can be used to derive a catalog of word and phrase pairs with known entailment relationships, for instance $\langle India, country, \sqsubset \rangle$. Using word alignments between our paraphrase pairs, we can establish what portions of a pair have labels in WordNet and retain corresponding features.

To leverage unlabeled data, we propose extending our notion of distributional similarity. Previously, we used cosine similarity to compare the signatures of two phrases. However, cosine similarity is a symmetric measure, and it is unlikely to prove helpful for determining the (asymmetric) entailment directionality of a paraphrase pair (i.e. whether it is a hypo- or hypernym relation). We therefore propose to extract a variety of asymmetric similarity features from distributional contexts. Specifically, we seek a measure that compares both the similarity and the "breadth" of two vectors. Assuming that wider breadth implies a hypernym, i.e. a $\sqsubset$-entailment, the scores produced by such a measure can be highly



Paraphrase rules			Entailment classification			
$CD \rightarrow$	twelve	12	twelve	$\equiv$	12
$JJ \rightarrow$	ϵ	editorial	ϵ	$\sqsupset$	editorial
$NNS \rightarrow$	illustrations	cartoons	illustrations	$\sqsupset$	cartoons
$JJ \rightarrow$	insulting	offensive	insulting	$\equiv / \sqsubset$	offensive
$NP \rightarrow$	the prophet	muhammad	the prophet	$\equiv$	muhammad
$VB \rightarrow$	caused	sparked	caused	$\sqsupset$	sparked
$NP \rightarrow$	unrest	riots	unrest	$\sqsupset$	riots
$PP \rightarrow$	ϵ	in Denmark	ϵ	$\sqsupset$	in Denmark
$NP \rightarrow$	$CD(\equiv)\ NNS(\sqsupset)$	$CD(\equiv)$ of the $NNS(\sqsupset)$	twelve illustrations	$\sqsupset$	12 of the cartoons

Figure 5: Our system will use synchronous parsing and paraphrase grammars to perform natural language inference. Each paraphrase transformation will be classified with a natural logic entailment relation. These will be joined bottom-up, as illustrated by the last rule, where the join of the smaller constituents $\equiv \bowtie \sqsupset$ results in $\sqsupset$ for the larger phrase pairs. This process will be propagated up the trees to determine if the hypothesis can be inferred from the premise.

informative for our classification problem. Asymmetric measures like Tversky indices (Tolias et al., 2001) appear well-suited to the problem. We will investigate application of Tversky indices to our distributional signatures and their usefulness for entailment relation classification.

4.1 Task-Based Evaluation

We propose evaluating the resulting system on textual entailment recognition. To do this, we cast the RTE task as a synchronous parsing problem, as illustrated in Figure 5. We will extend the notion of synchronous parsing towards resolving entailments, and define and implement a compositional join operator $\bowtie$ to compute entailment relations over synchronous derivations from the individual rule entailments.

While the assumption of a synchronous parse structure is likely to be valid for translations and paraphrases, we do not expect it to straightforwardly hold for entailment recognition. We will thus investigate the limits of the synchronous assumption over RTE data. Furthermore, to expand the system's coverage in a first step, we propose a simple relaxation of the synchronousness requirement via entailment-less "glue rules." These rules, similar to out-of-vocabulary rules in translation, will allow us to include potentially unrelated or unrecognized portions of the input into the synchronous parse.

5 Conclusion

We have described an extension of the state of the art in paraphrasing in a number of important ways: we leverage large bilingual data sets to extract linguistically expressive high-coverage paraphrases based on an SCFG formalism. On an example text-to-text generation task, sentence compression, we show that an easily adapted paraphrase system achieves state of the art meaning retention. Further, we include a complementary data source, monolingual corpora, to augment the quality of the previously obtained paraphrase grammar. The resulting system is shown to perform significantly better than the purely bilingual paraphrases, in both meaning retention and grammaticality, achieving results on par with the state of the art. Finally, we propose an extension of SCFG-based paraphrasing towards a more fine grained semantic representation using a classification-based approach. In extending the synchronous parsing methodology, we outline the expansion of the paraphraser towards a system capable of tackling entailment recognition tasks.

Acknowledgements

The ideas described in this paper were developed in collaboration with Benjamin Van Durme and Chris Callison-Burch. 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, the NSF, 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*.

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.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Poossin. 1990. A statistical approach to language translation. *Computational Linguistics*, 16(2), June.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

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

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

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:273–381.

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.

Juri Ganitkevitch, Chris Callison-Burch, and Benjamin Van Durme. 2013. Ppdb: The paraphrase database. In *Proceedings of HLT/NAACL*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

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*.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University.

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

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

Yannis A. Tolias, Stavros M. Panas, and Lefteri H. Tsoukalas. 2001. Generalized fuzzy indices for similarity matching. *Fuzzy Sets and Systems*, 120(2):255–270.