Monolingual Distributional Similarity for Texty Text Generation

Juri Ganitkevitch, Benjamin Van Durme and **Chris Callison-Burch**
Department of Computer Science, Johns Hopkins University
Baltimore, MD 21218, USA

Abstract

Previous work obtained collections of paraphrases by either relying on sentence-aligned parallel datasets, or by using distributional similarity metrics over large text corpora. Our approach combines these two orthogonal sources of information by directly integrating them into the decoding algorithm. We hope to report significant improvements in output quality on an array of text-to-text generation tasks.

1 Introduction

A wide variety of applications in natural language processing can be cast in terms of text-to-text generation. Given input in the form of natural language, a text-to-text generation system produces natural language output that fulfills previously defined constraints and objectives on both the text's surface form and meaning. Paraphrases, i.e. differing textual realizations of the same meaning, are a crucial components of text-to-text generation systems, and have been successfully applied to tasks such as multi-document summarization, query expansion, question answering, sentence compression and simplification (Barzilay et al., 1999; Barzilay, 2003; McKeown, 1979; Anick and Tipirneni, 1999; Ravichandran and Hovy, 2002; Riezler et al., 2007).

Recently, Ganitkevitch et al. (2011) presented an approach for sentential paraphrasing derived from state-of-the-art statistical machine translation systems. They described a large-scale extraction method for syntactically annotated paraphrases from bilingual parallel corpora, as well as an adaptation framework that allowed for straight-forward, non-naive adaptation of the system to any given sentential text-to-text generation task.

In this paper, we describe an extension of Ganitkevitch et al. (2011)'s approach by introducing a new component into the paraphrasing system that is not derived from the statistical machine translation domain and present an orthogonal source of information: monolingual distributional similarity. More specifically, we show that:

- Using monolingual distributional similarity features improves paraphrase quality past what we can achieve with features estimated from bilingual data. We demonstrate that different types of monolingual distributional information can be used to achieve differing effects such as improvements in grammaticality or word sense disambiguation, and discuss the trade-off between data sources with high-coverage versus smaller, more richly annotated corpora.
- We define the notion of distributional similarity for paraphrase patterns that contain multi-word gaps. This generalizes over previous approaches that defined the notion for contiguous phrases or single-word gaps.
- Finally, we compare the effectiveness of out method against a variety of baselines on an example text-to-text generation task, sentence compression. We show improvements in quality over both a purely bilingually sourced paraphrasing system and an ILP-based compression model.

In the following, we will give an overview of SCFG-based paraphrase extraction (Section 2) and

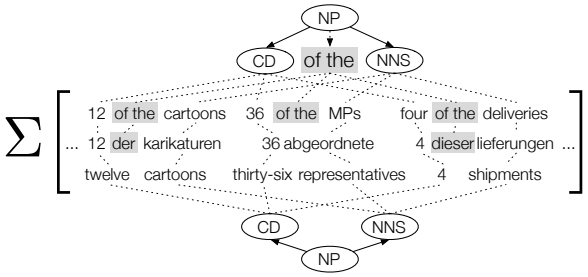

Figure 1: An example of syntactic paraphrase extraction and feature estimation via the pivoting approach.

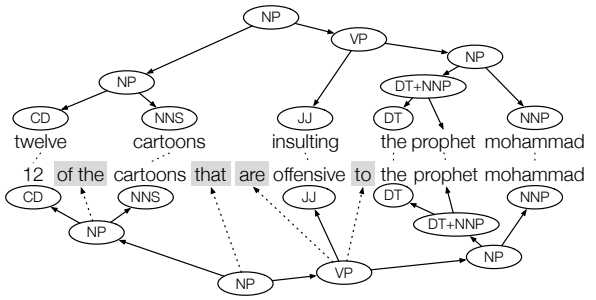

Figure 2: An example of a synchronous paraphrastic derivation.

monolingual distributional similarity (Section 3). Section 4 presents our rescoring model. We discuss the reranking results in Section 5. We relate our work to prior research in Section 6. Finally, Sections 7 and 8 present our experimental setup and the results obtained for the sentence compression task. We conclude in Section 9.

2 Synchronous Context-Free Grammars

Following Ganitkevitch et al. (2011), we formulate our paraphrases as a syntactically annotated *synchronous context-free grammar* (SCFG) (Aho and Ullman, 1972; Chiang, 2005). An SCFG rule has the form:

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

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 with an equal number of nonterminals. The function $\sim$ defines a one-to-one correspondency function between the nonterminals in f and e. Drawing on machine translation terminology, we refer to f as the *source* and e as the *target* side of the rule.

Each rule is annotated with a vector of feature functions $\vec{\varphi} = \{\varphi_1 ... \varphi_N\}$ that, using a corresponding weight vector $\vec{\lambda}$, are combined in a log-linear model to compute the *cost* of applying $\mathbf{r}$:

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

Typical features used in the statistical machine translation models that our system builds on are conditional phrasal, lexical and left-hand side label probabilities, as well as a variety of count and indicator features. We detail the feature set used in our experiments in Section 7.

To obtain a paraphrase grammar, we first must extract a translation grammar that translates any given foreign language into English. Then, for each pair of translation rules where the left-hand side C and foreign string f match:

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

we use the intuition that two english strings e_1 and e_2 that translate to the same foreign string f are equivalent in meaning, and *pivot* over f to create a paraphrase rule (Ganitkevitch et al., 2011; Callison-Burch, 2008; Bannard and Callison-Burch, 2005):

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

with a combined nonterminal correspondency function $\sim_p$. Similarly, 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:

$$p(e_2|e_1) = \sum_f p(e_2, f|e_1) \quad (2)$$
$$= \sum_f p(e_2|f, e_1) p(f|e_1) \quad (3)$$
$$\approx \sum_f p(e_2|f) p(f|e_1). \quad (4)$$

Figure 1 illustrates syntax-constrained pivoting and feature aggregation over multiple foreign language

translations for a paraphrase pattern. Figure 2 shows an example for a synchronous paraphrastic derivation produced as a result of applying our grammar in the decoding process.

The approach we outlined in this section relies on supervised sentence-level parallelism to identify phrases and patterns that are equivalent in meaning. When extracting paraphrases from monolingual text, we have to rely on an entirely different set of semantic cues and features.

3 Monolingual Distributional Similarity

In absence of other correspondency information, paraphrase extraction from monolingual corpora relies on contextual features. To describe a phrase e, we define a set of features that describe the context of an occurrence of e in our corpus. The resulting feature vectors $\vec{s}_{e,i}$ are aggregated 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 similar contexts, we can then identify e' as a paraphrase of e by computing the cosine similarity between their distributional signatures:

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

The features used to describe the context of a phrase differ by application and data source. Both Lin and Pantel (2001) and Church and Hanks (1991) use a rich feature set based on constituency and dependency parses of the text corpora they extract paraphrases from. In their work, a phrase is described by the various syntactic relations it has with lexical items in its context, such as the set of verbs it appears is seen as the subject of, or the set of adjectives that modify it.

However, when moving to vast text collections or collapsed representations of large text corpora, parsing can become impractical or even impossible. In these cases using simple features based on lexical n-grams has proven to be effective (Lapata and Keller, 2005; Bhagat and Ravichandran, 2008; Lin et al., 2010; Van Durme and Lall, 2010).

In order to investigate the impact of the feature set used, we chose to extract two collections of distributional similarity-based paraphrases. Using a web-scale n-gram corpus (Brants and Franz, 2006; Lin et al., 2010), we extract unigram features for the words to the left and right for phrases up to a length of 4. The features are weighed with the n-gram count given by the dataset. The resulting collection comprised context vectors for the 200 million most frequent 1- to 4-grams in the dataset.

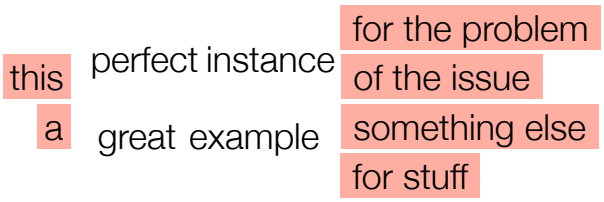

Figure 3: An example of immediate lexical context acquisition over n-gram corpora.

For contrast, we use the constituency- and dependency-parsed Los Angeles Times/Washington Post portion of the Gigaword corpus (Graff et al., 2003). The following feature set is used to compute phrase contexts over this dataset:

- Lexical and part-of-speech unigram and bigram features, drawn from a three-word window to the right and left of the phrase.
- Features based on dependencies for both links into and out of the phrase, labeled with the corresponding lexical item and POS. If the phrase is syntactically well-formed we additionally include lexical and POS features for its head.
- Syntactic features for constituents governing the phrase, as well as for CCG-style slashed constituent labels for the phrase, split by governing constituent and missing constituent.

Figure 4 illustrates our choice of feature set. As a result we obtain context information for over 12 million 1- to 4-gram phrases.

Much like Ravichandran et al. (2005) and Bhagat and Ravichandran (2008), we relied on Locality Sensitive Hashing (LSH), to make the use of these large collections practical. In order to avoid explicitly computing the feature vectors, which can be memory intensive for frequent phrases, we chose the online LSH variant described in (Van Durme and Lall, 2010). This method, based on the earlier work of Indyk and Motwani (1998) and Charikar (2002),

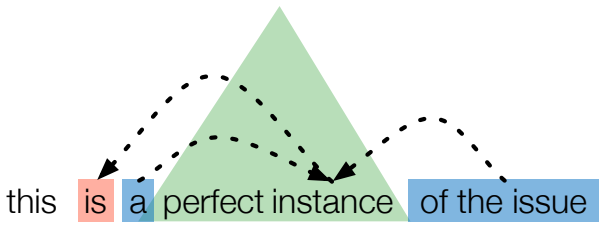

Figure 4: An example of rich monolingual distributional features over a parsed text corpus.

approximates the cosine similarity between two feature vectors based on the Hamming distance in a dimensionality-reduced bitwise representation. Two feature vectors u, v each of dimension d are first projected through a $d \times b$ random matrix populated with draws from $\mathcal{N}(0, 1)$. We then convert the resulting b-dimensional vectors into bit-vectors by setting each bit of the signature conditioned on whether the corresponding projected value is less than 0. Now, given the bit signatures $h(\vec{u})$ and $h(\vec{v})$, we approximate the cosine similarity of u and v as:

$$sim'(u, v) = \cos\left(\frac{D(h(\vec{u}), h(\vec{v}))}{b}\pi\right),$$

where $D()$ is the Hamming distance.

4 Incorporating Distributional Similarity

The approach to paraphrase extraction we outlined above relies on bitexts and the semantic cues we gather from the sentence-level alignments.

vast amounts of data

contextual information versus alignment

5 Paraphrase Ranking

6 Related Work

In this section we refer back to some prior work on

- Distributional similarity metrics in paraphrase acquisition.
- Use of distributional approaches in text-to-text and machine translation.

Some papers to be mentioning are here (Langkilde and Knight, 1998b; Langkilde and Knight, 1998a; Lin and Pantel, 2001; Bhagat and Ravichandran, 2008).

7 Experimental Setup

8 Sentence Compression Results

9 Conclusion

References

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

Peter G. Anick and Suresh Tipirneni. 1999. The paraphrase search assistant: terminological feedback for iterative information seeking. In *Proceedings of SIGIR*.

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

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of ACL*.

Regina Barzilay. 2003. *Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

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*.

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.

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*.

D. Graff, J. Kong, K. Chen, and K. Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*.

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

I. Langkilde and K. Knight. 1998a. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*,

pages 704–710. Association for Computational Linguistics.

I. Langkilde and K. Knight. 1998b. The practical value of n-grams in generation.

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*.

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

Deepak Ravichandran and Eduard Hovy. 2002. Learning sufrace text patterns for a question answering system. In *Proceedings of ACL.*

D. Ravichandran, P. Pantel, and E. Hovy. 2005. Randomized algorithms and nlp: Using locality sensitive hash functions for high speed noun clustering. In *Proceedings of ACL.*

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

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