Learning Structural and Sentential Paraphrases from Parallel Corpora

Juri Ganitkevitch and **Chris Callison-Burch**
Center for Language and Speech
Johns Hopkins University

Abstract

Later.

1 Introduction

Motivate paraphrasing within the field (same meaning, transformations, entailment).

2 Related Work

There have been many proposed approaches to paraphrase induction, most of which can be placed in two groups, according to what data they use to source their method. Monolingual approaches typically make use of vast amounts of English text to extract paraphrases. Due to the availability of good syntactic and dependency parsers, these methods are often able to infer *structural* paraphrases, i.e. paraphrastic patterns that capture either syntactic or semantic information and can be generalize via "slots". However, the coverage of such systems tends to be low.

...

Bilingually sourced approaches, on the other hand make use of the relative abundance of sentence-parallel corpora and extract bilingual tables of phrases (cite Chris) or patterns (cite Nitin, Zhao) from which paraphrases can be extracted by pivoting over the non-English side of the table. However, due to the nature of bilingual phrase tables, the resulting paraphrases are often restricted to surface-level. *not quite true, since Zhao does extract labled pattern based on dependency graphs. need to look into that and categorize, organize prior work properly.*

Should distinguish between phrasal and structural paraphrases, as well as bilingually and monolingually sourced approaches to extraction.

Perhaps: translation as bilingual paraphrasing, and how that thought brings about the pivot approach.

Reference monolingual approaches that are structural, pivot-based approaches that are phrasal, Nitin's stuff.

Note how our approach sort-of unifies the two; there's an analogy to Chris' syntactic constraints as well as the obvious step from Hiero to SAMT.

3 Sentential Paraphrasing

move to analysis?

We are interested in sentential paraphrases. Why are we interested? More powerful than locally constrained, gives us large-scale changes to sentential structure, which can be cruicial to applications such as detecting entailment or automatically creating significantly differing references. *However, phrasal decoder-based approaches should be able to achieve similar reordering effects (if not the generality, which we only implicitely achieve, really). Maybe we should add a phrase-based baseline in addition to Hiero? Did Nitin talk about this?*

While the definition of a phrasal paraphrase is intuitively clear, sentential paraphrases are much harder to define. When paraphrasing a sentence s into a new sentence t, the term suggests that we expect the changes to s to be above a certain threshold for t to be considered a sentential paraphrase.

4 To Add

Formal: Synchronous grammars, with the usual examples (one phrasal, one structural).

Formal: Log-linear model for features, weights will be optimized to some objective, we discuss those in later section.

Diagram for grammar extraction (two sentence pairs with trees and alignments, show how that gets us a paraphrase pattern).

5 Paraphrase Acquisition

The method we present in this work extends Nitin's pivot-based Hiero paraphrasing approach to the richer, syntax-informed SAMT formalism. Starting with a bilingual parallel corpus, we use the familiar MT and parsing (cite!) machinery to word-align the data and extract an SAMT-style grammar.

Elaborate on this, reference the appropriate work. This is more about the pivoting than it is about the MT-style application of the paraphrases.

In Section 8.1 we give a brief description of the data sets used in our experiments and outline the tools used to process them. Section 5.1 elaborates on the extraction of the bilingual translation grammars and their transformation into monolingual paraphrasing grammars.

5.1 Paraphrase Grammar Extraction

Some talk about the SAMT approach needs to go here.

We extract SAMT translation grammars for nine languages. Give some details on the pipeline, mention that the grammars are *gargantuan* (this word needs to be in the paper!), but that the whole process is very well-suited for MapReduce (even though, we didn't use it).

5.2 Creating Paraphrase Rules

5.2.1 Rule Body

To create paraphrase rules from bilingual translation rules, we pivot over the foreign side of the translation rule, with the additional constraint that the rule's head, i.e. the label that governs the rule.

Mention the proper mapping and flipping of nonterminals.

5.2.2 Rule Features

The SAMT grammars our paraphrasing system is based on provide a rich feature set that takes into account source- and target-side frequencies, reordering of NTs and lexical translation probablities for each rule. When transforming the bilingual grammars into a monolingual paraphrase grammar we preseve the feature set and

Shouldn't give details on every single feature, but point out some key approaches:

- probablistic features multiply
- target-side indicator features are inherited
- actually state what happens with other indicator features (re-ordering, punctuation, rareness etc)

6 Analysis

one of the motivations for our approach was the expectation of being able to acquire structural paraphrases, especially paraphrases taht are capable of learning long-distance transformations such as passivization, dative shift, possessive something and prepositional paraphrases.

7 Paraphrasing System

The paraphrase extraction described in the precious section yields an English-to-English paraphrase grammar

7.1 Training

We use MERT. It's great.

7.2 Objective Function

Och (MERT) has shown that it is best to tune to the objective function that will be used for evaluation..

7.2.1 Paraphrase BLEU

7.2.2 Summarization BLEU

yet to be thought out, but another example here would be great

get some data for summarization? other task?

8 Experimental Results

8.1 Data Collection

We use Europarl v. 5. Align with Berlekey and parse with The Parser.

8.2 Reference Expansion for SMT

A straightforward way to evaluate a paraphrasing system is by using to improve an SMT system's performance. Cite Chris and Nitin. Compare to Hiero baseline.

8.3 Possibly Another Paraphrase Application

Maybe summarization with examples? Is there any "hard" eval to be had? Human via MTurk?

9 Conclusion

Look, we unified everything in the field and made all this stuff from the previous section much better. Or did we?

References

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

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL*.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*.

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

Wauter Bosma and Chris Callison-Burch. 2006. Paraphrase substitution for recognizing textual entailment. In *Proceedings of CLEF*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT/NAACL*.

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

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

Trevor Cohn and Mirella Lapata. 2007. Large margin synchronous generation and its application to sentence compression. In *Proceedings of EMNLP-CoLing*.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research (JAIR)*, 34:637–674.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. to appear. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of 3rd International Workshop on Paraphrasing*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Mark Dras. 1997a. Reluctant paraphrase: Textual restructuring under an optimization model. In *PACLING-97*, pages 184–191.

Mark Dras. 1997b. Representing paraphrases using synchronous tree adjoining grammars. In *Proceedings of ACL*.

Mark Dras. 1998. Search in constraint-based paraphrasing. In *Proceedings of the Second International Conference on Natural Language Processing and Industrial Applications*, Moncton, Canada.

Mark Dras. 1999a. A meta-level grammar: Redefining synchronous TAGs for translation and paraphrase. In *Proceedings of ACL*, pages 98–104.

Mark Dras. 1999b. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Australia.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of ACL*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT/NAACL*.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*.

Lidija Iordanskaja, Richard Kittredge, and Alain Polgére. 1991. Lexical selection and paraphrase in a meaning text generation model. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of EMNLP*.

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

Irene Langkilde and Kevin Knight. 1998. The practical value of n-grams in generation. In *Workshop On Natural Language Generation*, Ontario, Canada.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules from text. *Natural Language Engineering*, 7(3):343–360.

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

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL/HLT*.

Bonnie Dorr Matthew Snover, Nitin Madnani and Richard Schwartz. 2009. Ter-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation Journal*.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*.

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

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

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

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008b. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL/HLT*.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006a. Re-evaluating machine translation results with paraphrase support. In *Proceedings of EMNLP*.

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

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the NAACL-2006 Workshop on Statistical Machine Translation (WMT-06)*, New York, New York.