Comparing Phrase-based and Syntax-based Paraphrase Generation
Wubben, S.; Marsi, E.C.; van den Bosch, A.; Krahmer, E.J.

Published in:
Proceedings of the ACL-HLT Workshop on Monolingual Text-To-Text Generation

Document version:
Publisher's PDF, also known as Version of record

Publication date:
2011

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Comparing Phrase-based and Syntax-based Paraphrase Generation

Sander Wubben
Tilburg University
P.O. Box 90135
5000 LE Tilburg
The Netherlands
s.wubben@uvt.nl

Erwin Marsi
NTNU
Sem Saelandsvei 7-9
NO-7491 Trondheim
Norway
emarsi@idi.ntnu.no

Antal van den Bosch
Tilburg University
P.O. Box 90135
5000 LE Tilburg
The Netherlands
antal.vdnbosch@uvt.nl

Emiel Krahmer
Tilburg University
P.O. Box 90135
5000 LE Tilburg
The Netherlands
e.j.krahmer@uvt.nl

Abstract

Paraphrase generation can be regarded as machine translation where source and target language are the same. We use the Moses statistical machine translation toolkit for paraphrasing, comparing phrase-based to syntax-based approaches. Data is derived from a recently released, large scale (2.1M tokens) paraphrase corpus for Dutch. Preliminary results indicate that the phrase-based approach performs better in terms of NIST scores and produces paraphrases at a greater distance from the source.

1 Introduction

One of the challenging properties of natural language is that the same semantic content can typically be expressed by many different surface forms. As the ability to deal with paraphrases holds great potential for improving the coverage of NLP systems, a substantial body of research addressing recognition, extraction and generation of paraphrases has emerged (Androutsopoulos and Malakasiotis, 2010; Madnani and Dorr, 2010). Paraphrase Generation can be regarded as a translation task in which source and target language are the same. Both Paraphrase Generation and Machine Translation (MT) are instances of Text-To-Text Generation, which involves transforming one text into another, obeying certain restrictions. Here these restrictions are that the generated text must be grammatically well-formed and semantically/translationally equivalent to the source text. Additionally Paraphrase Generation requires that the output should differ from the input to a certain degree.

The similarity between Paraphrase Generation and MT suggests that methods and tools originally developed for MT could be exploited for Paraphrase Generation. One popular approach – arguably the most successful so far – is Statistical Phrase-based Machine Translation (PBMT), which learns phrase translation rules from aligned bilingual text corpora (Och et al., 1999; Vogel et al., 2000; Zens et al., 2002; Koehn et al., 2003). Prior work has explored the use of PBMT for paraphrase generation (Quirk et al., 2004; Bannard and Callison-Burch, 2005; Madnani et al., 2007; Callison-Burch, 2008; Zhao et al., 2009; Wubben et al., 2010).

However, since many researchers believe that PBMT has reached a performance ceiling, ongoing research looks into more structural approaches to statistical MT (Marcus and Wong, 2002; Och and Ney, 2004; Khalilov and Fonollosa, 2009). Syntax-based MT attempts to extract translation rules in terms of syntactic constituents or subtrees rather than arbitrary phrases, presupposing syntactic structures for source, target or both languages. Syntactic information might lead to better results in the area of grammatical well-formedness, and unlike phrase-based MT that uses contiguous n-grams, syntax enables the modeling of long-distance translation patterns.

While the verdict on whether or not this approach leads to any significant performance gain is still out, a similar line of reasoning would suggest that syntax-based paraphrasing may offer similar advantages over phrase-based paraphrasing. Considering the fact that the success of PBMT can partly be attributed to the abundance of large parallel corpora,
and that sufficiently large parallel corpora are still lacking for paraphrase generation, using more linguistically motivated methods might prove beneficial for paraphrase generation. At the same time, automatic syntactic analysis introduces errors in the parse trees, as no syntactic parser is perfect. Likewise, automatic alignment of syntactic phrases may be prone to errors.

The main contribution of this paper is a systematic comparison between phrase-based and syntax-based paraphrase generation using an off-the-shelf statistical machine translation (SMT) decoder, namely Moses (Koehn et al., 2007) and the word-alignment tool GIZA++ (Och and Ney, 2003). Training data derives from a new, large scale (2.1M tokens) paraphrase corpus for Dutch, which has been recently released.

The paper is organized as follows. Section 2 reviews the paraphrase corpus from which provides training and test data. Next, Section 3 describes the paraphrase generation methods and the experimental setup. Results are presented in Section 4. In Section 5 we discuss our findings and formulate our conclusions.

2 Corpus

The main bottleneck in building SMT systems is the need for a substantial amount of parallel aligned text. Likewise, exploiting SMT for paraphrasing requires large amounts of monolingual parallel text. However, paraphrase corpora are scarce; the situation is more dire than in MT, and this has caused some studies to focus on the automatic harvesting of paraphrase corpora. The use of monolingual parallel text corpora was first suggested by Barzilay and McKeown (2001), who built their corpus using various alternative human-produced translations of literary texts and then applied machine learning or multi-sequence alignment for extracting paraphrases. In a similar vein, Pang et al. (2003) used a corpus of alternative English translations of Chinese news stories in combination with a syntax-based algorithm that automatically builds word lattices, in which paraphrases can be identified.

So-called comparable monolingual corpora, for instance independently written news reports describing the same event, in which some pairs of sentences exhibit partial semantic overlap have also been investigated (Shinyama et al., 2002; Barzilay and Lee, 2003; Shen et al., 2006; Wubben et al., 2009).

The first manually collected paraphrase corpus is the Microsoft Research Paraphrase (MSRP) Corpus (Dolan et al., 2004), consisting of 5,801 sentence pairs, sampled from a larger corpus of news articles. However, it is rather small and contains no sub-sentential alignments. Cohn et al. (2008) developed a parallel monolingual corpus of 900 sentence pairs annotated at the word and phrase level. However, all of these corpora are small from an SMT perspective.

Recently a new large-scale paraphrase corpus for Dutch, the DAESO corpus, was released. The corpus contains both samples of parallel and comparable text in which similar sentences, phrases and words are aligned. One part of the corpus is manually aligned, whereas another part is automatically aligned using a data-driven aligner trained on the first part. The DAESO corpus is extensively described in (Marsi and Krahmer, 2011); the summary here is limited to aspects relevant to the work at hand.

The corpus contains the following types of text: (1) alternative translations in Dutch of three literary works of fiction; (2) autocue text from television broadcast news as read by the news reader, and the corresponding subtitles; (3) headlines from similar news articles obtained from Google News Dutch; (4) press releases about the same news topic from two different press agencies; (5) similar answers retrieved from a document collection in the medical domain, originally created for evaluating question-answering systems.

In a first step, similar sentences were automatically aligned, after which alignments were manually corrected. In the case of the parallel book texts, aligned sentences are (approximate) paraphrases. To a lesser degree, this is also true for the news headlines. The autocue-subtitle pairs are mostly examples of sentence compression, as the subtitle tends to be a compressed version of the read autocue text. In contrast, the press releases and the QA answers, are characterized by a great deal of one-to-many sentence alignments, as well as sentences left unaligned, as is to be expected in comparable text. Most sentences in these types of text tend to have only partial overlap in meaning.
Table 1: Properties of the manually aligned corpus

<table>
<thead>
<tr>
<th></th>
<th>Autosub</th>
<th>Books</th>
<th>Headlines</th>
<th>News</th>
<th>QA</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>aligned trees</td>
<td>18,338</td>
<td>6,362</td>
<td>32,627</td>
<td>11,052</td>
<td>118</td>
<td>68,497</td>
</tr>
<tr>
<td>tokens</td>
<td>217,959</td>
<td>115,893</td>
<td>179,629</td>
<td>162,361</td>
<td>2,230</td>
<td>678,072</td>
</tr>
<tr>
<td>tokens/sent</td>
<td>11.89</td>
<td>18.22</td>
<td>5.51</td>
<td>14.69</td>
<td>18.90</td>
<td>9.90</td>
</tr>
<tr>
<td>nodes</td>
<td>365,157</td>
<td>191,636</td>
<td>318,399</td>
<td>271,192</td>
<td>3734</td>
<td>1,150,118</td>
</tr>
<tr>
<td>nodes/tree</td>
<td>19.91</td>
<td>30.12</td>
<td>9.76</td>
<td>24.54</td>
<td>31.64</td>
<td>16.79</td>
</tr>
<tr>
<td>uniquely aligned trees (%)</td>
<td>92.93</td>
<td>92.49</td>
<td>84.57</td>
<td>63.61</td>
<td>50.00</td>
<td>84.10</td>
</tr>
<tr>
<td>aligned nodes (%)</td>
<td>73.53</td>
<td>66.83</td>
<td>73.58</td>
<td>53.62</td>
<td>38.62</td>
<td>67.62</td>
</tr>
</tbody>
</table>

Next, aligned sentences were tokenized and parsed with the Alpino parser for Dutch (Bouma et al., 2001). The parser provides a relatively theory-neutral syntactic analysis which is a blend of phrase structure analysis and dependency analysis, with a backbone of phrasal constituents and arcs labeled with syntactic function/dependency labels.

The alignments not only concern paraphrases in the strict sense, i.e., expressions that are semantically equivalent, but extend to expressions that are semantically similar in less strict ways, for instance, where one phrase is either more specific or more general than the related phrase. For this reason, alignments are also labeled according to a limited set of semantic similarity relations. Since these relations were not used in the current study, we will not discuss them further here.

The corpus comprises over 2.1 million tokens, 678 thousand of which are manually annotated and 1,511 thousand are automatically processed.

To give a more complete overview of the sizes of different corpus segments, some properties of the manually aligned corpus are listed in Table 1. Properties of the automatically aligned part are similar, except for the fact that it only contains text of the news and QA type.

3 Paraphrase generation

Phrase-based MT models consider translation as a mapping of small text chunks, with possible reordering (Och and Ney, 2004). Operations such as insertion, deletion and many-to-one, one-to-many or many-to-many translation are all covered in the structure of the phrase table. Phrase-based models have been used most prominently in the past decade, as they have shown to outperform other approaches (Callison-Burch et al., 2009).

One issue with the phrase-based approach is that recursion is not handled explicitly. It is generally acknowledged that language contains recursive structures up to certain depths. So-called hierarchical models have introduced the inclusion of non-terminals in the mapping rules, to allow for recursion (Chiang et al., 2005). However, using a generic non-terminal X can introduce many substitutions in translations that do not make sense. By making the non-terminals explicit, using syntactic categories such as NPs and VPs, this phenomenon is constrained, resulting in syntax-based translation. Instead of phrase translations, translation rules in terms of syntactic constituents or subtrees are extracted, presupposing the availability of syntactic structures for source, target, or both languages.

Incorporating syntax can guide the translation process and unlike phrase-based MT syntax it enables the modeling of long-distance translation patterns. Syntax-based systems may parse the data on the target side (string-to-tree), source side (tree-to-string), or both (tree-to-tree).

In our experiments we use tree-to-tree syntax-based MT. We also experiment with relaxing the parses by a method proposed under the label of syntax-augmented machine translation (SAMT), described in (Zollmann and Venugopal, 2006). This method combines any neighboring nodes and labels previously unlabeled nodes, removing the syntactic constraint on the grammar.\(^1\)

We train all systems on the DAESO data (218,102 lines of aligned sentences) and test on a held-out set consisting of manually aligned headlines that ap-

---

\(^1\)This method is implemented in the Moses package in the program relax-parse as option SAMT 4
Table 2: Examples of output of the phrase-based and syntax-based systems

<table>
<thead>
<tr>
<th>Source</th>
<th>Phrase-based</th>
<th>Syntax-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>jongen (7) zwaargewond na aanrijding</td>
<td>boy (7) severely injured after crash</td>
<td>boy (7) severely injured after crash</td>
</tr>
<tr>
<td>7-jarige gewond na botsing</td>
<td>7-year-old injured after collision</td>
<td>7-year-old injured after collision</td>
</tr>
<tr>
<td>jongen (7) zwaar gewond na aanrijding</td>
<td>boy u.v severely injured after crash</td>
<td>boy u.v severely injured after crash</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Phrase-based</th>
<th>Syntax-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>jeugdwerkloosheid daalt vooral bij voldoende opleiding</td>
<td>youth-unemployment drops especially with adequate training</td>
<td>youth-unemployment drops especially with sufficient study</td>
</tr>
<tr>
<td>werkloosheid jongeren daalt, vooral bij voldoende studie</td>
<td>unemployment youths drops, especially with sufficient study</td>
<td>youth-unemployment drops especially in adequate training</td>
</tr>
<tr>
<td>* jeugdwerkloosheid daalt vooral in voldoende opleiding</td>
<td>criticism of fining-policy ns</td>
<td>criticism of controversial and youth-unemployment drops especially in adequate training</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Phrase-based</th>
<th>Syntax-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>weer bestuurders radboud weg</td>
<td>again directors radboud leaving</td>
<td>again directors leaving hospital</td>
</tr>
<tr>
<td>* weer de weg ziekenhuis</td>
<td>again the leaving hospital</td>
<td>yet-again directors hospital leaving</td>
</tr>
</tbody>
</table>

We test on 773 headlines that have three or more aligned paraphrasing reference headlines. We use an SRILM (Stolcke, 2002) language model trained on the Twente news corpus.

To investigate the effect of the amount of training data on results, we also train a phrase-based model on more data by adding more aligned headlines originating from data crawled in 2010 and aligned using tf.idf scores over headline clusters and Cosine similarity as described in (Wubben et al., 2009), resulting in an extra 612,158 aligned headlines.

Evaluation is based on the assumption that a good paraphrase is well-formed and semantically similar but structurally different from the source sentence. We therefore score the generated paraphrases not only by an MT metric (we use NIST scores), but also factor in the edit distance between the input sentence and the output sentence. We take the 10-best generated paraphrases and select from these the one most dissimilar from the source sentence in terms of Levenshtein distance on tokens. We then weigh NIST scores according to their corresponding sentence Levenshtein Distance, to calculate a weighted average score. This implies that we penalize systems that provide output at Levenshtein distance 0, which are essentially copies of the input, and not paraphrases. Formally, the score is computed as follows:

\[ NIST_{weighted_{LD}} = \frac{\sum_{i=LD(1..8)} (i * N_i * NIST_i)}{\sum_{i=LD(1..8)} (i * N_i)} \]

where \( \alpha \) is the percentage of output phrases that have a sentence Levenshtein Distance higher than 0. Instead of NIST scores, other MT evaluation scores can be plugged into this formula, such as METEOR (Lavie and Agarwal, 2007) for languages for which paraphrase data is available.

4 Results

Figure 1 shows NIST scores per Levenshtein Distance. It can be observed that overall the NIST score decreases as the distance to the input increases, indicating that more distant paraphrases are of less quality. The relaxed syntax-based approach (SAMT) performs mildly better than the standard syntax-based approach, but performs worse than the phrase-based approach. The distribution of generated paraphrases per Levenshtein Distance is shown in Figure 2. It reveals that the Syntax-based approaches tend to stay closer to the source than the phrase-based approaches.

In Table 2 a few examples of output from both Phrase- and Syntax-based systems are given. The
top two examples show sentences where the phrase-based approach scores better, and the bottom two show examples where the syntax-based approach scores better. In general, we observe that the phrase-based approach is often more drastic with its changes, as shown also in Figure 2. The syntax-based approach is less risky, and reverts more to single-word substitution.

The weighted NIST score for the phrase-based approach is 7.14 versus 6.75 for the syntax-based approach. Adding extra data does not improve the phrase-based approach, as it yields a score of 6.47, but the relaxed method does improve the syntax-based approach (7.04).

5 Discussion and conclusion

We have compared a phrase-based MT approach to paraphrasing with a syntax-based MT approach. The Phrase-based approach performs better in terms of NIST score weighted by edit distance of the output. In general, the phrase-based MT system performs more edits and these edits seem to be more reliable than the edits done by the Syntax-based approach. A relaxed Syntax-based approach performs better, while adding more data to the Phrase-based approach does not yield better results. To gain a better understanding of the quality of the output generated by the different approaches, it would be desirable to present the output of the different systems to human judges. In future work, we intend to compare the effects of using manual word alignments from the DAESO corpus instead of the automatic alignments produced by GIZA++. We also wish to further explore the effect of the nature of the data that we train on: the DAESO corpus consists of various data sources from different domains. Our aim is also to incorporate the notion of dissimilarity into the paraphrase model, by adding dissimilarity scores to the model.
References


32


