Creating and using large monolingual parallel corpora for sentential paraphrase generation

Sander Wubben, Antal van den Bosch, Emiel Krahmer
Tilburg University, Radboud University Nijmegen, Tilburg University
Warandelaan 2 Tilburg, Comeniuslaan 4 Nijmegen, Warandelaan 2 Tilburg
s.wubben@uvt.nl, a.vandenbosch@let.ru.nl, e.j.krahmer@uvt.nl

Abstract

In this paper we investigate the automatic generation of paraphrases by using machine translation techniques. Three contributions we make are the construction of a large paraphrase corpus for English and Dutch, a re-ranking heuristic to use machine translation for paraphrase generation and a proper evaluation methodology. A large parallel corpus is constructed by aligning clustered headlines that are scraped from a news aggregator site. To generate sentential paraphrases we use a standard phrase-based machine translation (PBMT) framework modified with a re-ranking component (henceforth PBMT-R). We demonstrate this approach for Dutch and English and evaluate by using human judgements collected from 76 participants. The judgments are compared to two automatic machine translation evaluation metrics. We observe that as the paraphrases deviate more from the source sentence, the performance of the PBMT-R system degrades less than that of the word substitution baseline system.

Keywords: paraphrasing, parallel corpora, monolingual machine translation

1. Introduction

Paraphrasing can be defined as transforming a word, phrase, sentence or longer text segment in a language from its original surface form to an alternative surface form in the same language that still expresses approximately the same semantic content as the original. The use of paraphrase generation has been demonstrated to be valuable for question answering (Lin and Pantel, 2001; Riezler et al., 2007), machine translation (Callison-Burch et al., 2006; Marton et al., 2009) and the evaluation thereof (Kauchak and Barzilay, 2006; Zhou et al., 2006; Pado et al., 2009). Adding certain constraints to paraphrasing allows for additional useful applications. When the constraint is specified that a paraphrase should be shorter than the input text, paraphrasing can be used for sentence compression (Knight and Marcu, 2002; Barzilay and Lee, 2003). Another specific task that can be approached this way is text simplification (Zhu et al., 2010; Woodsend and Lapata, 2011; Wubben et al., 2012), to convert for example medical terms into layperson's English (Elhadad and Sutaria, 2007; Deléger et al., 2009), or for subtitle generation (Daelemans et al., 2004).

Two important problems arise when developing a system that learns to generate paraphrases automatically from examples, namely how to obtain a sufficient number of examples to train the system on, and how to evaluate properly. We present a paraphrase corpus composed of data scraped from Google News to create a parallel corpus, and a standard PBMT framework modified with a re-ranking component (PBMT-R) to learn phrase alignments and generate paraphrases. We demonstrate this approach on Dutch and English and perform an extensive evaluation using human judgements collected from 76 participants, as well as two automatic machine translation evaluation metrics. Our approach can easily be adapted to other languages.

1.1. Phrase-based machine translation (PBMT) for paraphrasing

Sentential paraphrase generation can be approached as a monolingual machine translation task, where the source and target languages are the same
(Quirk et al., 2004; Bannard and Burch, 2005; Callison-Burch, 2008; Zhao et al., 2009; Wubben et al., 2010) and where the output should be different in form from the input but similar in meaning. Statistical machine translation (SMT) typically makes use of large parallel corpora to train a model on. These corpora need to be aligned at the sentence level. Large parallel corpora, such as the multilingual proceedings of the European Parliament (Europarl), are readily available for many languages.
Phrase-based machine translation (PBMT) is a form of SMT where the translation model aims to translate longer sequences of words ("phrases") in one go, solving part of the word ordering problem along the way that would be left to the decoder and the target language model in a word-based SMT system (Koehn et al., 2003). One advantage of PBMT is that it is adaptable to any language pair for which there is a parallel corpus available. The PBMT model makes use of a translation model, derived from the parallel corpus, and a language model, derived from a monolingual corpus in the target language. The language model is typically an n-gram model with smoothing. In principle, all of this should be transportable to a data-driven machine translation account of paraphrasing. For this to work, however, a preferably large collection of data is required, which in this case would be pairs of sentences that paraphrase each other.

1.2. Parallel corpora for paraphrasing

Two recently published surveys on paraphrasing address the need for paraphrase corpora to further develop research into paraphrasing (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010). (Androutsopoulos and Malakasiotis, 2010) observe that not many such parallel corpora currently exist, and that the ones that do exist are not even close to the size of corpora generally used to train statistical machine translation systems (Androutsopoulos and Malakasiotis, 2010). (Barzilay and McKeown, 2001) suggest building parallel paraphrase corpora by using multiple human translations of literary works originally written in a different language (Barzilay and McKeown, 2001). The fact that different translators may use different wordings can be exploited to find paraphrase pairs within a language. In general, for the machine translation approach to paraphrasing to work, first the texts need to be aligned at the sentence level to obtain sentence pairs that can be used in a parallel monolingual corpus, where each sentence in translation T_1 is ideally semantically equivalent to each sentence in translation T_2 in language L.

(Shinyama et al., 2002) use named entity recognition to extract paraphrases from various news articles describing the same event. The Microsoft Research Paraphrase Corpus (MSR) (Quirk et al., 2004; Dolan et al., 2004; Nelken and Shieber, 2006) is a paraphrase corpus constructed in an unsupervised manner. The MSR contains 5,801 pairs of sentences that were extracted from news sources on the Web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship. Of these sentences, the judges agreed that 67% were indeed paraphrases. (Cohn et al., 2008) developed a monolingual parallel corpus consisting of 900 sentence pairs annotated with alignments at the word and phrase level, which also contained sentences from the MSR (Cohn et al., 2008).

While these corpora are valuable, for a statistical paraphrasing approach to work they are generally several orders too small. Preferably, such systems are trained on hundreds of thousands to millions of parallel sentences, where the available paraphrase corpora contain several thousand sentences at best. One solution to this problem is to leverage the abundance of bilingual parallel corpora to find paraphrases. (Bannard and Burch, 2005) use a bilingual corpus and a pivot language to construct a monolingual phrase-table (Bannard and Burch, 2005). They do this by aligning phrases across the two languages and then harvesting all phrases aligned to one phrase in the pivot language as paraphrases. In contrast to using a pivot language, we demonstrate that it is possible to construct a sufficiently large parallel corpus without relying on a second language, but by harvesting different headlines for the same event. This has several advantages. One reason to use headlines is that these are abundant on the Web in many languages, and every day new ones appear describing real world events. The real world knowledge implicitly present in the system stays up to date this way: it will know that in this time frame (early 2014) *"Barack Obama"* can be paraphrased as *"The President of the United States"*. A more crucial reason is that there is much paraphrastic variety in headlines. Different journalists and news editors will try to come up with their own unique headlines that describe the same event. Another reason is that we have less of a problem dealing with sentence alignment between two texts to construct the parallel corpus, because headlines can be clustered relatively accurately by news aggregators such as Google News. Finally, headlines tend to be shorter than regular sentences and therefore words and phrases in them are easier to align.

1.3. Evaluation

As (Callison-Burch et al., 2008) argue, automatic evaluation of paraphrasing is problematic (Callison-Burch et al., 2008). The essence of paraphrasing is to be able to generate a sentence that paraphrases a source, but that is at the same time structurally different from that source. Automatic evaluation metrics in related fields such as standard multilingual machine translation (e.g. BLEU (Papineni et al., 2002)) operate on a notion of joint semantic and structural similarity, while paraphrasing aims to achieve semantic similarity, but also structural dissimilarity. As (Madnani and Dorr, 2010) rightfully observe, precision and recall are not suited when evaluating sentential paraphrase generation, because no exhaustive list of paraphrases can exist (Madnani and Dorr, 2010). There have been efforts to develop automatic metrics for the evaluation of paraphrases, such as ParaMetric (Callison-Burch, 2008) and PEM (Liu et al., 2010). ParaMetric is used to measure performance in alignment between two given sentences, and is not suited to measure the performance of a sentential paraphrase generation method given unseen sentences. PEM (Paraphrase Evaluation Metric) seems a promising approach in that it addresses the three crucial parts in paraphrase evaluation, namely fluency, adequacy and to some extent structural dissimilarity (PEM measures lexical dissimilarity). PEM makes no use of reference paraphrases; rather, it makes use of bilingual parallel corpora through the pivot approach. This suggests it might be biased towards paraphrasing approaches that use statistical machine translation and in particular pivot approaches. Another approach is to look at dissimilarity to the source sentence in addition to similarity to a collection of reference paraphrases. This is the approach we take and which has also been investigated by (Chen and Dolan, 2011). Chen and Dolan propose a new metric called PINC, which can be seen as a complement to BLEU: it measures the n-gram overlap between output and source sentence. The higher the overlap, the lower the PINC score. The idea is that good paraphrases show a high amount of overlap with reference paraphrases, and low overlap with the source sentence. We evaluate the output of our system by comparing it to a word substitution baseline, which uses a semantic lexicon and a language model to perform edit operations to construct a paraphrasing sentence, and a randomly selected human authored paraphrasing headline. We do this for Dutch and English and let 76 participants rate the paraphrases. We also take into account automatic machine translation evaluation metrics to see whether these correlate with human judgements, and show the results at different edit distances.

2. Data collection

For the development of our data collection method we use headline data from the DAESO corpus[1], a parallel monolingual treebank for Dutch (Marsi and Krahmer, 2007). Part

[1] http://daeso.uvt.nl/

of the data in the DAESO corpus consists of headline clusters scraped from Google News in the period April–August 2006. Google News uses clustering algorithms that consider the full text of each news article, as well as other features such as temporal and category cues, to produce sets of topically related articles. The scraper stores the headline and the first 150 characters of each news article scraped from the Google News Website. Roughly 13,000 clusters were retrieved. It is clear that although clusters deal roughly with one subject, the headlines can represent quite a different perspective on the content of the article; certain headlines are paraphrases, others are clearly not. To obtain only paraphrase pairs, the clusters need to be more coherent. In the DAESO project 865 clusters were manually subdivided into sub-clusters of headlines that show clear semantic overlap.

With these data we develop a method to extract paraphrase pairs from headline clusters. We divide the annotated 865 headline clusters in a development set of 40 clusters, while the remaining 825 are used as test data. The headlines are stemmed using the Porter stemmer for Dutch (Kraaij and Pohlmann, 1994). Instead of a word overlap measure as used by (Barzilay and Elhadad, 2003), we use a modified TF.IDF word score as suggested by (Nelken and Shieber, 2006).

2.1. Pairwise similarity

Our approach for aligning paraphrasing headlines is to directly calculate similarities for each pair of headlines within a cluster. If the similarity exceeds a certain threshold, the pair is accepted as a paraphrase pair. If it is below the threshold, it is rejected. However, as (Barzilay and Elhadad, 2003) have pointed out, this type of sentence alignment is only effective to a certain extent. Beyond that point, context is needed. With this in mind, we adopt two thresholds and the cosine similarity function to calculate the similarity between two sentences. If the similarity is higher than the upper threshold, it is accepted. If it is lower than the lower threshold, it is rejected. In the remaining case of a similarity between the two thresholds, similarity is calculated over the contexts of the two headlines, namely the text snippet that was retrieved with the headline. If this similarity exceeds the upper threshold, it is accepted. Threshold values as found by optimizing on the development data using again an $F_{0.25}$-score, are $Th_{lower} = 0.2$ and $Th_{upper} = 0.5$. An optional final step is to add transitive alignments. For instance, if headline A is paired with headline B, and headline B is aligned to headline C, headline A can be aligned to C as well. We do not add these alignments, because when one incorrect alignment is made, this process adds a large number of incorrect alignments, particularly in large clusters.

We extract paraphrasing headline pairs from new expanded datasets this way consisting of roughly 51,000 English headline clusters and 31,000 Dutch headline clusters, scraped from Google News in 2006 and in 2010. This method produces a collection of 9.3 million pairwise alignments of 1.9 million unique headlines for English and 841,588 pairwise alignments of 394,056 unique headlines for Dutch[2]. To our knowledge this new paraphrase source is several orders larger than existing paraphrase corpora.

3. Paraphrase generation

We use the collection of automatically obtained aligned headlines to train a paraphrase generation model using a phrase-based machine translation (PBMT) framework, extended with a post-hoc re-ranking model based on dissimilarity, resulting in our model PBMT-R. We compare this approach to a word substitution baseline. The generated paraphrases along with their source headlines are presented to human judges, whose ratings are compared to a collection of automatic machine translation evaluation metrics.

3.1. PBMT-R

We use the Moses software to train a PBMT model (Koehn et al., 2007).The GIZA++ statistical alignment package is used to perform the word alignments, which are later combined into phrase alignments in the Moses pipeline (Och and Ney, 2003) to build the paraphrase model. GIZA++ implements IBM Models 1 to 5 and an HMM word alignment model to find statistically motivated alignments between words. We first tokenize our data before training a re-caser. We then lowercase all data and use all unique headlines in the training data to train an n-gram language model with the SRILM toolkit (Stolcke, 2002). Then we invoke the GIZA++ aligner using the training paraphrase pairs. We run GIZA++ with standard settings and we perform no optimization. Finally, we use the Moses decoder to generate paraphrases for our test data.

We perform post-hoc re-ranking on the output based on dissimilarity to the input, as described earlier in (Wubben et al., 2012). We do this to select output that is as different as possible from the source sentence, so that ideally multiple phrases are paraphrased; at the same time, we base our re-ranking on a top-n of output candidates according to Moses, with a small n, to ensure that the quality of the output in terms of fluency and adequacy is also controlled for. Setting $n = 10$, for each source sentence we re-rank the ten best sentences as scored by the decoder according to the Levenshtein Distance (or edit distance) measure (Levenshtein, 1966) at the word level between the input and output sentence, counting the minimum number of edits needed to transform the source string into the target string, where the allowable edit operations are insertion, deletion, and substitution of a single word and casing is ignored. In case of a tie in Levenshtein Distance, we select the sequence with the better decoder score. When Moses is unable to generate ten different sentences, we select from the lower number of outputs. The resulting headlines are de-tokenized and recased using the previously trained re-caser.

3.2. Word substitution baseline

3.2.1. English

The PBMT-R results are compared with a word substitution baseline. For each noun, adjective and verb in the sentence this model takes that word and its part-of-speech tag and

[2] The aligned headline collection can be found at http://ilk.uvt.nl/ swubben/resources.html

retrieves from the English WordNet (Fellbaum, 1998) all synonyms from all synsets the word occurs in. The English WordNet contains over 200K word-sense pairs. The word is then replaced by all of its synset words, and each replacement is scored by the trained SRILM language model also used in the PBMT-R system. The highest scoring alternative is kept. If no relevant alternative is found, the word is left unaltered. We use the Memory Based Tagger (Daelemans et al., 1996) trained on the Brown corpus to compute the part-of-speech tags. The WordNet::QueryData[3] Perl module is used to query WordNet.

3.2.2. Dutch

The word substitution baseline for Dutch works similarly to the English baseline and relies on the Cornetto database instead of WordNet. Cornetto is a lexical semantic database for Dutch, similar to WordNet. It includes 40K entries, covering the most generic and central part of the Dutch language (Vossen et al., 2008). As with the English system, all synonyms for a given word are extracted and the synonym which scores best in the sentence according to the language model is kept. The SRILM language model is trained on the Dutch headline paraphrase corpus.

4. Evaluation

A human judgement study was set up to evaluate the generated paraphrases by both the baseline and the PBMT-R system, and to compare these with a human produced referent. The human judges rated both adequacy and fluency, and their judgements are compared to automatic evaluation measures in order to gain more insight into the automatic evaluation of paraphrasing.

4.1. Method

4.1.1. Participants

Participants were 76 students of Tilburg University, who participated for partial course credits. All were native speakers of Dutch, and all were proficient in English, having taken a course on Academic English at university level.

4.1.2. Materials

We randomly selected 1,000 headline clusters for Dutch and 1,000 headline clusters for English that appeared online in January 2011. Each cluster consisted of between 10 and 50 aligned paraphrasing headlines. We used these clusters as multiple references for our automatic evaluation measures to account for the diversity in real-world paraphrases, as the aligned paraphrased headlines in Figure **??** witness. For each participant we randomly selected 40 clusters, and from each cluster we randomly selected one headline as the source headline. Each headline was used as input for the word substitution baseline and the PBMT-R system, to generate two target paraphrases. In addition, we randomly selected one of the aligned headlines in a cluster to serve as the human produced upper bound to compare our systems with. For each source headline, we thus generated three target headlines (word substitution, PBMT-R, human-produced paraphrase). Each participant saw 40 different source headlines.

operation	sentences
single word replacement	50%
single word deletion or insertion	34%
word/phrase reordering	11%
phrase replacement	33%
sentence rewriting	2%

Table 1: Analysis of a sample of output from the English PBMT-R system indicating the number of sentences containing one or more of the specified edit operations.

4.1.3. Procedure

Participants were randomly assigned to the Dutch (N = 36) or English (N = 40) condition. In one version participants rated only Dutch target headlines, in the other they rated English ones. The instructions were otherwise identical for both versions. Participants were told that they participated in the evaluation of a system that could automatically generate headlines, and that they would see one source headline and three automatically generated paraphrases of that headline. Following earlier evaluation studies (Doddington, 2002; **?**), we asked participants to evaluate both the fluency and adequacy of the target headlines on a five point Likert scale. Fluency was defined in the instructions as the extent to which a sentence reads well. Adequacy was defined as the extent to which the sentence is a good paraphrase of the example sentence. Each source headline was presented on the computer screen, together with the three target headlines. The order of these targets on the screen was randomized, to prevent a bias towards one of the paraphrases. The experiment was individually performed, and self-paced; participants could take as much time as they required. On average the experiment lasted 33 minutes for English and 28 minutes for Dutch.

4.1.4. Data analysis

In total we collected 76 (participants) × 40 (clusters) × 3 (targets) = 9120 judgements. In practice, it turned out that the baseline system failed to generate a paraphrase in 12% of the cases for English and in 21% of the cases for Dutch. These could not be included in the analysis, so that the total number of collected judgements was lower. Since we are interested in the amount of edit operations a system performs and how these influence the evaluation, we computed the Levenshtein Distance (LD) from the source sentence of each target sentence at the word level ignoring casing. We created bins of LD 1, 2, 3, 4, and a collapsed bin of 5 or more to prevent data sparseness.

We performed two kinds of analyses. First we analyzed the human judgements in a by-item Multivariate Analysis of Variance (MANOVA) with Levenshtein Distance (levels: 1, 2, 3, 4, 5+), System (levels: word substitution, PBMT, human reference) and Language (levels: Dutch, English) as fixed factors and fluency and adequacy as dependent variables. Planned pairwise comparisons were made with the Bonferroni method.

Next, we evaluated the paraphrases using two automatic metrics, originating from the evaluation of machine translation: the BLEU (Papineni et al., 2002) and NIST (Dod-

[3] http://search.cpan.org/dist/WordNet-QueryData/QueryData.pm

dington, 2002) metrics. BLEU measures n-gram overlap between strings, and is expressed as a score between 0 and 1, with higher scores representing more overlap. Different scores are calculated for n-grams of different size, up to n-grams of four. NIST is a BLEU variant giving more importance to less frequent (and hence more informative) n-grams. For each of the target paraphrases used in the evaluation experiment we compute BLEU and NIST scores, which we submitted to a MANOVA with the same design as used for the human judgements. We used the remaining headlines for each cluster as the reference paraphrases for the automatic measures. In addition, we look at the correlations between the human judgements and the automatic metrics.

system	LD English	LD Dutch	fr En	fr Du
Word Sub	2.73	1.76	12%	21%
PBMT-R	2.57	2.88	0%	0%
Human	5.76	4.40	0%	0%

Table 2: Levenshtein distance and fail rate (fr) of output of the various systems

4.2. Results

Table 2 offers statistics showing the average LD of the target paraphrases in the cases where the system could find one, and the percentage of cases where the system was not able to generate a paraphrase of the source sentence. It can be observed that in general the PBMT-R system executes roughly equally many as the baseline for English, and more than the baseline for Dutch. Human produced paraphrases tend to differ more from the source. In addition, for 12 percent of the English sentences and 21 percent of the Dutch sentences the word substitution baseline could not provide a paraphrase. The PBMT-R system provided a paraphrase for every sentence.

4.2.1. Human judgements

Next, we analyzed the human judgements of fluency and adequacy of the target paraphrases. As expected, on both measures, the baseline word substitution system scored lowest (fluency: $M_f = 2.86$, adequacy: $M_a = 2.61$), and the randomly selected human reference scored highest (M_f = 4.18., M_a = 3.24) with the PBMT-R system sandwiched in between (M_f = 3.32, M_a = 2.86), showing a significant main effect for both fluency ($F(2, 7584) = 449.33, p < .001$) and adequacy ($F(2, 7584) = 95.48, p < .001$). All pairwise comparisons were statistically significant ($p < .001$). In addition, main effects were found for Language and Levenshtein Distance, but these are qualified by interactions with System. A significant interaction was found between Language and System, for both fluency ($F(2, 7584) = 147.93, p < .001$) and adequacy ($F(2, 7584) = 27.11, p < .001$). These interactions suggest that the effect of language is larger for the Baseline

than for the PBMT-R system, which might be due to the larger coverage of the English WordNet and the higher quality of the English language model. In addition, a significant interaction was found between Levenshtein Distance and System for fluency ($F(8, 7584) = 11.89, p < .001$), whereas the same interaction showed a trend towards significance for adequacy ($F(8, 7584) = 3.37, p = .071$). These effects are illustrated in Figure 1. First consider the results for fluency. It can be seen that fluency judgements of the human reference sentences do not vary with Levenshtein Distance, whereas the scores for the automatic systems show a steady decline as distance increases. Crucially, the performance of the PBMT-R system decreases less than the word substitution baseline beyond LD = 1. The picture for adequacy is slightly different: here all systems score lower as a function of LD, which is what one would expect given that the more distant a sentence is, the more likely it is that its content is also different. Crucially, however, while at LD = 1 the PBMT-R system scores roughly comparable to the baseline system, the two diverge more starting from LD = 2, and the PBMT-R system scores closer to the human reference than to the Baseline at LD = 5+.

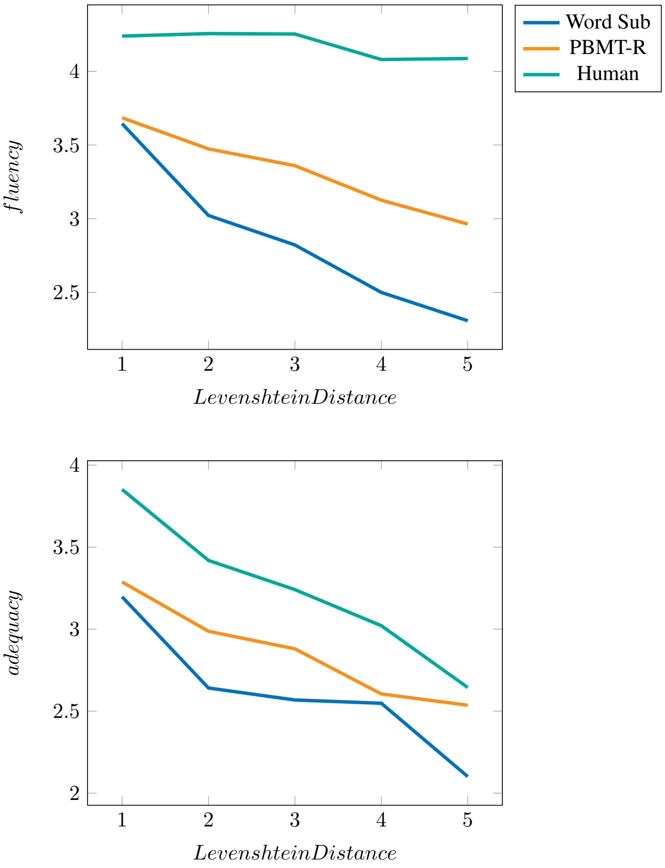

Figure 1: Fluency scores (top) and adequacy scores (bottom) per system as a function of Levenshtein Distance

4.2.2. Automatic measures

The results of the automatic evaluation metrics were analyzed next. We found that the baseline word substitution system attains the lowest scores (BLEU = 0.11, NIST = 7.00), and the randomly selected human reference scored highest (BLEU = 0.28, NIST = 8.19). We see that the PBMT-R system again scores between those two (BLEU = 0.18, NIST = 8.11), showing a significant effect for both BLEU ($F(2, 7584) = 200.91, p < .001$) and NIST ($F(2, 7584) = 105.54, p < .001$). In addition, main effects of Language and System are found, but these are again qualified by interactions.

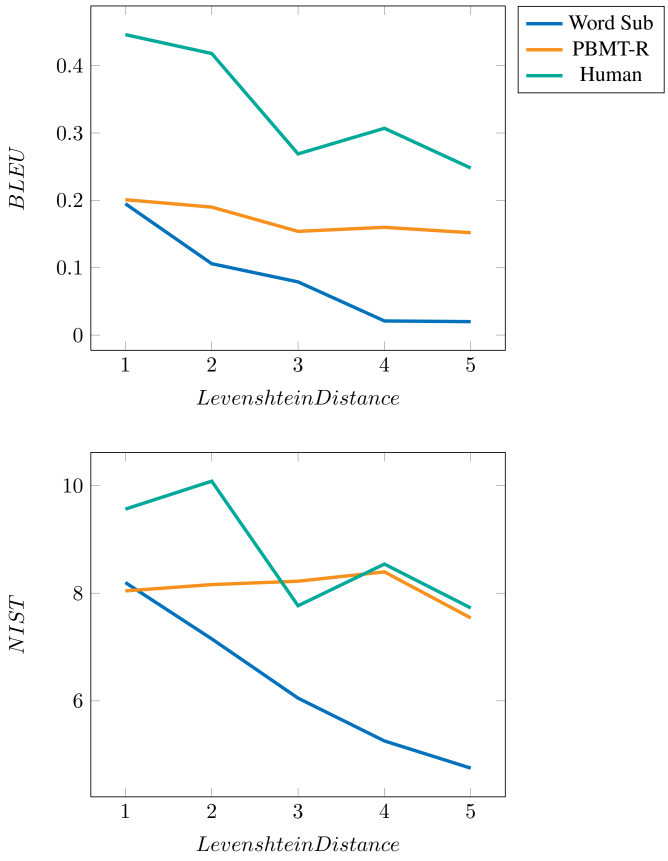

Figure 2: BLEU scores (top) and NIST scores (bottom) per system as a function of Levenshtein Distance

Significant interactions between Levenshtein Distance and System were found for both BLEU ($F(8, 7584) = 5.790, p < .001$) and NIST ($F(8, 7584) = 14.070, p < .001$). These interactions can be explained by looking at Figure 2: at LD = 1, the word substitution baseline and the PBMT system score roughly comparable and substantially lower than the human referent. However, when considering larger distances, the scores show a decreasing trend, but the scores for the PBMT-R system drop less than those of the word substitution baseline. At LD = 5 the PBMT-R system scores very comparable to the human baseline; this pattern is especially pronounced for the NIST scores. In addition, significant interactions were found of Language and System, for both BLEU ($F(4, 7584) = 3.781, p < .01$) and NIST ($F(4, 7584) = 4.329, p < .01$). This figure shows that, even though the PBMT-R system always scores higher than the word substitution system, the difference is more pronounced for English than for Dutch.

In general, it is fair to say that the results of the automatic evaluation mirror those of the human judgements. This is confirmed by a correlation analysis. We found a strong correlation between BLEU and NIST, as expected, but, more interestingly, we also found that both correlate significantly and positive with fluency ($r = .10$ for BLEU, and $r = .06$ for NIST, both $p < .001$) and adequacy ($r = .12$ for BLEU, and $r = .13$ for NIST, both $p < .001$).

Table 1 lists a breakdown of the paraphrasing operations the PBMT-R approach has performed. The number indicates the percentage of generated headlines out of a sample of 160 English generated headlines that contain one of the specified edit operations. Phrase replacements should be interpreted as a replacement involving multi-word phrases. Sentence rewriting means that the sentence is fundamentally changed in its entirety, for instance changing from passive to active and vice versa. We observe that even though the PBMT-R system is capable of manipulating multi-word phrases, the most frequent change is still single word replacement, and a majority of changes involve single word edits (replacements, insertions, or deletions). Yet, a substantial number of changes made by the PBMT-R system involve more complex phrasal manipulations and re-orderings.

5. Conclusion and discussion

In this paper we have presented a method to build a corpus of aligned sentential paraphrases. We used a standard PBMT framework with a dissimilarity component to generate the output paraphrases for two languages, English and Dutch, and compared this approach to a word substitution baseline.

In general, we found that the PBMT-R system outperforms the word substitution system on all dimensions of evaluation: it always succeeds in generating a paraphrase, while the baseline system fails to do so on 12% (English) to 21% (Dutch) of the source sentences. If we concentrate on the cases where the baseline system succeeds in generating a paraphrase, we find that the PBMT-R paraphrases are on average more dissimilar to the source sentences, as shown by their higher average Levenshtein distance. The human evaluators rated the output of the PBMT-R system higher than that of the baseline system, both in terms of adequacy and fluency. The human judgements show that while the performance of the baseline system drops substantially with higher Levenshtein distances, the PBMT-R system shows a less steep decline on both dimensions of evaluation. The automatic evaluation metrics (BLEU and NIST) reveal a similar pattern.

Human judges preferred the output of our PBMT-R system over the output of the word substitution system. However, it should be noted that the fluency of the PBMT-R system output is still very much below the fluency of human produced headlines. We have also addressed the problem of automatic paraphrase evaluation. We measured BLEU and NIST scores, and observed that these automatic scores correlate with human judgements to some degree. Overall they show the same picture: the selected human paraphrase scores best, followed by the PBMT-R system and the word substitution baseline comes in last. Because standard MT metrics such as BLEU and NIST do not take into account the notion of dissimilarity, these scores tend be high when few edits are made and drop as the paraphrases deviate more from the source sentence. When edit distance is considered, the decline of the scores of different systems can be compared.

We feel that our approach of using a corpus of scraped and aligned headlines together with an off-the-shelf PBMT package, modified to re-rank on dissimilarity (PBMT-R), despite the bias to the headline genre, is an important contribution in paraphrase research, as it allows the research to extend beyond English.

Our system, trained on the corpus of scraped and aligned headlines, may be usable in other domains and genres as well; it may be possible to train a language model on text from the new domain, and use the translation model acquired from the headlines to generate paraphrases for the new domain. We are also interested in capturing other monolingual text-to-text data, such as simplification or compression data, but acquiring monolingual parallel corpora for different domains is no trivial task.

6. References

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, 38:135–187, May.

Bannard, C. and Burch, C. C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Barzilay, R. and Lee, L. (2003). Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Callison-Burch, C., Cohn, T., and Lapata, M. (2008). Parametric: an automatic evaluation metric for paraphrasing. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, Morristown, NJ, USA. Association for Computational Linguistics.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 190–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing corpora for development and evaluation of paraphrase systems. *Computational Lingustics*, 34(4):597–614.

Daelemans, W., Zavrel, J., Berck, P., and Gillis, S. (1996). MBT: A memory-based part of speech tagger-generator. In *Proc. of Fourth Workshop on Very Large Corpora*, pages 14–27. ACL SIGDAT.

Daelemans, W., Hothker, A., and Tjong Kim Sang, E. (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Deléger, L., Merkel, M., and Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *J. of Biomedical Informatics*, 42:692–701, August.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 350, Morristown, NJ, USA. Association for Computational Linguistics.

Elhadad, N. and Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, May.

Kauchak, D. and Barzilay, R. (2006). Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA, June. Association for Computational Linguistics.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artif. Intell.*, 139(1):91–107.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Burch, C. C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL*. The Association for Computer Linguistics.

Kraaij, W. and Pohlmann, R. (1994). Porter's stemming algorithm for Dutch. In *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie*, pages 167–180.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lin, D. and Pantel, P. (2001). Dirt: Discovery of inference rules from text. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328, New York, NY, USA. ACM.

Liu, C., Dahlmeier, D., and Ng, H. T. (2010). Pem: a paraphrase evaluation metric exploiting parallel texts. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 923–932, Stroudsburg, PA, USA. Association for Computational Linguistics.

Madnani, N. and Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Marsi, E. and Krahmer, E. (2007). Annotating a parallel monolingual treebank with semantic similarity relations. In *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*, pages 85–96, Bergen, Norway.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore, August. Association for Computational Linguistics.

Nelken, R. and Shieber, S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy, 3–7 April.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pado, S., Galley, M., Jurafsky, D., and Manning, C. (2009). Robust machine translation evaluation with entailment features. In *Proceedings of ACL-IJCNLP*, pages 297–305.

Pang, Knight, and Marcu. (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *HLT-NAACL*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Quirk, C., Brockett, C., and Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V. O., and Liu, Y. (2007). Statistical machine translation for query expansion in answer retrieval. In *ACL*.

Shinyama, Y., Sekine, S., Sudo, K., and Grishman, R. (2002). Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference (HLT 2002)*, pages 313–318, San Diego, USA.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *In Proc. Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, Colorado.

Vossen, P., Maks, I., Segers, R., and VanderVliet, H. (2008). Integrating lexical units, synsets and ontology in the cornetto database. In Nicoletta Calzolari (Conference Chair), K. C., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Wubben, S., van den Bosch, A., and Krahmer, E. (2010). Paraphrase generation as monolingual translation: Data and evaluation. In J. Kelleher, B. M. N. and van der Sluis, I., editors, *Proceedings of the 10th International Workshop on Natural Language Generation (INLG 2010)*, pages 203–207, Dublin.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea, July. Association for Computational Linguistics.

Zhao, S., Lan, X., Liu, T., and Li, S. (2009). Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 834–842, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhou, L., Lin, C.-Y., and Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia, July. Association for Computational Linguistics.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.