

# Monolingual Alignment by Edit Rate Computation on Sentential Paraphrase Pairs

Houda Bouamor

Aurélien Max

Anne Vilnat

LIMSI-CNRS  
Univ. Paris Sud  
Orsay, France

{firstname.lastname}@limsi.fr

## Abstract

In this paper, we present a novel way of tackling the monolingual alignment problem on pairs of sentential paraphrases by means of edit rate computation. In order to inform the edit rate, information in the form of subsentential paraphrases is provided by a range of techniques built for different purposes. We show that the tunable TER-PLUS metric from Machine Translation evaluation can achieve good performance on this task and that it can effectively exploit information coming from complementary sources.

## 1 Introduction

The acquisition of subsentential paraphrases has attracted a lot of attention recently (Madnani and Dorr, 2010). Techniques are usually developed for extracting paraphrase candidates from specific types of corpora, including monolingual parallel corpora (Barzilay and McKeown, 2001), monolingual comparable corpora (Deléger and Zweigenbaum, 2009), bilingual parallel corpora (Bannard and Callison-Burch, 2005), and edit histories of multi-authored text (Max and Wisniewski, 2010). These approaches face two main issues, which correspond to the typical measures of *precision*, or how appropriate the extracted paraphrases are, and of *recall*, or how many of the paraphrases present in a given corpus can be found effectively. To start with, both measures are often hard to compute in practice, as 1) the definition of what makes an acceptable paraphrase pair is still a research question, and 2) it is often impractical to extract a complete set of acceptable paraphrases

from most resources. Second, as regards the precision of paraphrase acquisition techniques in particular, it is notable that most works on paraphrase acquisition are not based on *direct observation* of larger paraphrase pairs. Even monolingual corpora obtained by pairing very closely related texts such as news headlines on the same topic and from the same time frame (Dolan et al., 2004) often contain unrelated segments that should not be aligned to form a subsentential paraphrase pair. Using bilingual corpora to acquire paraphrases indirectly by pivoting through other languages is faced, in particular, with the issue of phrase polysemy, both in the source and in the pivot languages.

It has previously been noted that highly parallel monolingual corpora, typically obtained via multiple translation into the same language, constitute the most appropriate type of corpus for extracting high quality paraphrases, in spite of their rareness (Barzilay and McKeown, 2001; Cohn et al., 2008; Bouamor et al., 2010). We build on this claim here to propose an original approach for the task of subsentential alignment based on the computation of a minimum edit rate between two sentential paraphrases. More precisely, we concentrate on the alignment of *atomic paraphrase pairs* (Cohn et al., 2008), where the words from both paraphrases are aligned as a whole to the words of the other paraphrase, as opposed to *composite paraphrase pairs* obtained by joining together adjacent paraphrase pairs or possibly adding unaligned words. Figure 1 provides examples of atomic paraphrase pairs derived from a word alignment between two English sentential paraphrases.

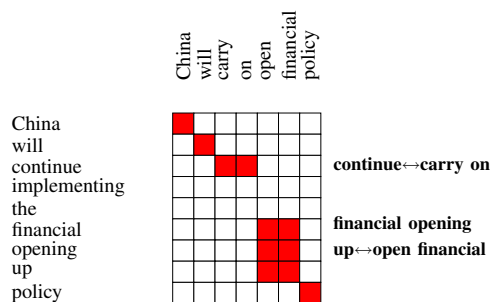


Figure 1: Reference alignments for a pair of English sentential paraphrases and their associated list of atomic paraphrase pairs extracted from them. Note that *identity pairs* (e.g. *China* ↔ *China*) will never be considered in this work and will not be taken into account for evaluation.

The remainder of this paper is organized as follows. We first briefly describe in section 2 how we apply edit rate computation to the task of atomic paraphrase alignment, and we explain in section 3 how we can inform such a technique with paraphrase candidates extracted by additional techniques. We present our experiments and discuss their results in section 4 and conclude in section 5.

## 2 Edit rate for paraphrase alignment

TER-PLUS (Translation Edit Rate Plus) (Snover et al., 2010) is a score designed for evaluation of Machine Translation (MT) output. Its typical use takes a system hypothesis to compute an optimal set of word edits that can transform it into some existing reference translation. Edit types include exact word matching, word insertion and deletion, block movement of contiguous words (computed as an approximation), as well as variants substitution through stemming, synonym or paraphrase matching. Each edit type is parameterized by at least one weight which can be optimized using e.g. hill climbing. TER-PLUS is therefore a tunable metric. We will henceforth design as  $TER_{MT}$  the TER metric (basically, without variants matching) optimized for correlation with human judgment of accuracy in MT evaluation, which is to date one of the most used metrics for this task.

While this metric was not designed explicitly for the acquisition of word alignments, it produces as a by-product of its approximate search a list of alignments involving either individual words or phrases, potentially fitting with the previous definition of atomic paraphrase pairs. When applying it on a MT system hypothesis and a reference translation, it computes how much effort would be needed to obtain the reference from the hypothesis, possibly independently of the appropriateness of the alignments produced. However, if we consider instead a pair of sentential paraphrases, it can be used to reveal what subsentential units can be aligned. Of course, this relies on information that will often go beyond simple exact word matching. This is where the capability of exploiting paraphrase matching can come into play: TER-PLUS can exploit a table of paraphrase pairs, and defines the cost of a phrase substitution as “a function of the probability of the paraphrase and the number of edits needed to align the two phrases without the use of phrase substitutions”. Intuitively, the more parallel two sentential paraphrases are, the more atomic paraphrase pairs will be reliably found, and the easier it will be for TER-PLUS to correctly identify the remaining pairs. But in the general case, and considering less apparently parallel sentence pairs, its work can be facilitated by the incorporation of candidate paraphrase pairs in its paraphrase table. We consider this possible type of hybridation in the next section.

## 3 Informing edit rate computation with other techniques

In this article, we use three baseline techniques for paraphrase pair acquisition, which we will only briefly introduce (see (Bouamor et al., 2010) for more details). As explained previously, we want to evaluate whether and how their candidate paraphrase pairs can be used to improve paraphrase acquisition on sentential paraphrases using TER-PLUS. We selected these three techniques for the complementarity of types of information that they use: statistical word alignment without *a priori* linguistic knowledge, symbolic expression of linguistic variation exploiting *a priori* linguistic knowledge, and syntactic similarity.

**Statistical Word Alignment** The GIZA++ tool (Och and Ney, 2004) computes statistical word alignment models of increasing complexity from parallel corpora. While originally developed in the bilingual context of Machine Translation, nothing prevents building such models on monolingual corpora. However, in order to build reliable models it is necessary to use enough training material including minimal redundancy of words. To this end, we will be using monolingual corpora made up of multiply-translated sentences, allowing us to provide GIZA++ with all possible sentence pairs to improve the quality of its word alignments (note that following common practice we used symmetrized alignments from the alignments in both directions). This constitutes an advantage for this technique that the following techniques working on each sentence pair independently do not have.

**Symbolic expression of linguistic variation** The FASTR tool (Jacquemin, 1999) was designed to spot term variants in large corpora. Variants are described through metarules expressing how the morphosyntactic structure of a term variant can be derived from a given term by means of regular expressions on word categories. Paradigmatic variation can also be expressed by defining constraints between words to force them to belong to the same morphological or semantic family, both constraints relying on preexisting repertoires available for English and French. To compute candidate paraphrase pairs using FASTR, we first consider all the phrases from the first sentence and search for variants in the other sentence, do the reverse process and take the intersection of the two sets.

**Syntactic similarity** The algorithm introduced by Pang et al. (2003) takes two sentences as input and merges them by top-down syntactic fusion guided by compatible syntactic substructure. A lexical blocking mechanism prevents sentence constituents from fusioning when there is evidence of the presence of a word in another constituent of one of the sentence. We use the Berkeley Probabilistic parser (Petrov and Klein, 2007) to obtain syntactic trees for English and its Bonsai adaptation for French (Candito et al., 2010). Because this process is highly sensitive to syntactic parse errors, we use  $k$ -best parses (with  $k = 3$  in our experiments) and

retain the most compact fusion from any pair of candidate parses.

## 4 Experiments and discussion

We used the methodology described by Cohn et al. (2008) for constructing evaluation corpora and assessing the performance of various techniques on the task of paraphrase acquisition. In a nutshell, pairs of sentential paraphrases are hand-aligned and define a set of reference atomic paraphrase pairs at the level of words or blocks or words, denoted as  $\mathcal{R}_{\text{atom}}$ , and also a set of reference *composite* paraphrase pairs obtained by joining adjacent atomic paraphrase pairs (up to a given length), denoted as  $\mathcal{R}$ . Techniques output word alignments from which atomic candidate paraphrase pairs, denoted as  $\mathcal{H}_{\text{atom}}$ , as well as composite paraphrase pairs, denoted as  $\mathcal{H}$ , can be extracted. The usual measures of *precision*, *recall* and *f-measure* can then be defined in the following way:

$$p = \frac{|\mathcal{H}_{\text{atom}} \cap \mathcal{R}|}{|\mathcal{H}_{\text{atom}}|} \quad r = \frac{|\mathcal{H} \cap \mathcal{R}_{\text{atom}}|}{|\mathcal{R}_{\text{atom}}|} \quad f_1 = \frac{2pr}{p+r}$$

To evaluate our individual techniques and their use by the tunable TER-PLUS technique (henceforth TERP), we measured results on two different corpora in French and English. In each case, a held-out development corpus of 150 paraphrase pairs was used for tuning the TERP hybrid systems towards precision ( $\rightarrow p$ ), recall ( $\rightarrow r$ ), or F-measure ( $\rightarrow f_1$ ).<sup>1</sup> All techniques were evaluated on the same test set consisting of 375 paraphrase pairs. For English, we used the MTC corpus described in (Cohn et al., 2008), which consists of multiply-translated Chinese sentences into English, with an average lexical overlap<sup>2</sup> of 65.91% (all tokens) and 63.95% (content words only). We used as our reference set both the alignments marked as “Sure” and “Possible”. For French, we used the CESTA corpus of news articles<sup>3</sup> obtained by translating into French from various languages with an average lexical overlap of 79.63% (all tokens) and 78.19% (content words only). These

<sup>1</sup>*Hill climbing* was used for tuning as in (Snover et al., 2010), with uniform weights and 100 random restarts.

<sup>2</sup>We compute the percentage of lexical overlap between the vocabularies of two sentences  $S_1$  and  $S_2$  as :  $|S_1 \cap S_2| / \min(|S_1|, |S_2|)$

<sup>3</sup><http://www.elda.org/article125.html>

|                       | Individual techniques |          |              |                 |                      |            |                         | Hybrid systems (TERP <sub>para+x</sub> ) |              |                         |              |              |                         |              |            |                         |            |            |                         |
|-----------------------|-----------------------|----------|--------------|-----------------|----------------------|------------|-------------------------|--|--------------|-------------------------|--------------|--------------|-------------------------|--------------|------------|-------------------------|------------|------------|-------------------------|
|                       | Giza++                | Fastr    | Pang         | T <sub>MT</sub> | TERP <sub>para</sub> |            |                         | +G                                       |              |                         | +F           |              |                         | +P           |            |                         | +G + F + P |            |                         |
|                       | <i>G</i>              | <i>F</i> | <i>P</i>     |                 | → <i>p</i>           | → <i>r</i> | → <i>f</i> <sub>1</sub> | → <i>p</i>                               | → <i>r</i>   | → <i>f</i> <sub>1</sub> | → <i>p</i>   | → <i>r</i>   | → <i>f</i> <sub>1</sub> | → <i>p</i>   | → <i>r</i> | → <i>f</i> <sub>1</sub> | → <i>p</i> | → <i>r</i> | → <i>f</i> <sub>1</sub> |
|                       | <b>French</b>         |          |              |                 |                      |            |                         | <b>French</b>                            |              |                         |              |              |                         |              |            |                         |            |            |                         |
| <i>p</i>              | 28.99                 | 52.48    | <b>62.50</b> | 25.66           | 31.35                | 30.26      | 31.43                   | 41.99                                    | 30.55        | 41.14                   | 36.74        | 29.65        | 34.84                   | <b>54.49</b> | 20.94      | 33.89                   | 42.27      | 27.06      | 42.80                   |
| <i>r</i>              | <b>45.98</b>          | 8.59     | 8.65         | 41.15           | 44.22                | 44.60      | 44.10                   | 35.88                                    | <b>45.67</b> | 35.25                   | 40.96        | 43.85        | 44.41                   | 13.61        | 40.40      | 40.46                   | 31.36      | 44.10      | 31.61                   |
| <i>f</i> <sub>1</sub> | 35.56                 | 14.77    | 15.20        | 25.66           | 36.69                | 36.05      | <b>36.70</b>            | 38.70                                    | 36.61        | 37.97                   | 38.74        | 35.38        | <b>39.05</b>            | 21.78        | 27.58      | 36.88                   | 36.01      | 33.54      | 36.37                   |
|                       | <b>English</b>        |          |              |                 |                      |            |                         | <b>English</b>                           |              |                         |              |              |                         |              |            |                         |            |            |                         |
| <i>p</i>              | 18.28                 | 33.02    | <b>36.66</b> | 20.41           | 31.19                | 19.14      | 19.35                   | 26.89                                    | 19.85        | 21.25                   | <b>41.57</b> | 20.81        | 22.51                   | 31.32        | 18.02      | 18.92                   | 29.45      | 16.81      | 29.42                   |
| <i>r</i>              | 14.63                 | 5.41     | 2.23         | 17.37           | 2.31                 | 19.38      | <b>19.69</b>            | 11.92                                    | 18.47        | 17.10                   | 6.94         | <b>21.02</b> | 20.28                   | 3.41         | 18.94      | 16.44                   | 13.57      | 19.30      | 16.35                   |
| <i>f</i> <sub>1</sub> | 16.25                 | 9.30     | 4.21         | 18.77           | 4.31                 | 19.26      | <b>19.52</b>            | 16.52                                    | 19.14        | 18.95                   | 11.91        | 20.92        | <b>21.33</b>            | 6.15         | 18.47      | 17.59                   | 18.58      | 17.96      | 21.02                   |

Figure 2: Results on the test set on French and English for the individual techniques and TERP hybrid systems. Column headers of the form “→ *c*” indicate that TERP was tuned on criterion *c*.

figures reveal that the French corpus tends to contain more literal translations, possibly due to the original languages of the sentences, which are closer to the target language than Chinese is to English. We used the YAWAT (Germann, 2008) interactive alignment tool and measure inter-annotator agreement over a subset and found it to be similar to the value reported by Cohn et al. (2008) for English.

Results for all individual techniques in the two languages are given on Figure 2. We first note that all techniques fared better on the French corpus than on the English corpus. This can certainly be explained by the fact that the former results from more literal translations, which are consequently easier to word-align.

TER<sub>MT</sub> (i.e. TER tuned for Machine Translation evaluation) performs significantly worse on all metrics for both languages than our tuned TERP experiments, revealing that the two tasks have different objectives. The two linguistically-aware techniques, FASTR and PANG, have a very strong precision on the more parallel French corpus, and also on the English corpus to a lesser extent, but fail to achieve a high recall (note, in particular, that they do not attempt to report preferentially *atomic* paraphrase pairs). GIZA++ and TERP<sub>para</sub> perform in the same range, with acceptable precision and recall, TERP<sub>para</sub> performing overall better, with e.g. a 1.14 advantage on f-measure on French and 3.27 on English. Recall that TERP works independently on each paraphrase pair, while GIZA++ makes use of

artificial repetitions of paraphrases of the same sentence.

Figure 3 gives an indication of how well each technique performs depending on the difficulty of the task, which we estimate here as the value  $(1 - \text{TER}(\text{para}_1, \text{para}_2))$ , whose low values correspond to sentences which are costly to transform into the other using TER. Not surprisingly, TERP<sub>para</sub> and GIZA++, and PANG to a lesser extent, perform better on “more parallel” sentential paraphrase pairs. Conversely, FASTR is not affected by the degree of parallelism between sentences, and manages to extract synonyms and more generally term variants, at any level of difficulty.

We have further tested 4 hybrid configurations by providing TERP<sub>para</sub> with the output of the other individual techniques and of their union, the latter simply obtained by taking paraphrase pairs output by at least one of these techniques. On French, where individual techniques achieve good performance, any hybridation improves the F-measure over both TERP<sub>para</sub> and the technique used, the best performance, using FASTR, corresponding to an improvement of respectively +2.35 and +24.28 over TERP<sub>para</sub> and FASTR. Taking the union of all techniques does not yield additional gains: this might be explained by the fact that incorrect predictions are proportionally more present and consequently have a greater impact when combining techniques without weighting them, possibly at the level of each

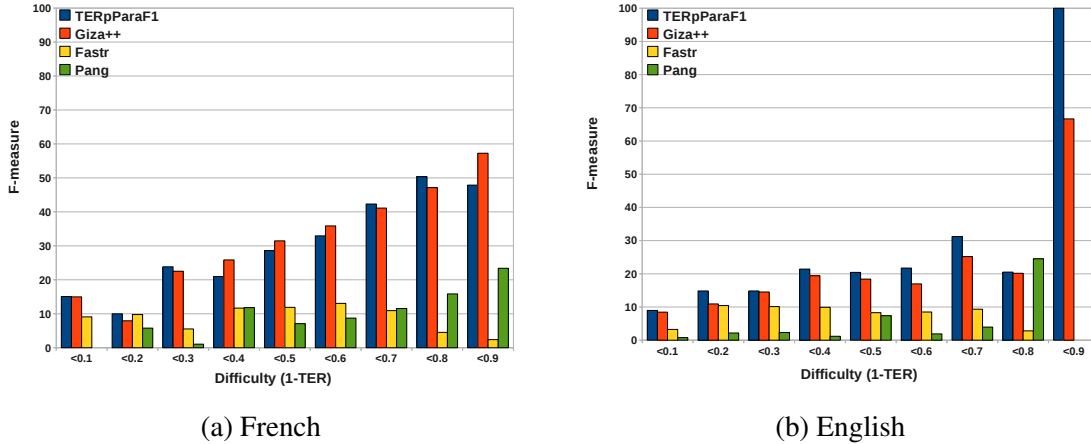


Figure 3: F-measure values for our 4 individual techniques on French and English depending on the complexity of paraphrase pairs measured with the (1-TER) formula. Note that each value corresponds to the average of F-measure values for test examples falling in a given difficulty range, and that all ranges do not necessarily contain the same number of examples.

prediction.<sup>4</sup> Successful hybridation on English seem harder to obtain, which may be partly attributed to the poor quality of the individual techniques relative to  $TER_{para}$ . We however note anew an improvement over  $TER_{para}$  of +1.81 when using  $F_{ASTR}$ . This confirms that some types of linguistic equivalences cannot be captured using edit rate computation alone, even on this type of corpus.

## 5 Conclusion and future work

In this article, we have described the use of edit rate computation for paraphrase alignment at the sub-sentential level from sentential paraphrases and the possibility of informing this search with paraphrase candidates coming from other techniques. Our experiments have shown that in some circumstances some techniques have a good complementarity and manage to improve results significantly. We are currently studying *hard-to-align* sub-sentential paraphrases from the type of corpora we used in order to get a better understanding of the types of knowledge required to improve automatic acquisition of these units.

<sup>4</sup>Indeed, measuring the precision on the union yields a poor performance of 23.96, but with the highest achievable value of 50.56 for recall. Similarly, the maximum value for precision with a good recall can be obtained by taking the intersection of the results of  $TER_{para}$  and  $GIZA++$ , which yields a value of 60.39.

Our future work also includes the acquisition of paraphrase patterns (e.g. (Zhao et al., 2008)) to generalize the acquired equivalence units to more contexts, which could be both used in applications and to attempt improving further paraphrase acquisition techniques. Integrating the use of patterns within an edit rate computation technique will however raise new difficulties.

We are finally also in the process of conducting a careful study of the characteristics of the paraphrase pairs that each technique can extract with high confidence, so that we can improve our hybridation experiments by considering confidence values at the paraphrase level using Machine Learning. This way, we may be able to use an edit rate computation algorithm such as  $TER-PLUS$  as a more efficient system combiner for paraphrase extraction methods than what was proposed here. A potential application of this would be an alternative proposal to the paraphrase evaluation metric  $PARAMETRIC$  (Callison-Burch et al., 2008), where individual techniques, outputting word alignments or not, could be evaluated from the ability of the informed edit rate technique to use correct equivalence units.

## Acknowledgments

This work was partly funded by a grant from LIMSI. The authors wish to thank the anonymous reviewers for their useful comments and suggestions.

## References

- Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL*, Ann Arbor, USA.
- Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*, Toulouse, France.
- Houda Bouamor, Aurélien Max, and Anne Vilnat. 2010. Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. In *Proceedings of IceTAL*, Reykjavik, Iceland.
- Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of COLING*, Manchester, UK.
- Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of LREC*, Valletta, Malta.
- Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4).
- Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, Singapore.
- Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004*, pages 350–356, Geneva, Switzerland.
- Ulrich Germann. 2008. Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, Columbus, USA.
- Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL*, pages 341–348, College Park, USA.
- Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3).
- Aurélien Max and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia’s Revision History. In *Proceedings of LREC*, Valletta, Malta.
- Franz Josef Och and Herman Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).
- Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL-HLT*, Edmonton, Canada.
- Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL-HLT*, Rochester, USA.
- Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2010. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).
- Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proceedings of ACL-HLT*, Columbus, USA.