Web-based validation for contextual targeted paraphrasing

Houda Bouamor
LIMSI-CNRS
Univ. Paris Sud
hbouamor@limsi.fr

Aurélien Max
LIMSI-CNRS
Univ. Paris Sud
amax@limsi.fr

Gabriel Illouz
LIMSI-CNRS
Univ. Paris Sud
gabrieli@limsi.fr

Anne Vilnat
LIMSI-CNRS
Univ. Paris Sud
anne@limsi.fr

Abstract

In this work, we present a scenario where contextual targeted paraphrasing of sub-sentential phrases is performed automatically to support the task of text revision. Candidate paraphrases are obtained from a preexisting repertoire and validated in the context of the original sentence using information derived from the Web. We report on experiments on French, where the original sentences to be rewritten are taken from a rewriting memory automatically extracted from the edit history of Wikipedia.

1 Introduction

There are many instances where it is reasonable to expect machines to produce text automatically. Traditionally, this was tackled as a concept-to-text realization problem. However, such needs apply sometimes to cases where a new text should be derived from some existing texts, an instance of text-to-text generation. The general idea is not anymore to produce a text from data, but to transform a text so as to ensure that it has desirable properties appropriate for some intended application (Zhao et al., 2009). For example, one may want a text to be shorter (Cohn and Lapata, 2008), tailored to some reader profile (Zhu et al., 2010), compliant with some specific norms (Max, 2004), or more adapted for subsequent machine processing tasks (Chandrasekar et al., 1996). The generation process must produce a text having a meaning which is compatible with the definition of the task at hand (e.g. strict paraphrasing for document normalization, relaxed paraphrasing for text simplification), while ensuring that it remains grammatically correct. Its complexity, compared with concept-to-text generation, mostly stems from the fact that the semantic relationship between the original text and the new one is more difficult to control, as the mapping from one text to another is very dependent on the rewriting context. The wide variety of techniques for acquiring phrasal paraphrases, which can subsequently be used by text paraphrasing techniques (Madnani and Dorr, 2010), the inherent polysemy of such linguistic units and the pragmatic constraints on their uses make it impossible to ensure that potential paraphrase pairs will be substitutable in any context, an observation which was already made at a lexical level (Zhao et al., 2007). Hence, automatic contextual validation of candidate rewritings is a fundamental issue for text paraphrasing with phrasal units.

In this article, we tackle the problem of what we call *targeted paraphrasing*, defined as the rewriting of a subpart of a sentence, as in e.g. (Resnik et al., 2010) where it is applied to making parts of sentences easier to translate automatically. While this problem is simpler than full sentence rewriting, its study is justified as it should be handled correctly for the more complex task to be successful. Moreover, being simpler, it offers evaluation scenarios which make the performance on the task easier to assess. Our particular experiments here aim to assist a Wikipedia contributor in revising a text to improve its quality. For this, we use a collection of phrases that have been rewritten in Wikipedia, and test the substitutability of paraphrases coming from a repertoire of sub-sentential paraphrases acquired

Workshop on Monolingual Text-To-Text Generation, pages 10–19,

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 10–19,
Portland, Oregon, 24 June 2011.

from different sources. We thus consider that preexisting repertoires of sub-sentential paraphrase pairs are available, and that each potential candidate has to be tested in the specific context of the desired rewriting. Due to the large variety of potential phrases and their associated known paraphrases, we do not rely on precomputed models of substitutability, but rather build them on-the-fly using information derived from web queries.[1]

This article is organized as follows. In section 2, we first describe the task of text revision, where a subpart of a sentence is rewritten, as an instance of targeted paraphrasing. Section 3 presents previous works on the acquisition of sub-sentential paraphrases and describes the knowledge sources that we have used in this work. We then describe in section 4 how we estimate models of phrase substitution in context by exploiting information coming from the web. We present our experiments and their results in section 5, and finally discuss our current results and future work in section 6.

2 Targeted paraphrasing for text revision

One of the important processes of text revision is the rewording of parts of sentences. Some rewordings are not intended to alter meaning significantly, but rather to make text more coherent and easier to comprehend. Those instances which express close meanings are sub-sentential paraphrases: in their simpler form, they can involve synonym substitution, but they can involve more complex deeper lexical-syntactic transformations.

Such rephrasings are commonly found in recordings of text revisions, which now exist in large quantities in the collaborative editing model of Wikipedia. In fact, revision histories of the encyclopedia contain a significant amount of sub-sentential paraphrases, as shown by the study of (Dutrey et al., 2011). This study also reports that there is an important variety of rephrasing phenomena, as illustrated by the difficulty of reaching a good identification coverage using a rule-based term variant identification engine.

The use of automatic targeted paraphrasing as an authoring aid has been illustrated by the work of Max and Zock (2008), in which writers are presented with potential paraphrases of sub-sentential fragments that they wish to reword. The automatic paraphrasing technique used is a contextual variant of bilingual translation pivoting (Bannard and Callison-Burch, 2005). It has also been proposed to externalize various text editing tasks, including proofreading, by having crowdsourcing functions on text directly from word processors (Bernstein et al., 2010).

Text improvements may also be more specifically targeted for automatic applications. In the work by Resnik *et al.* (2010), rephrasings for specific phrases are acquired through crowdsourcing. Difficult-to-translate phrases in the source text are first identified, and monolingual contributors are asked to provide rephrasings in context. Collected rephrasings can then be used as input for a Machine Translation system, which can positively exploit the increased variety in expression to produce more confident translations for better estimated source units (Schroeder et al., 2009).[2] For instance, the phrase in bold in the sentence *The number of people* ***known to have died*** *has now reached 358* can be rewritten as 1) *who died*, 2) *identified to have died* and 3) *known to have passed away*. All such rephrasings are grammatically correct, the first one being significantly shorter, and they all convey a meaning which is reasonably close to the original wording.

The task of rewriting complete sentences has also been addressed in various works (e.g. (Barzilay and Lee, 2003; Quirk et al., 2004; Zhao et al., 2010)). It poses, however, numerous other challenges, in particular regarding how it could be correctly evaluated. Human judgments of whole sentence transformations are complex and intra- and inter-judge coherence is difficult to attain with hypotheses of comparable quality. Using sentential paraphrases to support a given task (e.g. providing alternative reference translations for optimizing Statistical Machine Translation systems (Madnani et al., 2008))

[1] Note that using the web may not always be appropriate, or that at least it should be used in a different way than what we propose in this article, in particular in cases where the desired properties of the rewritten text are better described in controlled corpora.

[2] It is to be noted that, in the scenario presented in (Resnik et al., 2010), monolingual contributors cannot predict how useful their rewritings will be to the underlying Machine Translation engine used.

can be seen as a proxy for extrinsic evaluation of the quality of paraphrases, but it is not clear from published results that improvements on the task are clearly correlated with the quality of the produced paraphrases. Lastly, automatic metrics have been proposed for evaluating the grammaticality of sentences (e.g. (Mutton et al., 2007)). Automatic evaluation of sentential paraphrases has not produced any consensual results so far, as they do not integrate task-specific considerations and can be strongly biased towards some paraphrasing techniques.

In this work, we tackle the comparatively more modest task of sub-sentential paraphrasing applied to text revision. In order to use an unbiased task, we use a corpus of naturally-occurring rewritings from an authoring memory of Wikipedia articles. We use the WiCoPaCo corpus (Max and Wisniewski, 2010), a collection of local rephrasings from the edit history of Wikipedia which contains instances of lexical, syntactical and semantic rephrasings (Dutrey et al., 2011), the latter type being illustrated by the following example:

Ce vers de Nuit rhénane d'Apollinaire ***[qui paraît presque sans structure rythmique → dont la césure est comme masquée]****...*[3]

The appropriateness of this corpus for our work is twofold: first, the fact that it contains naturally-occurring rewritings provides us with an interesting source of text spans in context which have been rewritten. Moreover, for those instances where the meaning after rewriting was not significantly altered, it provides us with at least one candidate rewriting that should be considered as a correct paraphrase, which can be useful for training validation algorithms.

3 Automatic sub-sentential paraphrase acquisition and generation

The acquisition of paraphrases, and in particular of sub-sentential paraphrases and paraphrase patterns, has attracted a lot of works with the advent of data-intensive Natural Language Processing (Madnani and Dorr, 2010). The techniques proposed have a strong relationship to the type of text corpus used for acquisition, mainly:

- pairs of sentential paraphrases (**monolingual parallel corpora**) allow for a good precision but evidently a low recall (e.g. (Barzilay and McKeown, 2001; Pang et al., 2003; Cohn et al., 2008; Bouamor et al., 2011))
- pairs of bilingual sentences (**bilingual parallel corpora**) allow for a comparatively better recall (e.g. (Bannard and Callison-Burch, 2005; Kok and Brockett, 2010))
- pairs of related sentences (**monolingual comparable corpora**) allow for even higher recall but possibly lower precision (e.g. (Barzilay and Lee, 2003; Li et al., 2005; Bhagat and Ravichandran, 2008; Deléger and Zweigenbaum, 2009)

Although the precision of such techniques can in some cases be formulated with regards to a predefined reference set (Cohn et al., 2008), it should more generally be assessed in the specific context of some use of the paraphrase pair. This refers to the problem of *substituability in context* (e.g. (Connor and Roth, 2007; Zhao et al., 2007)), which is a well studied field at the lexical level and the object of evaluation campains (McCarthy and Navigli, 2009). Contextual phrase substitution poses the additional challenge that phrases are rarer than words, so that building contextual and grammatical models to ensure that the generated rephrasings are both semantically compatible and grammatical is more complicated (e.g. (Callison-Burch, 2008)).

The present work does not aim to present any original technique for paraphrase acquisition, but rather focusses on the task of sub-sentential paraphrase validation in context. We thus resort to some existing repertoire of phrasal paraphrase pairs. As explained in section 2, we use the WiCoPaCo corpus as a source of sub-sentential paraphrases: the phrase after rewriting can thus be used as a potential paraphrase in context.[4] To obtain other candidates of various quality, we used two knowledge sources. The first uses automatic pivot translation (Bannard and Callison-Burch, 2005), where a state-of-the-art

[3] *This verse from Apollinaire's Nuit Rhénane* ***[which seems almost without rhythmic structure → whose cesura is as if hidden]****...*

[4] Note, however, that in our experiments we will ask our human judges to assess anew its paraphrasing status in context.

general-purpose Statistical Machine Translation system is used in a two-way translation. The second uses manual acquisition of paraphrase candidates. Web-based acquisition of this type of knowledge has already been done before (Chklovski, 2005; España Bonet et al., 2009), and could be done by crowdsourcing, a technique growing in popularity in recent years. We have instead formulated manual acquisition as a web-based game. Players can take parts in two parts of the game, illustrated on Figure 3.

First, players propose sub-sentential paraphrases in context for selected text spans in web documents (top of Figure 3), and then raters can take part in assessing paraphrases proposed by other players (bottom of Figure 3). In order to avoid any bias, players cannot evaluate games in which they played. Evaluation is sped up by using a compact word lattice view for eliciting human judgments, built using the syntactic fusion algorithm of (Pang et al., 2003). Data acquisition was done in French to remain coherent with our experiments on the French corpus of WiCoPaCo, and both players and raters were native speakers. An important point is that in our experiments the context of acquisition and of evaluation were different: players were asked to generate paraphrases in contexts that are different from those of the WiCoPaCo corpus used for evaluation. To this end, web snippets were automatically retrieved for the various phrases of our dataset without contexts, so that sentences from the Web (but not from Wikipedia) were used for manual paraphrase acquisition. This allows us to simulate the availability of a preexisting repertoire of (contextless) sub-sentential paraphrases, and to assess the performance of our contextual validation techniques on a possibly incompatible context.

4 Web-based contextual validation

Given a repertoire of potential phrasal paraphrases and a context for a naturally-occurring rewriting, our task consists in deciding automatically which potential paraphrases can be substituted with good confidence for the original phrase. A concrete instantiation of it could correspond to the proposal of Max and Zock (2008), where such candidate rephrasings could be presented in order of decreasing suitability to a word processor user, possibly during the revision of a Wikipedia article.

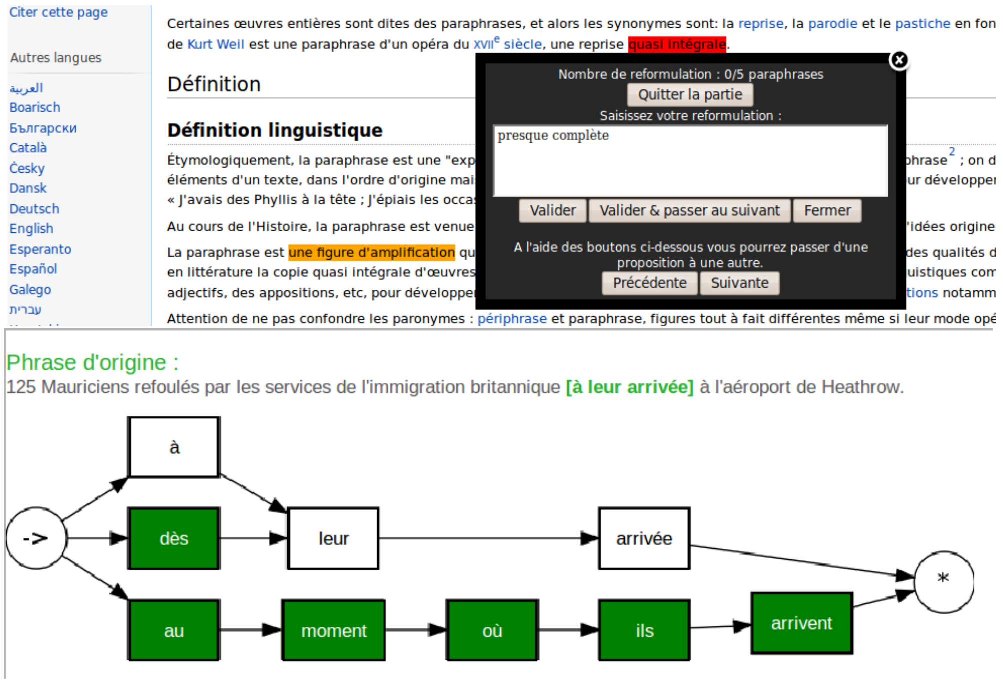
The specific nature of the text units that we are dealing with calls for a careful treatment: in the general scenario, it is unlikely that any supervised corpus would contain enough information for appropriate modeling of the substituability in context decision. It is therefore tempting to consider using the Web as the largest available information source, in spite of several of its known limitations, including that data can be of varying quality. It has however been shown that a large range of NLP applications can be improved by exploiting *n*-gram counts from the Web (using Web document counts as a proxy) (Lapata and Keller, 2005).

Paraphrase identification has been addressed previously, both using features computed from an offline corpus (Brockett and Dolan, 2005) and features computed from Web queries (Zhao et al., 2007). However, to our knowledge previous work exploiting information from the Web was limited to the identification of lexical paraphrases. Although the probability of finding phrase occurrences significantly increases by considering the Web, some phrases are still very rare or not present in search engine indexes.

As in (Brockett and Dolan, 2005), we tackle our paraphrase identification task as one of monolingual classification. More precisely, considering an original phrase *p* within the context of sentence *s*, we seek to determine whether a candidate paraphrase *p'* would be a grammatical paraphrase of *p* within the context of *s*. We make use of a Support Vector Machine (SVM) classifier which exploits the features described in the remainder of this section.

Edit distance model score Surface similarity on phrase pairs can be a good indicator that they share semantic content. In order to account for the cost of transforming one string into the other, rather than simply counting common words, we use the score produced by the Translation Edit Rate metric (Snover et al., 2010). Furthermore, we perform this computation on strings of lemmas rather than surface forms:[5]

[5] Note that because we computed the TER metric on French strings, stemming and semantic matching through WordNet were not activated.



Citer cette page

Autres langues

العربية
Boarisch
Български
Català
Česky
Dansk
Deutsch
English
Esperanto
Español
Galego
עברית

Certaines œuvres entières sont dites des paraphrases, et alors les synonymes sont: la reprise, la parodie et le pastiche en fon
de Kurt Weil est une paraphrase d'un opéra du XVII^{e} siècle, une reprise quasi intégrale.

Définition

Définition linguistique

Étymologiquement, la paraphrase est une "exp … phrase[2] ; on d
éléments d'un texte, dans l'ordre d'origine mai … ur développer
« J'avais des Phyllis à la tête ; J'épiais les occa

Au cours de l'Histoire, la paraphrase est venue … 'idées origine

La paraphrase est une figure d'amplification qu … des qualités d
en littérature la copie quasi intégrale d'œuvres … uistiques com
adjectifs, des appositions, etc, pour développe … tions notamm

Attention de ne pas confondre les paronymes : périphrase et paraphrase, figures tout à fait différentes même si leur mode opé

Nombre de reformulation : 0/5 paraphrases

Quitter la partie

Saisissez votre reformulation :

presque complète

Valider | Valider & passer au suivant | Fermer

A l'aide des boutons ci-dessous vous pourrez passer d'une proposition à une autre.

Précédente | Suivante

Phrase d'origine :

125 Mauriciens refoulés par les services de l'immigration britannique [à leur arrivée] à l'aéroport de Heathrow.

Figure 1: Interface of our web-based game for paraphrase acquisition and evaluation. On the top, players reformulate all text spans highlighted by the game creator on any webpage (a Wikipedia article on the example). On the bottom, raters evaluate paraphrases proposed by sets of players using a compact word-lattice view. Note that in its standard definition, the game attributes higher scores to paraphrase candidates that are highly rated and rarer.

$$h_{edit} = \text{TER}(Lem_{orig}, Lem_{para}) \tag{1}$$

Note that this model is not derived from information from the Web, in contrast to all the models described next.

Language model score The likelihood of a sentence can be a good indicator of its grammaticality (Mutton, 2006). Language model probabilities can now be obtained from Web counts. In our experiments, we used the Microsoft Web N-gram Service[6] for research (Wang et al., 2010) to obtain log likelihood scores for text units.[7] However, this score is certainly not sufficient as it does not take the original wording into account. We therefore used a ratio of the language model score of the paraphrased sentence with the language model score of the original sentence, after normalization by sentence length of the language model scores (Onishi et al., 2010):

$$h_{LM\_ratio} = \frac{LM(para)}{LM(orig)} = \frac{lm(para)^{1/length(para)}}{lm(orig)^{1/length(orig)}} \tag{2}$$

Contextless thematic model scores Cooccurring words are used in distributional semantics to account for common meanings of words. We build vector representations of cooccurrences for both the original phrase p and its paraphrase p'. Our contextless thematic model is built in the following fashion: we query a search engine to retrieve the top N document snippets for phrase p. We then count frequencies for all content words in these snippets, and keep the set W of words appearing more than a fraction of N. We then build a vector T (thematic profile) of dimension $|W|$ where values are computed by the following formula:

$$T_{orig}^{nocont}[w] = \frac{count(p, w)}{count(p)} \tag{3}$$

[6] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx

[7] Note that in order to query on French text, we had to remove all diacritics for the service to behave correctly, independently of encodings: careful examination of ranked hypotheses showed that this trick allowed us to obtain results coherent with expectations.

where $count(x)$ correspond to the number of documents containing a given exact phrase or word according to the search engine used and $count(x, y)$ correspond to the number of documents containing simultaneously both. We then compute the same thematic profile for the paraphrase p', using only the subset of words W:

$$T_{para}^{nocont}[w] = \frac{count(p', w)}{count(p)} \tag{4}$$

Finally, we compute a similarity between the two profiles by taking the cosinus between their two vectors:

$$h_{them}^{nocont} = \frac{T_{orig}^{nocont} \cdot T_{para}^{nocont}}{||T_{orig}^{nocont}|| * ||T_{para}^{nocont}||} \tag{5}$$

In all our experiments, we used the Yahoo! Search BOSS[8] Web service for obtaining Web counts and retrieving snippets. Assuming that the distribution of words in W is not biased by the result ordering of the search engine, our model measures some similarity between the most cooccurring content words with p and the same words with p'.

Context-aware thematic model scores Our context-aware thematic model takes into account the words of sentence s in which the substitution of p with p' is attempted. We now consider the set of content words from s (s being the part of the sentence without phrase p) in lieu of the previous set of cooccurring words W, and compute the same profile vectors and similarity between that of the original sentence and that of the paraphrased sentence:

$$h_{them}^{cont} = \frac{T_{orig}^{cont} \cdot T_{para}^{cont}}{||T_{orig}^{cont}|| * ||T_{para}^{cont}||} \tag{6}$$

However, words from s might not be strongly cooccurring with p. In order to increase the likelihood of finding thematically related words, we also build an extended context model, $h_{them}^{extcont}$ where content words from s are supplemented with their most cooccurring words. This is done using the same procedure as that previously used for finding content words cooccurring with p.

5 Experiments

In this section we report on experiments conducted to assess the performance of our proposed approach for validating candidate sub-sentential paraphrases using information from the Web.

5.1 Data used

We randomly extracted 150 original sentences in French and their rewritings from the WiCoPaCo corpus which were marked as paraphrases. Of those, we kept 100 for our training corpus and the remaining 50 for testing. The number of original phrases of each length is reported on Figure 2.

phrase length	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
original phrases	0	3	29	8	6	2	2	0
paraphrases	39	64	74	36	21	10	5	1

Figure 2: Distribution of number of phrases per phrase length in tokens for the test corpus

For each original sentence, we collected 5 candidate paraphrases to simulate the fact that we had a repertoire of paraphrases with the required entries:[9]

- Wicopaco: the original paraphrase from the WiCoPaCo corpus;
- Game: two candidate paraphrases from users of our Web-based game;
- Pivot$_{\text{ES}}$ and Pivot$_{\text{ZH}}$: two candidate paraphrases obtained by translation by pivot, using the Google Translate[10] online SMT system and one language close to French as pivot (Spanish), and another one more distant (Chinese).

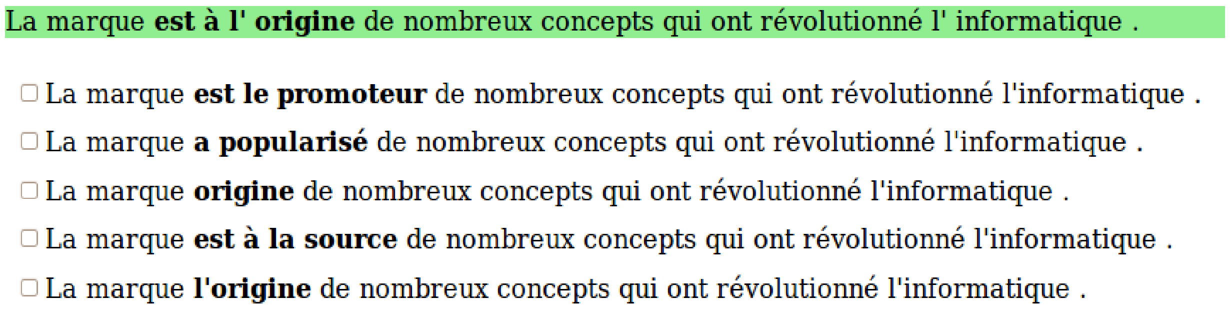
We then presented the original sentence and its 5 paraphrases (in random order) to two judges. Four native speakers took part in our experiments: they all took part in the data collection for one half of the sentences of the training and test corpora and to the evaluation of paraphrases for the other half. For the annotation with two classes (paraphrase vs. not paraphrase), we obtain as inter-judge agreement[11] a

[8] http://developer.yahoo.com/search/boss/

[9] Note that, as a consequence, we did not carry any experiment related to the recall of any technique here.

[10] http://translate.google.com

[11] We used R (http://www.r-project.org) to compute this Cohen's κ value.



La marque **est à l' origine** de nombreux concepts qui ont révolutionné l' informatique .

- □ La marque **est le promoteur** de nombreux concepts qui ont révolutionné l'informatique .
- □ La marque **a popularisé** de nombreux concepts qui ont révolutionné l'informatique .
- □ La marque **origine** de nombreux concepts qui ont révolutionné l'informatique .
- □ La marque **est à la source** de nombreux concepts qui ont révolutionné l'informatique .
- □ La marque **l'origine** de nombreux concepts qui ont révolutionné l'informatique .

Figure 3: Example of an original sentence and its 5 associated candidate paraphrases. The phrase in bold from the original sentence (*The brand* ***is at the origin*** *of many concepts that have revolutionized computing.*) is paraphrased as *est le promoteur* (*is the promoter*), *a popularisé* (*popularized*), *origine* (*origin*), *est à la source* (*is the source*), and *l'origine* (*the origin*).

value of $\kappa = 0.65$, corresponding to a *substantial* agreement according to the literature. An example of the interface used is provided in Figure 3.

We considered that our technique could not propose reliable results when web phrase counts were too low. From the distribution of counts of phrases and paraphrases from our training set (see Figure 4), we empirically chose a threshold of 10 for the minimum count of any phrase. Our corpus was consequently reduced from 750=150*5 to 434 examples for the training corpus, and from 250=50*5 to 215 for the test corpus.

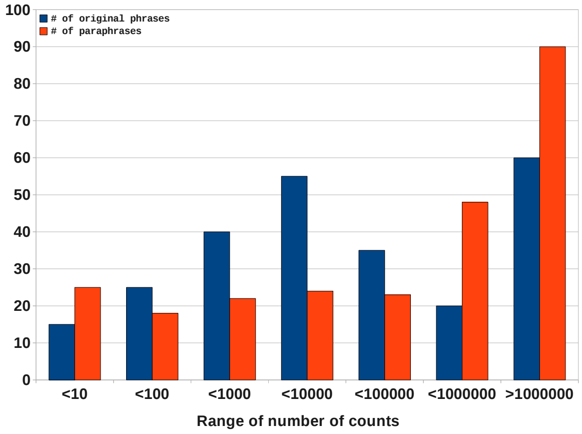


Figure 4: Number of phrases and paraphrases per web count range

Results will be reported for three conditions:

- **Possible**: the gold standard for instances where at least one of the judges indicated "paraphrases" records the pair as a paraphrase. In this condition, the test set has 116 instances that are paraphrases and 99 that are not.
- **Sure**: the gold standard for instances where not all judges indicated "paraphrases" records the pair as not paraphrase. In this condition, the test set has 76 instances that are paraphrases and 139 that are not.
- **Surer**: only those instances where both judges agree are recorded. This reduces our training and test set to respectively 287 and 175 examples. Thus, results on this subcorpora will not be directly comparable with the other results. In this condition, the test set has 76 instances that are paraphrases and 99 that are not.

5.2 Baseline techniques

Web-count based baselines We used two baselines based on simple Web counts. The first one, WEBLM, considers a candidate sentence a paraphrase of the original sentence whenever its Web language model score is higher than that of the original phrase. The second one, BOUNDLM, considers a sentence as a paraphrase whenever the counts for the bigrams crossing the left and right boundary of the sub-sentential paraphrase is higher than 10.

Syntactic dependency baseline When rewriting a subpart of a sentence, the fact that syntactic dependencies between the rewritten phrase and its context are the same than those of the original phrase and the same context can provide some information about the grammatical and semantic substituability of the two phrases (Zhao et al., 2007; Max and Zock, 2008). We thus build syntactic dependencies for both the original and rewritten sentence, using the French version (Candito et al., 2010) of the Berkeley probabilistic parser (Petrov and Klein, 2007), and consider the subset of dependencies for the two sentences that exist between a word inside the phrase under focus and a word outside it (Dep_{orig} and Dep_{para}). Our ContDep baseline considers a sentence as a paraphrase iff $Dep_{para} = Dep_{orig}$.

5.3 Evaluation results

We used the models described in Section 4 to build a SVM classifier using the libsvm package (Chang and Lin, 2001). Accuracy results are reported on Figure 5.

	WebLM	BoundLM	ContDep	Classifier
Possible	**62.79**	54.88	48.53	57.67
Sure	68.37	36.27	51.90	**70.69**
Surer	56.79	51.41	42.69	**62.85**

Figure 5: Accuracy results for the three baselines and our classifier on the test set for the three conditions. Note that the Surer condition cannot be directly compared with the other two as the number of training and test examples are not the same.

The first notable observation is that our task is not surprisingly a difficult one. The best performance achieved is an accuracy of 70.69 with our system in the Sure condition. There are, however, some important variations across conditions, with a result as low as 57.67 for our system in the Possible condition (recall that in this condition candidates are considered paraphrases when only one of the two judges considered it a paraphrase, i.e. when the two judges disagreed).

Overall, the WebLM baseline and our system appear as stronger than the two other baselines. The two lower baselines, BoundLM and ContDep, attempt to model local grammatical constraints, which are not surprisingly not sufficient for paraphrase identification. WebLM is comparatively a much more competitive baseline, but its accuracy in the Surer condition is not very strong. As this latter condition considers only consensual judgements for the two judges, we can hypothesize that the interpretation of its results is more reliable. In this condition, our system obtains the best performance, with a +6.06 advantage over WebLM. As found in other works (e.g. (Bannard and Callison-Burch, 2005)), using language models for paraphrase validation is not sufficient as it cannot model meaning preservation, and our results show that this is also true even when counts are estimated from the Web. Using a ratio of normalized LM scores may have improved the situation a bit.[12]

	Wicopaco	Gamers	Pivot_{ES}	Pivot_{ZH}
Possible	**89.33**	67.00	47.33	20.66
Sure	**64.00**	44.50	31.33	10.66
Surer	**86.03**	57.34	37.71	12.60

Figure 6: Paraphrase accuracy of our different paraphrase acquisition methods for the three conditions.

Lastly, we report in Figure 6 the paraphrase accuracy of each individual acquisition technique (i.e. source of paraphrases from the preexisting repertoire). The original rewritting from WiCoPaCo obtains not surprisingly a very high paraphrase accuracy, in particular in the Possible and Surer conditions. Paraphrases obtained through our Web-based game have an acceptable accuracy: the numbers confirm that paraphrase pairs are highly context-dependent, because the pairs which were likely to be paraphrases in the context of the game are not necessarily so in a different context. This, of course, may be due to a number of reasons that we will have to investigate. Lastly, there is a significant drop in accuracy for the automatic pivot paraphrasers, but pivoting through Spanish obtained, not suprisingly again, a much better performance than pivoting through Chinese.

6 Discussion and future work

We have presented an approach to the task of targeted paraphrasing in the context of text revision, a scenario which was supported by naturally-occurring data from the rephrasing memory of Wikipedia. Our framework takes a repertoire of existing sub-sentential paraphrases, coming from pos-

[12] A possible explanation for the relative good performance of WebLM may lie in the fact that our two automatic paraphrasers using Google Translate as a pivot translation engine tend to produce strings that are very likely according to the language models used by the translation system, which we assume to be very comparable to those that were used in our experiments.

sibly any source including manual acquisition, and validates all candidate paraphrases using information from the Web. Our experiments have shown that the current version of our classifier outperforms several baselines when considering paraphrases with consensual judgements in the gold standard reference.

Although our initial experiments are positive, we believe that they can be improved in a number of ways. We intend to broaden our exploration of the various characteristics at play. We will try more features, including e.g. a model of syntactic dependencies derived from the Web, and extend our work to new languages. We will also attempt to analyze more precisely our results to identify problematic cases, some of which could turn to be almost impossible to model without resorting to world knowledge, which was beyond our attempted modeling. Finally, we will also be interested in considering the applicability of this approach as a framework for the evaluation of paraphrase acquisition techniques.

Acknowledgments

This work was partly supported by ANR project Trace (ANR-09-CORD-023). The authors would like to thank the anonymous reviewers for their helpful questions and comments.

References

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

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

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

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the ACM symposium on User interface software and technology*.

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

Houda Bouamor, Aurélien Max, and Anne Vilnat. 2011. Monolingual alignment by edit rate computation on sentential paraphrase pairs. In *Proceedings of ACL, Short Papers session*, Portland, USA.

Chris Brockett and William B. Dolan. 2005. Support vector machines for paraphrase identification and corpus construction. In *Proceedings of The 3rd International Workshop on Paraphrasing IWP*, Jeju Island, South Korea.

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

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of LREC*, Valletta, Malta.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of COLING*, Copenhagen, Denmark.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Timothy Chklovski. 2005. Collecting paraphrase corpora from volunteer contributors. In *Proceedings of KCAP 2005*, Banff, Canada.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of COLING*, Manchester, UK.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4):597–614.

Michael Connor and Dan Roth. 2007. Context sensitive paraphrasing with a global unsupervised classifier. In *Proceedings of ECML*, Warsaw, Poland.

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.

Camille Dutrey, Houda Bouamor, Delphine Bernhard, and Aurélien Max. 2011. Local modifications and paraphrases in wikipedia's revision history. *SEPLN journal*, 46:51–58.

Cristina España Bonet, Marta Vila, M. Antònia Martí, and Horacio Rodríguez. 2009. Coco, a web interface for corpora compilation. *SEPLN journal*, 43.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Proceedings of NAACL-HLT*, Los Angeles, USA.

Mirella Lapata and Frank Keller. 2005. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31.

Weigang Li, Ting Liu, Yu Zhang, Sheng Li, and Wei He. 2005. Automated generalization of phrasal paraphrases from the web. In *Proceedings of the IJCNLP Workshop on Paraphrasing*, Jeju Island, South Korea.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3).

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of AMTA*, Waikiki, USA.

Aurélien Max and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of LREC 2010*, Valletta, Malta.

Aurélien Max and Michael Zock. 2008. Looking up phrase rephrasings via a pivot language. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*, Manchester, United Kingdom.

Aurélien Max. 2004. From controlled document authoring to interactive document normalization. In *Proceedings of COLING*, Geneva, Switzerland.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2).

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of ACL*, Prague, Czech Republic.

Andrew Mutton. 2006. *Evaluation of sentence grammaticality using Parsers and a Support Vector Machine*. Ph.D. thesis, Macquarie University.

Takashi Onishi, Masao Utiyama, and Eiichiro Sumita. 2010. Paraphrase Lattice for Statistical Machine Translation. In *Proceedings of ACL, Short Papers session*, Uppsala, Sweden.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignement 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.

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

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving translation via targeted paraphrasing. In *Proceedings of EMNLP*, Cambridge, MA.

Josh Schroeder, Trevor Cohn, and Philipp Koehn. 2009. Word lattices for multi-source translation. In *Proceedings of EACL*, Athens, Greece.

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).

Kuansan Wang, Chris Thrasher, Evelyne Viegas, Xiaolong Li, and Bo-june (Paul) Hsu. 2010. An Overview of Microsoft Web N-gram Corpus and Applications. In *Proceedings of the NAACL-HLT Demonstration Session*, Los Angeles, USA.

Shiqi Zhao, Ting Liu, Xincheng Yuan, Sheng Li, and Yu Zhang. 2007. Automatic acquisition of context-specific lexical paraphrases. In *Proceedings of IJCAI 2007*, Hyderabad, India.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint ACL-IJCNLP*, Singapore.

Shiqi Zhao, Haifeng Wang, Ting Liu, , and Sheng Li. 2010. Leveraging multiple mt engines for paraphrase generation. In *Proceedings of COLING*, Beijing, China.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of COLING*, Beijing, China.