Validation of sub-sentential paraphrases acquired from parallel monolingual corpora

Houda Bouamor **Aurélien Max** **Anne Vilnat**

LIMSI-CNRS & Univ. Paris Sud
Orsay, France
firstname.lastname@limsi.fr

Abstract

The task of paraphrase acquisition from related sentences can be tackled by a variety of techniques making use of various types of knowledge. In this work, we make the hypothesis that their performance can be increased if candidate paraphrases can be validated using information that characterizes paraphrases independently of the set of techniques that proposed them. We implement this as a bi-class classification problem (i.e. *paraphrase* vs. *not paraphrase*), allowing any paraphrase acquisition technique to be easily integrated into the combination system. We report experiments on two languages, English and French, with 5 individual techniques on parallel monolingual parallel corpora obtained *via* multiple translation, and a large set of classification features including surface to contextual similarity measures. Relative improvements in F-measure close to 18% are obtained on both languages over the best performing techniques.

1 Introduction

The fact that natural language allows messages to be conveyed in a great variety of ways constitutes an important difficulty for NLP, with applications in both text analysis and generation. The term *paraphrase* is now commonly used in the NLP litterature to refer to textual units of equivalent meaning at the phrasal level (including single words). For instance, the phrases *six months* and *half a year* form a paraphrase pair applicable in many different contexts, as they would appropriately denote the same concept. Although one can envisage to manually build high-coverage lists of synonyms, enumerating meaning equivalences at the level of phrases is too daunting a task for humans. Because this type of knowledge can however greatly benefit many NLP applications, automatic acquisition of such paraphrases has attracted a lot of attention (Androutsopoulos and Malakasiotis, 2010; Madnani and Dorr, 2010), and significant research efforts have been devoted to this objective (Callison-Burch, 2007; Bhagat, 2009; Madnani, 2010).

Central to acquiring paraphrases is the need of assessing the quality of the candidate paraphrases produced by a given technique. Most works to date have resorted to human evaluation of paraphrases on the levels of *grammaticality* and *meaning equivalence*. Human evaluation is however often criticized as being both costly and non reproducible, and the situation is even more complicated by the inherent complexity of the task that can produce low inter-judge agreement. Task-based evaluation involving the use of paraphrasing into some application thus seem an acceptable solution, provided the evaluation methodologies for the given task are deemed acceptable. This, in turn, puts the emphasis on observing the impact of paraphrasing on the targeted application and is rarely accompanied by a study of the intrinsic limitations of the paraphrase acquisition technique used.

The present work is concerned with the task of sub-sentential paraphrase acquisition from pairs of related sentences. A large variety of techniques have been proposed that can be applied to this task. They typically make use of different kinds of automatically or manually acquired knowledge. We make the hypothesis that their performance can be increased if candidate para-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 716–725,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

phrases can be validated using information that characterize paraphrases in complement to the set of techniques that proposed them. We propose to implement this as a bi-class classification problem (i.e. *paraphrase* vs. *not paraphrase*), allowing any paraphrase acquisition technique to be easily integrated into the combination system. In this article, we report experiments on two languages, English and French, with 5 individual techniques based on *a*) statistical word alignment models, *b*) translational equivalence, *c*) handcoded rules of term variation, *d*) syntactic similarity, and *e*) edit distance on word sequences. We used parallel monolingual parallel corpora obtained *via* multiple translation from a single language as our sources of related sentences, and a large set of features including surface to contextual similarity measures. Relative improvements in F-measure close to 18% are obtained on both languages over the best performing techniques.

The remainder of this article is organized as follows. We first briefly review previous work on sub-sentential paraphrase acquisition in section 2. We then describe our experimental setting in section 3 and the individual techniques that we have studied in section 4. Section 5 is devoted to our approach for validating paraphrases proposed by individual techniques. Finally, section 6 concludes the article and presents some of our future work in the area of paraphrase acquisition.

2 Related work

The hypothesis that if two words or, by extension, two phrases, occur in similar contexts then they may be interchangeable has been extensively tested. The *distributional hypothesis*, attributed to Zellig Harris, was for example applied to syntactic dependency paths in the work of Lin and Pantel (2001). Their results take the form of equivalence patterns with two arguments such as $\{$*X asks for Y*, *X requests Y*, *X's request for Y*, *X wants Y*, *Y is requested by X*, ...$\}$.

Using comparable corpora, where the same information probably exists under various linguistic forms, increases the likelihood of finding very close contexts for sub-sentential units. Barzilay and Lee (2003) proposed a multi-sequence alignment algorithm that takes structurally similar sentences and builds a compact lattice representation that encodes local variations. The work by Bhagat and Ravichandran (2008) describes an application of a similar technique on a very large scale.

The hypothesis that two words or phrases are interchangeable if they share a common translation into one or more other languages has also been extensively studied in works on sub-sentential paraphrase acquisition. Bannard and Callison-Burch (2005) described a pivoting approach that can exploit bilingual parallel corpora in several languages. The same technique has been applied to the acquisition of local paraphrasing patterns in Zhao et al. (2008). The work of Callison-Burch (2008) has shown how the monolingual context of a sentence to paraphrase can be used to improve the quality of the acquired paraphrases.

Another approach consists in modelling local paraphrasing identification rules. The work of Jacquemin (1999) on the identification of term variants, which exploits rewriting morphosyntactic rules and descriptions of morphological and semantic lexical families, can be extended to extract the various forms corresponding to input patterns from large monolingual corpora.

When parallel monolingual corpora aligned at the sentence level are available (e.g. multiple translations into the same language), the task of sub-sentential paraphrase acquisition can be cast as one of word alignment between two aligned sentences (Cohn et al., 2008). Barzilay and McKeown (2001) applied the distributionality hypothesis on such parallel sentences, and Pang *et al.* (2003) proposed an algorithm to align sentences by recursive fusion of their common syntactic constituants.

Finally, they has been a recent interest in automatic evaluation of paraphrases (Callison-Burch et al., 2008; Liu et al., 2010; Chen and Dolan, 2011; Metzler et al., 2011).

3 Experimental setting

We used the main aspects of the methodology described by Cohn et al. (2008) for constructing evaluation corpora and assessing the performance of techniques on the task of sub-sentential paraphrase acquisition. Pairs of related sentences are hand-aligned to define a set of reference *atomic* paraphrase pairs at the level of words or phrases, denoted as $\mathcal{R}_{\text{atom}}$[1].

[1] Note that in this study we do not distinguish between "Sure" and "Possible" alignments, and when reusing anno-

	single language translation	**multiple language translation**	**video descriptions**	**multiply-translated subtitles**	**news headlines**
# tokens	4,476	4,630	1,452	2,721	1,908
# unique tokens	656	795	357	830	716
% aligned tokens (excluding identities)	60.58	48.80	23.82	29.76	14.46
lexical overlap (tokens)	77.21	61.03	59.50	32.51	39.63
lexical overlap (lemmas content words)	83.77	71.04	64.83	39.54	45.31
translation edit rate (TER)	0.32	0.55	0.76	0.68	0.62
penalized *n*-gram prec. (BLEU)	0.33	0.15	0.13	0.14	0.39

Table 1: Various indicators of sentence pair comparability for different corpus types. Statistics are reported for French on sets of 100 sentence pairs.

We conducted a small-scale study to assess different types of corpora of related sentences:

1. **single language translation** Corpora obtained by several independent human translation of the same sentences (e.g. (Barzilay and McKeown, 2001)).

2. **multiple language translation** Same as above, but where a sentence is translated from 4 different languages into the same language (Bouamor et al., 2010).

3. **video descriptions** Descriptions of short YouTube videos obtained *via* Mechanical Turk (Chen and Dolan, 2011).

4. **multiply-translated subtitles** Aligned multiple translations of contributed movie subtitles (Tiedemann, 2007).

5. **comparable news headlines** News headlines collected from Google News clusters (e.g. (Dolan et al., 2004)).

We collected 100 sentence pairs of each type in French, for which various comparability measures are reported on Table 1. In particular, the "**% aligned tokens**" row indicates the proportion of tokens from the sentence pairs that could be manually aligned by a native-speaker annotator.[2] Obviously, the more common tokens two sentences from a pair contain, the fewer subsentential paraphrases may be extracted from that pair. However, high lexical overlap increases the probability that two sentences be indeed paraphrases, and in turn the probability that some of their phrases be paraphrases. Furthermore, the presence of common token may serve as useful clues to guide paraphrase extraction.

For our experiments, we chose to use parallel monolingual corpora obtained by single language translation, the most direct resource type for acquiring sub-sentential paraphrase pairs. This allows us to define acceptable references for the task and resort to the most consensual evaluation technique for paraphrase acquisition to date. Using such corpora, we expect to be able to extract *precise* paraphrases (see Table 1), which will be natural candidates for further validation, which will be addressed in section 5.3.

Figure 1 illustrates a reference alignment obtained on a pair of English sentential paraphrases and the list of atomic paraphrase pairs that can be extracted from it, against which acquisition techniques will be evaluated. Note that we do not consider pairs of identical units during evaluation, so we filter them out from the list of reference paraphrase pairs.

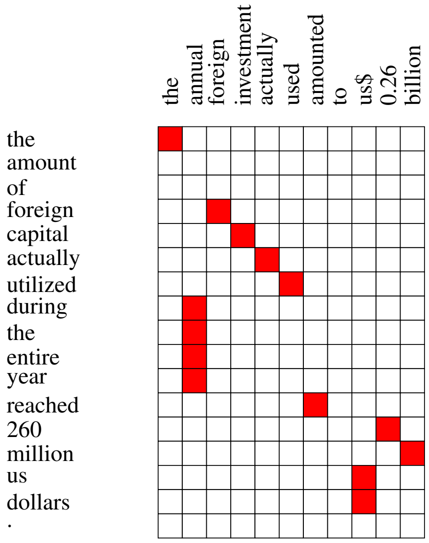
The example in Figure 1 shows different cases that point to the inherent complexity of this task, even for human annotators: it could be argued, for instance, that a correct atomic paraphrase pair should be *reached* ↔ *amounted to* rather than *reached* ↔ *amounted*. Also, aligning independently *260* ↔ *0.26* and *million* ↔ *billion* is assuredly an error, while the pair *260 million* ↔ *0.26 billion* would have been appropriate. A case of alignment that seems non trivial can be observed in the provided example (*during the entire year* ↔ *annual*). The abovementioned reasons will explain in part the difficulties in reaching high performance values using such gold standards.

Reference *composite* paraphrase pairs (denoted as $\mathcal{R}$), obtained by joining adjacent atomic paraphrase pairs from $\mathcal{R}_{\text{atom}}$ up to 6 tokens[3], will

tated corpora using them we considered all alignments as being correct.

[2] The same annotator hand-aligned the 5*100=500 paraphrase pairs using the YAWAT (Germann, 2008) manual alignment tool.

[3] We used standard biphrase extraction heuristics (Koehn



capital ↔ investment
utilized ↔ used
during the entire year ↔ annual
reached ↔ amounted
260 ↔ 0.26
million ↔ billion
us dollars ↔ us$

Figure 1: Reference alignments for a pair of English sentential paraphrases from the annotation corpus of Cohn et al. (2008) (note that possible and sure alignments are not distinguished here) and the list of atomic paraphrase pairs extracted from these alignments.

also be considered when measuring performance. Evaluated techniques have to output atomic candidate paraphrase pairs (denoted as $\mathcal{H}_{\text{atom}}$) from which composite paraphrase pairs (denoted as $\mathcal{H}$) are computed. The usual measures of *precision* (P), *recall* (R) and *F-measure* (F_1) can then be defined in the following way (Cohn et al., 2008):

$$P = \frac{|\mathcal{H}_{\text{atom}} \cap \mathcal{R}|}{|\mathcal{H}_{\text{atom}}|} \; R = \frac{|\mathcal{H} \cap \mathcal{R}_{\text{atom}}|}{|\mathcal{R}_{\text{atom}}|} \; F_1 = \frac{2pr}{p+r}$$

We conducted experiments using two different corpora in English and French. In each case, a held-out development corpus of 150 sentential paraphrase pairs was used for development and tuning, and all techniques were evaluated on the same test set consisting of 375 sentential paraphrase pairs. For English, we used the MTC corpus described in (Cohn et al., 2008), consisting of multiply-translated Chinese sentences into English, and used as our gold standard both the alignments marked as "Sure" and "Possible". For French, we used the CESTA corpus of news articles[4] obtained by translating into French from English.

We used the YAWAT (Germann, 2008) manual alignment tool. Inter-annotator agreement values (averaging with each annotation set as the gold standard) are 66.1 for English and 64.6 for French, which we interpret as acceptable values. Manual inspection of the two corpora reveals 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.

4 Individual techniques for paraphrase acquisition

As discussed in section 2, the acquisition of sub-sentential paraphrases is a challenging task that has previously attracted a lot of work. In this work, we consider the scenario where sentential paraphrases are available and words and phrases from one sentence can be aligned to words and phrases from the other sentence to form atomic paraphrase pairs. We now describe several techniques that perform the task of sub-sentential unit alignment. We have selected and implemented five techniques which we believe are representative of the type of knowledge that these techniques use, and have reused existing tools, initially developed for other tasks, when possible.

4.1 Statistical learning of word alignments (Giza)

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 Statistical 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 provided GIZA++ with all possible sentence pairs from our mutiply-translated corpus to improve the quality of its word alignments (note that

et al., 2007) : all words from a phrase must be aligned to at least one word from the other and not to words outside, but unaligned words at phrase boundaries are not used.

[4] http://www.elda.org/article125.html

we used symmetrized alignments from the alignments in both directions). This constitutes a significant advantage for this technique that techniques working on each sentence pair independently do not have.

4.2 Translational equivalence (Pivot)

Translational equivalence can be exploited to determine that two phrases may be paraphrases. Bannard and Callison-Burch (2005) defined a paraphrasing probability between two phrases based on their translation probability through all possible *pivot* phrases as:

$$P_{para}(p_1, p_2) = \sum_{piv} P_t(piv|p_1) P_t(p_2|piv)$$

where P_t denotes translation probabilies. We used the Europarl corpus[5] of parliamentary debates in English and French, consisting of approximately 1.7 million parallel sentences : this allowed us to use the same resource to build paraphrases for English, using French as the pivot language, and for French, using English as the pivot language. The GIZA++ tool was used for word alignment and the MOSES Statistical Machine Translation toolkit (Koehn et al., 2007) was used to compute phrase translation probabilities from these word alignments. For each sentential paraphrase pair, we applied the following algorithm: for each phrase, we build the entire set of paraphrases using the previous definition. We then extract its best paraphrase as the one exactly appearing in the other sentence with maximum paraphrase probability, using a minimal threshold value of 10^{-4}.

4.3 Linguistic knowledge on term variation (Fastr)

The FASTR tool (Jacquemin, 1999) was designed to spot term/phrase 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 morphosyntactic categories. Paradigmatic variation can also be expressed by expressing constraints between words, imposing that they be of the same morphological or semantic family. Both constraints rely on preexisting repertoires available for English and French. To compute candidate paraphrase pairs using FASTR, we first consider all phrases from the first sentence and search for variants in the other sentence, then do the reverse process and finally take the intersection of the two sets.

4.4 Syntactic similarity (Synt)

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 constituents from fusionning when there is evidence of the presence of a word in another constituent of one of the sentence. We use the Berkeley Probabilistic parser (Klein and Manning, 2003) to obtain syntactic trees for English and its adapted version for French (Candito et al., 2010). Because this process is highly sensitive to syntactic parse errors, we use in our implementation k-best parses and retain the most compact fusion from any pair of candidate parses.

4.5 Edit rate on word sequences (TER$_p$)

TER$_p$ (Translation Edit Rate Plus) (Snover et al., 2010) is a score designed for the evaluation of Machine Translation 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 optionally variants substitution through stemming, synonym or paraphrase matching.[6] Each edit type is parameterized by at least one weight which can be optimized using e.g. hill climbing. TER$_p$ being a tunable metric, our experiments will include tuning TER$_p$ systems towards either precision ($\rightarrow P$), recall ($\rightarrow R$), or F-measure ($\rightarrow F_1$).[7]

4.6 Evaluation of individual techniques

Results for the 5 individual techniques are given on the left part of Table 2. It is first apparent that all techniques but TER$_p$ 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 (from

[5] http://statmt.org/europarl

[6] Note that for these experiments we did not use the stemming module, the interface to WordNet for synonym matching and the provided paraphrase table for English, due to the fact that these resources were available for English only.

[7] *Hill climbing* was used for all tunings as done by Snover et al. (2010), and we used one iteration starting with uniform weights and 100 random restarts.

	Individual techniques							Combinations	
	GIZA	PIVOT	FASTR	SYNT	TER$_p$			union	validation
					$\rightarrow P$	$\rightarrow R$	$\rightarrow F_1$		
	English								
P	31.01	31.78	37.38	**52.17**	50.00	29.15	33.37	21.44	50.51
R	38.30	18.50	6.71	2.53	5.83	45.19	**45.37**	**60.87**	41.19
F_1	34.27	23.39	11.38	4.83	10.44	35.44	**38.46**	31.71	**45.37**
	French								
P	28.99	29.53	52.48	**62.50**	31.35	30.26	31.43	17.58	40.77
R	**45.98**	26.66	8.59	8.65	44.22	44.60	44.10	**63.36**	45.85
F_1	35.56	28.02	14.77	15.20	**36.69**	36.05	**36.70**	27.53	**43.16**

Table 2: Results on the test set on English and French for the 5 individual paraphrase acquisition techniques (left part) and for the 2 combination techniques (right part).

English to French, compared with from Chinese to English), which should be consequently easier to word-align. This is for example clearly shown by the results of the statistical aligner GIZA, which obtains a 7.68 advantage on recall for French over English.

The two linguistically-aware techniques, FASTR and SYNT, have a very strong precision on the more parallel French corpus, but fail to achieve an acceptable recall on their own. This is not surprising : FASTR metarules are focussed on term variant extraction, and SYNT requires two syntactic trees to be highly comparable to extract sub-sentential paraphrases. When these constrained conditions are met, these two techniques appear to perform quite well in terms of precision.

GIZA and TER$_p$ perform roughly in the same range on French, with acceptable precision and recall, TER$_p$ performing overall better, with e.g. a 1.14 advantage on F-measure on French and 4.19 on English. The fact that TER$_p$ performs comparatively better on English than on French[8], with a 1.76 advantage on F-measure, is not contradictory: the implemented edit distance makes it possible to align reasonably distant words and phrases independently from syntax, and to find alignments for close remaining words, so the differences of performance between the two languages are not necessarily expected to be comparable with the results of a statistical alignment technique. English being a poorly-inflected language, alignment clues between two sentential paraphrases are expected to be more numerous than for highly-inflected French.

PIVOT is on par with GIZA as regards precision, but obtains a comparatively much lower recall (differences of 19.32 and 19.80 on recall on French and English respectively). This may first be due in part to the paraphrasing score threshold used for PIVOT, but most certainly to the use of a bilingual corpus from the domain of parliamentary debates to extract paraphrases when our test sets are from the news domain: we may be observing differences inherent to the domain, and possibly facing the issue of numerous "out-of-vocabulary" phrases, in particular for named entities which frequently occur in the news domain.

Importantly, we can note that we obtain at best a recall of 45.98 on French (GIZA) and of 45.37 on English (TER$_p$). This may come as a disappointment but, given the broad set of techniques evaluated, this should rather underline the inherent complexity of the task. Also, recall that the metrics used do not consider identity paraphrases (e.g. *at the same time* ↔ *at the same time*), as well as the fact that gold standard alignment is a very difficult process as shown by interjudge agreement values and our example from section 3. This, again, confirms that the task that is addressed is indeed a difficult one, and provides further justification for initially focussing on *parallel* monolingual corpora, albeit scarce, for conducting fine-grained studies on sub-sentential paraphrasing.

Lastly, we can also note that precision is not very high, with (at best, using TER$_{p\rightarrow P}$) average values for all techniques of 40.97 and 40.46 on French and English, respectively. Several facts may provide explanations for this observation. First, it should be noted that none of those techniques, except SYNT, was originally developed

[8] Recall that all specific linguistic modules for English only from TER$_p$ had been disabled, so the better performance on English cannot be explained by a difference in terms of resources used.

for the task of sub-sentential paraphrase acquisition from monolingual parallel corpora. This results in definitions that are at best closely related to this task.[9] Designing new techniques was not one of the objectives of our study, so we have reused existing techniques, originally developed with different aims (bilingual parallel corpora word alignment (GIZA), term variant recognition (FASTR), Machine Translation evaluation (TER$_p$)). Also, techniques such as GIZA and TER$_p$ attempt to align as many words as possible in a sentence pair, when gold standard alignments sometimes contain gaps.[10] Finally, the metrics used will count as false small variations of gold standard paraphrases (e.g. missing function word): the acceptability or not of such candidates could be either evaluated in a scenario where such "acceptable" variants would be taken into account, and could be considered in the context of some actual use of the acquired paraphrases in some application. Nonetheless, on average the techniques in our study produce more candidates that are not in the gold standard: this will be an important fact to keep in mind when tackling the task of combining their outputs. In particular, we will investigate the use of features indicating the combination of techniques that predicted a given paraphrase pair, aiming to capture consensus information.

5 Paraphrase validation

5.1 Technique complementarity

Before considering combining and validating the outputs of individual techniques, it is informative to look at some notion of "complementarity" between techniques, in terms of how many correct paraphrases a technique would add to a combined set. The following formula was used to account for the complementarity between the set of candidates from some technique i, t_i, and the set for some technique j, t_j:

$$C(t_i, t_j) = \text{recall}(t_i \cup t_j) - \max(\text{recall}(t_i), \text{recall}(t_j))$$

Results on the test set for the two languages are given in Table 3. A number of pairs of techniques have strong complementarity values, the strongest one being for GIZA and TER$_p$ for both languages. According to these figures, PIVOT identify paraphrases which are slightly more similar to those of TER$_p$ than those of GIZA. Interestingly, FASTR and SYNT exhibit a strong complementarity, where in French, for instance, they only have a very small proportion of paraphrases in common. Considering the set of all other techniques, GIZA provides the more new paraphrases on French and TER$_p$ on English.

	GIZA	PIVOT	FASTR	SYNT	TER$_{p\rightarrow R}$	*all others*
	English					
GIZA	-	4.65	2.83	0.59	**10.31**	8.31
PIVOT	**4.65**	-	2.30	1.88	3.12	3.72
FASTR	**2.83**	2.30	-	2.42	1.71	0.53
SYNT	0.59	1.88	**2.42**	-	0.59	0.00
TER$_{p\rightarrow R}$	**10.31**	3.12	1.71	0.59	-	12.20
	French					
GIZA	-	9.79	3.64	2.20	**10.73**	8.91
PIVOT	**9.79**	-	2.26	5.22	7.84	3.39
FASTR	3.64	2.26	-	**7.28**	3.01	0.19
SYNT	2.20	5.22	**7.28**	-	1.76	0.44
TER$_{p\rightarrow R}$	**10.73**	7.84	3.01	1.76	-	5.65

Table 3: Values of *complementarity* on the test set for both languages, where the following formula was used for the set of technique outputs $T = \{t_1, t_2, ..., t_n\}$: $C(t_i, t_j) = \text{recall}(t_i \cup t_j) - \max(\text{recall}(t_i), \text{recall}(t_j))$. Complementarity values are computed between all pairs of individual techniques, and each individual technique and the set of all other techniques. Values in bold indicate highest values for the technique of each row.

5.2 Naive combination by union

We first implemented a naive combination obtained by taking the union of all techniques. Results are given in the first column of the right part of Table 2. The first result is quite encouraging: in both languages, more than 6 paraphrases from the gold standard out of 10 are found by at least one of the techniques, which, given our previous discussion, constitutes a good result and provide a clear justification for combining different techniques for improving performance on this task. Precision is mechanically lowered to account for roughly 1 correct paraphrase over 5 candidates for both languages. F-measure values are much lower than those of TER$_p$ and GIZA, showing that the union of all techniques is only interesting for recall-oriented paraphrase acquisition. In

[9] Recall, however, that our best performing technique on F-measure, TER$_p$, was optimized to our task using a held out development set.

[10] It is arguable whether such cases should happen in sentence pairs obtained by translating the same original sentence into the same language, but this clearly depends on the interpretation of the expected level of annotation by the annotators.

the next section, we will show how the results of the union can be validated using machine learning to improve these figures.

5.3 Paraphrase validation via automatic classification

A natural improvement to the naive combination of paraphrase candidates from all techniques can consist in validating candidate paraphrases by using several models that may be good indicators of their paraphrasing status. We can therefore cast our problem as one of biclass classification (i.e. "paraphrase" vs. "not paraphrase").

We have used a maximum entropy classifier[11] with the following features, aiming at capturing information on the paraphrase status of a candidate pair:

Morphosyntactic equivalence (POS) It may be the case that some sequences of part-of-speech can be rewritten as different sequences, e.g. as a result of verb nominalization. We therefore use features to indicate the sequences of part-of-speech for a pair of candidate paraphrases. We used the preterminal symbols of the syntactic trees of the parser used for SYNT.

Character-based distance (CAR) Morphological variants often have close word forms, and more generally close word forms in sentential paraphase pairs may indicate related words. We used features for discretized values of the edit distance between the two phrases of a candidate paraphrase pair as measured by the Levenshtein distance.

Stem similarity (STEM) Inflectional morphology, which is quite productive in languages such as French, can increase vocabulary size significantly, while in sentential paraphrases common stems may indicate related words. We used a binary feature indicating whether the stemmed phrases of a candidate paraphrase pair match.[12]

Token set identity (BOW) Syntactic rearrangements may involve the same sets of words in various orders. We used discretized features indicating the proportion of common tokens in the set of tokens for the two phrases of a candidate paraphrase pair.

Context similarity (CTXT) It can be derived from the distributionality hypothesis that the more two phrases will be seen in similar contexts, the more they are likely to be paraphrases. We used discretized features indicating how similar the contexts of occurrences of two paraphrases are. For this, we used the full set of bilingual English-French data available for the translation task of the Workshop on Statistical Machine Translation[13], totalling roughly 30 million parallel sentences: this again ensures that the same resources are used for experiments in the two languages. We collect all occurrences for the phrases in a pair, and build a vector of content words cooccurring within a distance of 10 words from each phrase. We finally compute the cosine between the vectors of the two phrases of a candidate paraphrase pair.

Relative position in a sentence (REL) Depending on the language in which parallel sentences are analyzed, it may be the case that subsentential paraphrases occur at close locations in their respective sentence. We used a discretized feature indicating the relative position of the two phrases in their original sentence.

Identity check (COOC) We used a binary feature indicating whether one of the two phrases from a candidate pair, or the two, occurred at some other location in the other sentence.

Phrase length ratio (LEN) We used a discretized feature indicating phrase length ratio.

Source techniques (SRC) Finally, as our setting validates paraphrase candidates produced by a set of techniques, we used features indicating which combination of techniques predicted a paraphrase candidate. This can allow learning that paraphrases in the intersection of the predicted sets for some techniques may produce good results.

We used a held out training set consisting of 150 sentential paraphrase pairs from the same corpora as our previous developement and test sets for both languages. Positive examples were taken from the candidate paraphrase pairs from any of

[11] We used the implementation available at: `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

[12] We use the implementations of the Snowball stemmer from English and French available from: `http://snowball.tartarus.org`

[13] `http://www.statmt.org/wmt11/translation-task.html`

the 5 techniques in our study which belong to the gold standard, and we used a corresponding number of negative examples (randomly selected) from candidate pairs not in the gold standard. The right part of Table 2 provides the results for our validation experiments of the union set for all previous techniques.

We obtain our best results for this study using the output of our validation classifier over the set of all candidate paraphrase pairs. On French, it yields an improvement in F-measure (43.16) of +6.46 over the best individual technique (TER$_p$) and of +15.63 over the naive union from all individual techniques. On English, the improvement in F-measure (45.37) is for the same conditions of respectively +6.91 (over TER$_p$) and +13.66. We unfortunately observe an important decrease in recall over the naive union, of respectively -17.54 and -19.68 for French and English. Increasing our amount of training data to better represent the full range of paraphrase types may certainly overcome this in part. This would indeed be sensible, as better covering the variety of paraphrase types as a one-time effort would help all subsequent validations. Figure 2 shows how performance varies on French with number of training examples for various feature configurations. However, some paraphrase types will require integration of more complex knowledge, as is the case, for instance, for paraphrase pairs involving some anaphora and its antecedent (e.g. *China* ↔ *it*).

While these results, which are very comparable for the two languages studied, are already satisfying given the complexity of our task, further inspection of false positives and negatives may help us to develop additional models that will help us obtain a better classification performance.

6 Conclusions and future work

In this article, we have addressed the task of combining the results of sub-sentential paraphrase acquition from parallel monolingual corpora using a large variety of techniques. We have provided justifications for using highly parallel corpora consisting of multiply translated sentences from a single language. All our experiments were conducted on both English and French using comparable resources, so although the results cannot be directly compared they give some acceptable comparison points. The best recall of any individual technique is around 45 for both language, and F-measure in the range 36-38, indicating that the task under study is a very challenging one. Our validation strategy based on bi-class classification using a broad set of features applicable to all candidate paraphrase pairs allowed us to obtain a 18% relative improvement in F-measure over the best individual technique for both languages.

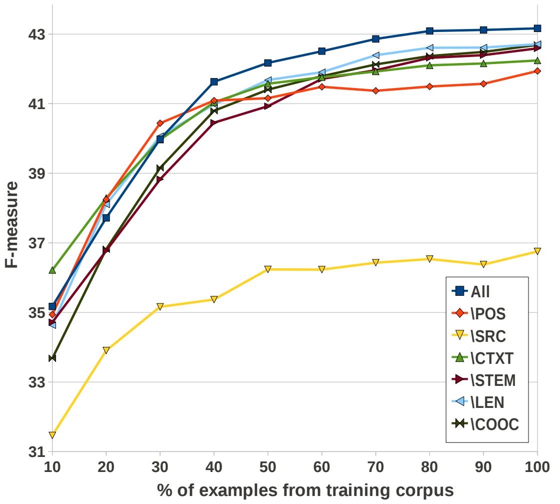


Figure 2: Learning curves obtained on French by removing features individually.

Our future work include performing a deeper error analysis of our current results, to better comprehend what characteristics of paraphrase still defy current validation. Also, we want to investigate adding new individual techniques to provide so far unseen candidates. Another possible approach would be to submit all pairs of sub-sentential paraphrase pairs from a sentence pair to our validation process, which would obviously require some optimization and devising sensible heuristics to limit time complexity. We also intend to collect larger corpora for all other corpus types appearing in Table 1 and conducting anew our acquisition and validation tasks.

Acknowledgements

The authors would like to thank the reviewers for their comments and suggestions, as well as Guillaume Wisniewski for helpful discussions. This work was partly funded by ANR project Edylex (ANR-09-CORD-008).

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38:135–187.

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 R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*, Toulouse, France.

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

Rahul Bhagat. 2009. *Learning Paraphrases from Text*. Ph.D. thesis, University of Southern California.

Houda Bouamor, Aurélien Max, and Anne Vilnat. 2010. Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. In *Proceedings of IceTAL*, Rejkavik, Iceland.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of COLING*, Manchester, UK.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.

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.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of ACL*, Portland, USA.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4).

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*, Geneva, Switzerland.

Ulrich Germann. 2008. Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-HLT, demo session*, Columbus, USA.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL*, College Park, USA.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*, Sapporo, Japan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. PEM: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of EMNLP*, Cambridge, USA.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, University of Maryland College Park.

Donald Metzler, Eduard Hovy, and Chunliang Zhang. 2011. An empirical evaluation of data-driven paraphrase generation techniques. In *Proceedings of ACL-HLT*, Portland, USA.

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 alignement of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of NAACL-HLT*, Edmonton, Canada.

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).

Jörg Tiedemann. 2007. Building a Multilingual Parallel Subtitle Corpus. In *Proceedings of the Conference on Computational Linguistics in the Netherlands*, Leuven, Belgium.

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.