Towards a General and Extensible Phrase-Extraction Algorithm

Wang Ling, Tiago Luís, João Graça, Luísa Coheur and Isabel Trancoso

L^2F Spoken Systems Lab
INESC-ID Lisboa
{wang.ling,tiago.luis,joao.graca,luisa.coheur,imt}@l2f.inesc-id.pt

Abstract

Phrase-based systems deeply depend on the quality of their phrase tables and therefore, the process of phrase extraction is always a fundamental step. In this paper we present a general and extensible phrase extraction algorithm, where we have highlighted several control points. The instantiation of these control points allows the simulation of previous approaches, as in each one of these points different strategies/heuristics can be tested. We show how previous approaches fit in this algorithm, compare several of them and, in addition, we propose alternative heuristics, showing their impact on the final translation results. Considering two different test scenarios from the IWSLT 2010 competition (BTEC, Fr-En and DIALOG, Cn-En), we have obtained an improvement in the results of 2.4 and 2.8 BLEU points, respectively.

1. Introduction

Modern statistical translation models depend crucially on the minimal translation units that are available during decoding. However, the evolution of statistical models – from Word-based (Wb) [1] to Phrase-based (Pb) [2, 3] or Syntax-based (Sb) models [4, 5] – increased the difficulty of understanding what makes a good translation unit, and increased as well their acquisition process. Concerning this acquisition process, the original and widely used approach [3] to aquire minimal units for Pb models consists of a pipeline that starts with a set of word alignments and implements a series of heuristics to build the final phrase table. In recent years, several studies have tried to improve this pipeline. Thus, better alignment models were created [6, 7, 8], better combination of different alignments were tested [9], the posterior distribution over the alignments was used instead of a single best alignment [7, 10], different features were added to the phrase table [11], selective selection of phrases was tested [12, 13] and phrase tables were pruned [14], among others.

In this paper we follow the Pb paradigm and present a general and extensible phrase extraction algorithm. We have highlighted several control points during the creation of phrase tables, that represent fundamental steps in this process. In each one of these points, different strategies/heuristics can be tested with minimal implementation efforts. In this paper, we also show how previous approaches fit in this algorithm and compare several of them. Moreover, we propose different heuristics and show their impact on the final translation results. Our experiments ran on two different test scenarios from the IWSLT 2010 competition (`http://iwslt2010.fbk.eu/`).

This paper is organized as follows: in Section 2 we describe some background, in Section 3 we present the proposed algorithm for phrase extraction and in Section 4 we describe the used corpora. In Section 5 we present and discuss the obtained results and we conclude in Section 6, where we also propose some future work.

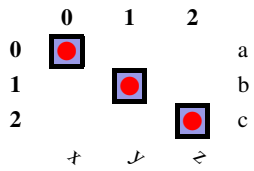
2. Background

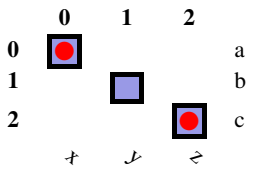
The Pb model is currently the most commonly used and one of the best performing models in Machine Translation (MT). It extends Wb models by translating chunks of words, called phrases, at a time, instead of single words, which leads to several advantages over Wb systems. In fact, Pb models capture both the translation of compound expressions and local reordering (and thus simplify the reordering step). Furthermore, they also simplify the search space.

Pb systems rely on a phrase table, a collection of minimal translation units, with the phrases in the source language and their correspondent translation in the target language. Given a phrase table, the translation process can be broken down into three steps: segment the source sentence into phrases, translate each source phrase into a target phrase, and reorder the target phrases. The quality of Pb systems depend directly on the quality of their phrase-table and therefore, phrase extraction is always a fundamental step in Statistical MT. Extracting all possible phrase-pairs is not a valid option since an exponential number of phrases would be extracted, most of which linguistically irrelevant. Learning a phrase table directly from a bilingual corpus has also been tried previously [15, 16], but these methods failed to compete with heuristic methods that we will describe briefly in the following section. In [16] some problems that result from learning a phrase table directly from data using the EM algorithm are identified. In general terms, phrase pairs with different segmentations (potentially all equally good) compete for probability mass (as opposed to learning bilingual lexicons where the competition is only based on bilingual word pairs).

The most common phrase extraction algorithm [3]

Foreign	Source	Points
x	a	0-0
x y	a b	0-0 1-1
x y z	a b c	0-0 1-1 2-2
y	b	1-1
y z	b c	1-1 2-2
z	c	2-2





Foreign	Source	Points
x	a	0-0
x y	a	0-0
x	a b	0-0
x y	a b	0-0
x y z	a b c	0-0 2-2
z	c	2-2
y z	c	2-2
z	b c	2-2
y z	b c	2-2

Figure 1: Machine Translation phrase extraction from word alignments example.

uses word alignment information to constraint the possible phrases that can be extracted. Given a word alignment, all phrase pairs consistent with that word alignment are extracted from the parallel sentence (a phrase pair is consistent with a word alignment if all words in one language contained in the phrase pair are aligned only to words in the other language which are also contained in the phrase pair). That is to say, all phrase pairs that include at least one aligned point, but do not contradict an alignment by including an aligned word in one language without its translation in the other language, are extracted. So on the one hand if there are too many incorrect alignment points forming a cluster, the correct phrases cannot be extracted without the spurious words, leading to missing words/phrases from the phrase table. In addition, unaligned words act as wild cards that can be aligned to every word in the neighborhood, thus increasing the size of the phrase table. Another undesirable effect of unaligned words is that they will only appear (in the phrase table) in the context of the surrounding words. Moreover, the spurious phrase pairs will change both the phrase probability as well as the lexical weight feature. The work by [17] concludes that the factor with most impact was the degradation of the translation probabilities due to noisy phrase pairs.

Figure 1 shows the phrase tables extracted from two word alignments for the same sentence pair. These alignments only differ in one point: *y-b*. However, the nonexistence of this point in the second word alignment results in the removal of the phrase *y-b* in the second phrase table. Hence we would not be able to translate *y* as *b* except in the contexts shown in that table.

Figure 2 shows an example of a word alignment where a rare source word is aligned to a group of target words, an effect known as garbage collector, which causes that the word *baldwin* cannot be extracted without the incorrect surrounding context. This will make the pair *baldwin, baldwin* unavailable outside the given context.

Most word alignment models are asymmetric in nature only allowing 1-n alignments; moreover, when training the same language pair, switching the source/target languages results in very different alignments. For phrase extraction we

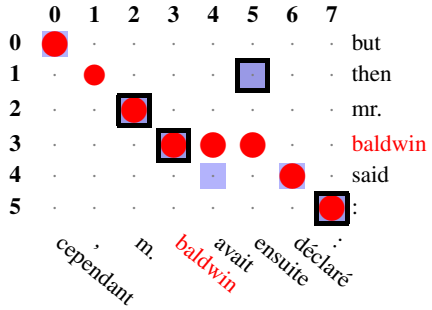


Figure 2: Example of a word alignment suffering from the garbage collector effect.

are interested in a single alignment per sentence so the two directional alignments are combined to form a single alignment. Several approaches have been proposed to symmetrize these word alignments. The most commonly used is called **grow diagonal final** [18]. It starts with the intersection of the sets of aligned points and adds points around the diagonal that are in the union of the two sets of aligned points. The resulting alignment has high recall relative to the intersection and only slightly lower recall than the union. Other common approaches include the **intersection** and **union** of the directional alignments.

However, the exact relationship between word alignment quality and machine translation quality is not straightforward. Despite recent work showing that better word alignments can result in better MT [6, 19], namely by reducing the garbage collector effect and hence increasing the phrase table coverage, there is evidence that the two are only indirectly connected [17, 20]. There are several reasons that explain these facts. First during the pipeline, one keeps committing to the best options available: best word alignment, best symmetrized alignment, etc. Second, the symmetrization heuristics tend to obfuscate the quality of the resulting alignment and the original ones by creating alignments close to the diagonal. Third, current phrase extraction weights each phrase independently of the quality of the underlying align-

ments (all phrases are given the weight of 1).

Several approaches have been proposed to mitigate this problem. **Soft Union** [9] uses the knowledge about the posterior distributions of each directional model. It includes a point in the final alignment if the average of the posteriors under the two models for that point is above a certain threshold. However, this approach still produces a single alignment to be used by the phrase extraction. An alternative is to not commit to any particular alignment, but either use n-best lists [21] or the posterior distribution over the alignments to extract phrases [7, 10]. Both of these approaches increase the coverage of the phrase table. Moreover, by using the posterior weight of each phrase as its score instead of using 1 for all sentences, one can better estimate the phrase probabilities.

3. General Phrase Extraction

In this section we present a general phrase extraction algorithm (described in Algorithm 1). We have highlighted several entry points in this algorithm that when instantiated, implemented the different phrase extraction methods described in the previous section.

Algorithm 1 General Phrase Table Extraction

```
Require: Bilingual Corpus
Require: MaximumPhraseSize - max
  for each sentence pair (s, t) in Corpus do
    extractedPhrasePairs = extractPhrasePairs(s, t, max)
    for each phrase pair p in extractedPhrasePairs do
      phraseTable.add(p)
    end for
  end for
  computeGlobalPhraseStats
  pruneGlobalPhraseStats
  savePhraseTable
```

In general terms, Algorithm 1 iterates through all sentences in the bilingual corpus and for each sentence pair, a set of phrase pairs are extracted, according to algorithm 2. Then all phrase pairs *p* are grouped by the method *phraseTable.add(p)*. Phrase pairs' features are calculated in the method *computeGlobalPhraseStats*, that is then followed by the methods *pruneGlobalPhraseStats* and *savePhraseTable*, which are responsible for pruning (using global statistics from the phrase table, for instance) and saving the phrase table (several saving options can be taken at this point, such as deciding to generate different phrase tables, according with some criteria), respectively.

Different methods for *pruneGlobalPhraseStats* and *savePhraseTable* are out of the scope of this paper. We leave for future work the implementation of alternative methods, such as the one described in the work done by [14], concerning the pruning of the phrase table. Here we will focus on the methods *computeGlobalPhraseStats* and *extractPhrasePairs(s, t, max)*.

The method *computeGlobalPhraseStats* calculates features, such as the phrase translation probability and the lexical probability based on the counts collected for each occurrence of each phrase pair. This is also the point where different types of smoothing for the phrase table can be applied (for instance, we can implement at this step the Knesser-Ney smoothing for bilingual phrases [22]). In this paper, we use as features both the phrase probability and lexical probability in the general phrase extraction [3] [10].

Algorithm 2 Extract Phrase Pairs

```
Require: Bilingual sentence s
  fl = s.foreignLen
  sl = s.sourceLen
  extractedPhrasePairs = {}
  for fp = 0; fp ≤ fl; fp++ do
    for fd = 1; fd ≤ maxDuration; fd++ do
      if ForeignPhraseAcceptor.accept(s, fp, fd) then
        for sp = 0; sp ≤ sl; sp++ do
          for sd = 1; sd ≤ maxDuration; sd++ do
            if SourcePhraseAcceptor.accept(s, sp, sd)
            then
              PhrasePair p = phrase pair from s from (fp,
              sp) to (fd, sd)
              LocalPhrasePairFeaturesCreator.addFeatures(p)
              if PhrasePairAcceptor.accept(p) then
                extractedPhrasePairs.add(p)
              end if
            end if
          end for
        end for
      end if
    end for
  end for
  return extractedPhrasePairs
```

The extraction of phrases is handled in Algorithm 2. This algorithm iterates through all word combinations in a given sentence and extracts a set of phrase pairs. The key control points in this algorithm are *ForeignPhraseAcceptor* and *SourcePhraseAcceptor* which decide if a monolingual phrase makes up for a good translation unit. In this work we accept all phrases, although a methodology for rejecting spurious phrases can be easily implemented.

Moreover, the method *LocalPhrasePairFeaturesCreator* collects a set of features for a given phrase pair and *PhrasePairAcceptor*, based on the collected features, decides if a phrase pair should be accepted or not. This leads us to the main target of this work, which is to compare different acceptors. The baseline system accepts a phrase if it is consistent with a word alignment (we call this the *KoehnAcceptor*) and has a score feature equal to 1 for each extracted phrase, meaning that all phrases are weighted equally. We have also extracted features from the posterior weight of each alignment, as described in [10]. In this work, a weighted align-

Data	Lang.	Sentences	Words	Avg. Length
BTEC				
Train	fr	19972	200614	10.04
	en	19972	189020	9.46
Dev	fr	506	4066	8.04
	en	506	3804	7.52
Test	fr	500	4068	8.14
	en	8000	55017	6.87
DIALOG				
Train	cn	30033	274057	9.13
	en	30033	333400	11.10
Dev	cn	200	2140	10.70
	en	200	2435	12.18
Test	cn	506	3354	6.62
	en	8096	67608	8.35

Table 1: Data statistics of the datasets. There are 85 unknown words in the BTEC test set and 102 unknown words in the DIALOG test set.

ment matrix is used for each phrase pair, with the probabilities of each word in the source to be aligned with each target word, and the phrase is scored based on quality of this alignment matrix. Afterwards, all phrase pairs that score lower than a given score are rejected(the *posteriorAcceptor*). Furthermore, the acceptors and feature extractors can be used together. Using this, we have tested different acceptors based on the existing punctuation inside each phrase, and based on the differences in length of each phrase pair.

4. Corpus

Our experiments were performed over two datasets, the BTEC and DIALOG parallel corpus from the IWSLT 2010 evaluation. The BTEC corpus is a multilingual speech corpus that contains tourism-related sentences, like the ones found in phrasebooks for tourists going abroad. The DIALOG corpus is a collection of human-mediated cross-lingual dialogs in travel situations, that also contains some parts of the BTEC corpus.

The experiments performed with the BTEC corpus were done on the French-English direction, while on the DIALOG corpus were on the Chinese-English direction. The training corpus contains about 19K sentences for the BTEC corpus and 30K for the DIALOG corpus. The development corpus has about 500 sentences for the BTEC corpus and 200 sentences for the DIALOG corpus. For test purposes, we used one of the multiple development corpus provided. All the test corpora were evaluated against multiple references (both BTEC and DIALOG test corpora have a total of 16 different references). Table 1 show present some statistics taken from these corpora.

5. Experimental Results

In this section we compare different methods regarding Algorithm 1. We follow the steps described in the workshop on statistical machine translation (`http://www.statmt.org/wmt09/baseline.html`). Therefore, for both corpora we start by tokenizing and lowercasing them with the provided scripts, and building the corresponding language models. For Chinese, we also replace the punctuation ("。", "，", "？", "！") with the respective latin punctuation. Furthermore, we leave the segmentation of Chinese characters as the one given in the corpus.

Method	Fr-En	Cn-En
Moses IBM M4	61.05	40.29
Moses HMM	59.87	38.54
HMM	59.93	38.49
BHMM	62.45	38.17
SHMM	**62.46**	**41.42**

Table 2: Bleu using the default moses extraction algorithm (one, kohen, moses) for the different alignment models for three different scenarions.

At the end of the pipeline, we detokenize and recase the translation, so that the evaluation is performed according to the IWLST task. The recasing is done using a maximum entropy-based capitalization system [23]. For all experiments we use the Moses decoder (`http://www.statmt.org/moses/`), and before decoding the test set, we tune the weights of the phrase table using Minimum Error Rate Training (MERT). The results are evaluated using the BLEU metric [24].

We also follow the baseline described in the workshop above, which creates directional alignments produced by GIZA++ using the IBM M4 model (the train conditions are the same as the default moses training scripts: 5 iterations of IBM M1, 5 iterations of HMM and 5 iterations of IBM M4). Also, we combine these alignments using the grow-diagonal-final heuristic, and use the default phrase extraction algorithm [3]. We will refer to the baseline as "Moses IBM M4". Moreover, we also compare three different alignment models: the regular HMM [25] (referred to as "HMM"), the same model, but using the posterior regularization framework with bijective constraints (refered to as "BHMM"), and with symmetry constraints (refered to as "SHMM") [19]. The latter constraint takes into account that if a certain unit a is aligned to unit b in the source to target alignment model, b should also be aligned to a in the target to source alignment model. In these alignments we use a threshold of 0.4 for accepting an alignment point. We trained these models using the conditions described in [19]. For each model, we initialize the translation tables with the results produced by IBM M1 with 5 iterations. The "HMM" model was run for 5 iterations, while the "BHMM" and "SHMM" were run for 2 iterations. Both "BHMM" and "SHMM" require two parameters: the constraint set slack and convergence stopping criteria. For both, we used the value 0.001 (we refer the reader for the original paper for an explanation of the meaning of these parameters).

Table 2 compares the different word alignments using the

Method	Fr-En	Cn-En
HMM	59.93	38.49
BHMM	62.45	38.17
SHMM	62.46	41.42
HMM-post	61.74	39.48
BHMM-post	62.74	40.69
SHMM-post	63.07	42.15

Table 3: Different weighting of each phrase using the score from weighted alignment matrix.

Method	Fr-En	Cn-En
HMM	83	105
BHMM	2	7
SHMM	10	11
HMM-post	2	14
BHMM-post	0	6
SHMM-post	3	3

Table 4: Number of unknown words during decoding that exist in the training corpus.

default phrase extraction. As a sanity check, we compare our implementation of the default phrase extraction, with the one provided by moses ("HMM" vs "Moses HMM"), which yielded very close results. The small difference in values is due to an implementation detail difference in the alignments used, when calculating the lexical weighting of phrase pairs that were generated from multiple alignments. When this happens, we need to choose which alignment to use to compute the lexical weighting. In the case of Moses, it picks the most frequent alignment, while we select the alignment from the first phrase pair that is selected. Comparing the results for the different alignment models we see that the constrained-based alignments perform better than the regular HMM and IBM M4. This is not specially surprising since this was observed before, specially under small data conditions [19], as the ones used here.

For the next experiment we use information about the posterior distributions in the alignments. Given the posterior distribution for an alignment link, we use the soft union heuristic (the average of each link) to obtain a symmetrized alignment with link posteriors. Given these alignments links, we calculate the phrase translation probability and the link probability using the approach proposed for weighted alignment matrixes [10]. We only accept a phrase if its phrase posterior probability is above a particular threshold. For both the BTEC and DIALOG corpora we use a threshold of 0.1. We set the values based on the results of the original paper [10] and leave the tuning of this particular threshold as future work, as lowering does not always yields better results.

Table 3 shows the differences between using the default phrase extraction and using information about the alignment posterior. We note that, for every scenario, using the posteriors improves the translation quality over using a single alignment. Table 4 shows the number of words that exist in the training corpus, but the translation system does not know how to translate. For the default heuristic this is mainly due to the garbage collector effect which does not allows a phrase to exist outside the context where it was seen in the training corpus. The constrained alignment models partially solve this problem by correctly dealing with the garbage collector effect. This is further improved since a word pair can be extracted even when it is not consistent with the existing alignments. For instance, when translating the French sentence *"qu'est-ce qu'elle disait?"*, our models created with the default phrase extraction, unlike the ones created with alignment posteriors, do not have any word translation for the word *"disait"*. The phrases extracted with the default extraction method only contain the following right context *"disait six heures"* and, therefore, this context does not allow the translation of the sentence above. The words that are left unknown are due to the threshold being too high.

We also add new features and new acceptors to address some observed problems.

In the first experiment we add a part of speech feature that calculates the phrase probabilities and the lexical probabilities based on the part of speech of each word. We use the unsupervised POS system described by [26] (using the source code available at the authors website), and cluster the words into 50 different groups. We then tag each word with the attained tag. The intuition behind this feature is that a given word can be translated differently if it is being used as a noun or as a verb, and different POS sequences tend to generate different translations even if the words are the same. However, this approach produces worse results than the baseline. Two possible reasons are the use of an unsupervised system, whose accuracy is not very high. Furthermore, this system does not allows word ambiguity (the same word cannot have two different parts of speech). We plan to test this same feature but using a supervised system.

The second test consists of adding an acceptor that discards phrase pairs whose difference in the source and target phrase length are higher than a given threshold. The intuition behind this acceptor is that translations are mostly word by word, specially between English and French. Hence, if there is a huge difference between phrases, this is probably due to the unaligned words, and will lead to a lot of spurious phrases. Furthermore, using this heuristic, we are able to generate smaller phrase tables, raising the translation performance.

Therefore, we perform two tests to discard phrases whose difference is larger than 2 (length-diff2) and 4 (length-diff4). Figure 3 shows the distribution of length difference of the phrase pairs in the phrase table, and Figure 4 shows the distributions of length difference of the phrase pairs actually used by the decoder. Finally, Table 6 shows the percentual reduction in the phrase table size using each heuristic.

As expected, although there are a lot of phrase pairs with large length differences, the decoder only uses 1 sentence

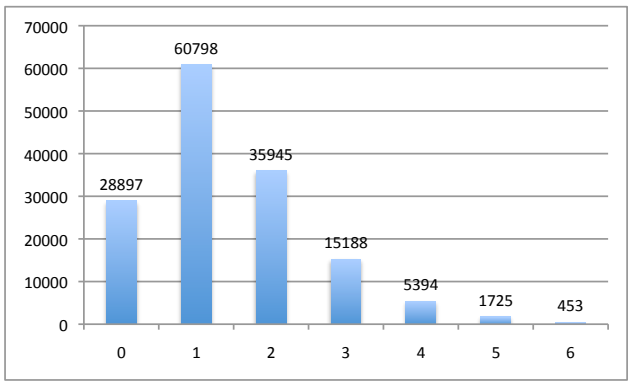


Figure 3: Distribution of phrase-pairs in the phrase table by the difference between lenghts for source and target phrases for the BTEC corpus.

with length difference of 6, as well as 4 phrase-pairs with a difference of 5. However, looking at Table 5 we see that both heuristics (length-diff2 and length-diff4) hurt the performance. Table 7 show the number of times phrase-pairs with a given length difference were used. A first observation is that the unique phrase pair whose length difference was 6 was used almost 300 times which explains why discarding this phrase pair hinders the results. This particular case is the translation of the english word *"could"* by the French sentence *"pourriez-vous , s' il vous plaît ,"*, which is obviously not a good translation, since the correct translation of the French sentence is *"could you, please,"*. However, due to different writing styles used in both languages this translation occurs very often, and when phrase pairs with length larger than 4 are cut of the translation table we could not perform such a translation.

An alternative approach for the same idea is to only remove a phrase pair with large differences in length if it can be re-written by using smaller existing phrases. This approach

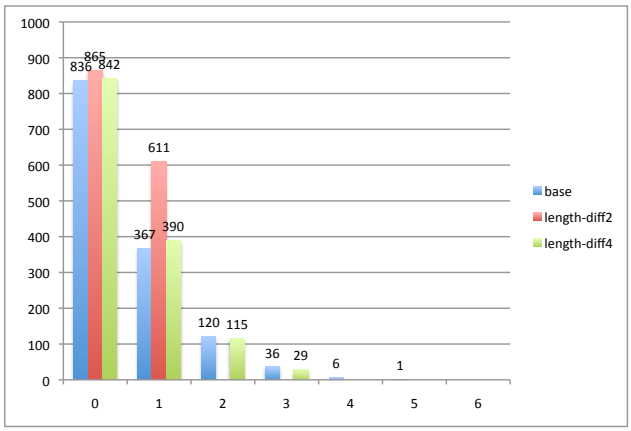


Figure 4: Distribution of unique phrase-pairs used during decoding for the different heuristics for the BTEC corpus.

is similar to the approach proposed in [27] which only includes phrases of length larger than a given threshold if they cannot be realized by using smaller phrases.

In the Chinese test set, we actually obtain better results by discarding phrase pairs with a length difference of 4. The reason for this is that although most phrase pairs with large size difference are not very likely to be the best translations, their probabilities are very high, since the number of occurrences of the larger phrase is usually very small as larger sequences of words tend to occur less. Despite the fact that the lexical weighting helps to a certain extent, in the majority of the cases the system prefers to use phrase pairs with larger difference in characters. By eliminating these sentences, we force the system to use smaller translation units which can be better reordered and which allows more translation options for each sentence. For instance, the sentence 需要预订吗？", which means *"Do I need a reservation?"*, is wrongly translated to *"I need to make a reservation ?"* using the baseline, mainly because "需要预订" is translated to *"I need to make a reservation"*, since it did not have an entry with the question form with that source phrase. On the other hand, with this heuristic the system is able to translate "需要" to *"do I need"* and then, "预订" to *"a reservation"*.

Finally, we built special acceptors that deal with punctuation. The idea is that punctuation is normally translated one to one. Moreover, we observed that spurious punctuation tend to appear in the translation due to incorrect phrases. To this end we tried three different acceptors. Acceptor *no-punct* rejects all phrases that contains punctuation. This is the more radical approach and produces worst results. The reason is that some types of punctuation, commas for instance, are used in different contexts from one language to another. For instance, the sentence *"s'il vous plaît, cherchez mon non encore une fois"*, does not have the comma in the translation *"Please look for my name again"*. By discarding these phrases, these commas could never be translated, hence decreasing performance. The second heuristic *no-terminal-punct* rejects all phrases that contain a terminal punctuation. This heuristic produces small improvement for French to English but not for Chinese to English. This happens since Chinese has a lot of particles such as "吗", "呢", "啊" and "吧", specially in spoken text, which are characters that, are not aligned with any words in English. Since we do not use null translations, these must be aligned with something in the phrase table. In the case of "吗" and "呢", because they are question particles, they tend to appear before question marks, so they are generally aligned like "吗？ " to "?", which would work like a null translation for the particle, but because of our heuristic, we would not allow the resulting phrase pair to be extracted. Thus, when the character "吗" is translated, the chosen translation is picked from phrase pairs that are created from incorrect alignments for the particle. We tried another test by removing the majority of these particles, totalizing 113 particles, and obtained better results.

SHMM-post	Fr-En	Cn-En
base	63.07	42.15
pos	62.22	-
length-diff2	59.26	40.39
length-diff4	62.27	43.06
no-punct	62.75	40.87
no-terminal-punct	63.41	41.44
no-punct(w/o particles)	-	41.20
no-terminal-punct (w/o particles)	-	42.28

Table 5: Experiments with different features and acceptors for phrase extraction.

SHMM-post-lex	Fr-En	Cn-En
base	100%	100%
length-diff2	63%	50%
length-diff4	94%	87%
no-punct	63%	53%
no-terminal-punct	74%	64%

Table 6: Phrase table size comparison.

6. Conclusions And Future Work

In this paper we presented an empirical evaluation of different methods to extract and score minimal translation units for phrase-based translation. We presented a general algorithm for phrase extraction, with well identified control points, and showed how we can replicate several proposed approaches to phrase extraction using this algorithm. The framework is easily extensible and, by providing it to the community as a replacement for the moses training scripts, we hope that more people will devote some attention to the extraction of minimal units, a crucial step in statistical machine translation. The code and scripts used on this paper will be made available and can be used as a direct replacement of the moses training scripts.

Different types of word alignment algorithms were compared and we showed how they affect the performance of the end translation. We also showed that the results attained by running these different algorithms are even better when the alignments probabilities are used to extract and weight the respective phrases. We have described some simple extraction heuristics that combined with the existing ones lead to an overall improvement of 2.4 BLEU point on the BTEC English to French translation task and 2.8 BLEU points improvement over the DIALOG English to Chinese translation task.

As future work we intend to pursue the study of the effect of phrase extraction on different corpora and different corpora sizes, because these differences become more apparent as the size of the corpus increases. Moreover, we will try different features and selectors based on other linguistic resources. Finally, we think that a similar idea can be used for the extraction of hierarchical phrases and syntax rules.

	0	1	2	3	4	5
BTEC						
base	1183	387	138	772	7	287
length-diff2	1304	731	0	0	0	0
length-diff4	1281	414	138	733	0	0
DIALOG						
base	2063	1130	247	107	4	0
length-diff2	1194	1416	0	0	0	0
length-diff4	1985	1257	276	161	0	0

Table 7: Translation units used during decoding for the different heuristics.

7. Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Tiago Luís is supported by FCT grant SFRH/BD/62151/2009. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.

8. References

[1] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[2] F. Och and H. Ney, "The Alignment Template Approach to Statistical Machine Translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-based Translation," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.

[4] K. Yamada and K. Knight, "Syntax-based Statistical Translation Model," July 3 2002, uS Patent App. 10/190,298.

[5] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a Translation Rule," in *Proceedings of HLT/NAACL*, vol. 4, 2004, pp. 273–280.

[6] K. Ganchev, J. V. Graça, and B. Taskar, "Better Alignments = Better Translations?" in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 986–993.

[7] Y. Deng and W. Byrne, "HMM word and phrase alignment for statistical machine translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 494–507, 2008.

[8] A. Fraser and D. Marcu, "Getting the Structure Right for Word Alignment: LEAF," in *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, June 2007, pp. 51–60.

[9] J. DeNero and D. Klein, "Tailoring Word Alignments to Syntactic Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 17–24.

[10] Y. Liu, T. Xia, X. Xiao, and Q. Liu, "Weighted Alignment Matrices for Statistical Machine Translation," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 1017–1026.

[11] M. Costa-jussà and J. Fonollosa, "Improving Phrase-based Statistical Translation by Modifying Phrase Extraction and Including Several Features," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics, 2005, pp. 149–154.

[12] L. Zettlemoyer and R. Moore, "Selective Phrase Pair Extraction for Improved Statistical Machine Translation," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers on XX*. Association for Computational Linguistics, 2007, pp. 209–212.

[13] Y. Deng, J. Xu, and Y. Gao, "Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair," *Proceedings of ACL-08: HLT*, pp. 81–88, 2008.

[14] H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving Translation Quality by Discarding Most of the Phrasetable," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 967–975.

[15] D. Marcu and W. Wong, "A Phrase-based, Joint Probability Model for Statistical Machine Translation," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, p. 139.

[16] J. DeNero, D. Gillick, J. Zhang, and D. Klein, "Why generative phrase models underperform surface heuristics," in *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2006, pp. 31–38.

[17] A. Lopez and P. Resnik, "Word-based alignment, phrase-based translation: What´s the link?" in *Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA): visions for the future of machine translation*, Boston, MA, 2006, pp. 90–99.

[18] A. Axelrod, R. B. Mayne, C. Callison-burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 iwslt speech translation evaluation," in *In Proc. International Workshop on Spoken Language Translation (IWSLT*, 2005.

[19] J. Graça, K. Ganchev, and B. Taskar, "Learning Tractable Word Alignment Models with Complex Constraints," *Comput. Linguist.*, vol. 36, pp. 481–504.

[20] D. Vilar, M. Popović, and H. Ney, "Aer: Do we need to "improve" our alignments?" in *Proc. IWSLT*, 2006.

[21] A. Venugopal, A. Zollmann, N. Smith, and S. Vogel, "Wider pipelines: N-best alignments and parses in mt training," in *Proceedings of AMTA*, 2008.

[22] G. Foster, R. Kuhn, and H. Johnson, "Phrasetable smoothing for statistical machine translation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 53–61.

[23] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[24] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[25] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Translation," in *Proceedings of the 16th conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1996, pp. 836–841.

[26] A. Clark, "Combining distributional and morphological information for part of speech induction," in *Proc. EACL*, 2003.

[27] J. Marino, R. Banchs, J. Crego, A. de Gispert, P. Lambert, J. Fonollosa, and M. Ruiz, "Bilingual n-gram statistical machine translation," *Proc. of Machine Translation Summit X*, pp. 275–82, 2005.