# UPC-BMIC-VDU system description for the IWSLT 2010: testing several collocation segmentations in a phrase-based SMT system

Carlos A. Henríquez Q.*, Marta R. Costa-jussà†, Vidas Daudaravicius‡, Rafael E.Banchs^æ, José B. Mariño*

*TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
{carlos.henriquez,jose.marino}@upc.edu
†Barcelona Media Innovation Center, Barcelona
marta.ruiz@barcelonamedia.org
‡Vytautas Magnus University, Kaunas
vidas@donelaitis.vdu.lt
^æInstitute for Infocomm Research, Singapore
rembanchs@i2r.a-star.edu.sg

## Abstract

This paper describes the UPC-BMIC-VMU participation in the IWSLT 2010 evaluation campaign. The SMT system is a standard phrase-based enriched with novel segmentations. These novel segmentations are computed using statistical measures such as Log-likelihood, T-score, Chi-squared, Dice, Mutual Information or Gravity-Counts.

The analysis of translation results allows to divide measures into three groups. First, Log-likelihood, Chi-squared and T-score tend to combine high frequency words and collocation segments are very short. They improve the SMT system by adding new translation units. Second, Mutual Information and Dice tend to combine low frequency words and collocation segments are short. They improve the SMT system by smoothing the translation units. And third, Gravity-Counts tends to combine high and low frequency words and collocation segments are long. However, in this case, the SMT system is not improved.

Thus, the road-map for translation system improvement is to introduce new phrases with either low frequency or high frequency words. It is hard to introduce new phrases with low and high frequency words in order to improve translation quality. Experimental results are reported in the French-to-English IWSLT 2010 evaluation where our system was ranked 3rd out of nine systems.

## 1. Introduction

The Universitat Politècnica de Catalunya (UPC), Barcelona Media Innovation Center (BMIC) and Vytautas Magnus University (VMU) participated together in the IWSLT 2010 evaluation campaign. This paper describes the UPC-BMIC-VMU system, which is basically a statistical phrase-based system enriched with collocation segmentation information. Adding a novel segmentation in an SMT system allows to enrich the translation dictionary and/or to smooth the existing translation probabilities. Basically, we extend the work presented in the WMT 2010 evaluation for Spanish-to-English [7] by experimenting with different statistical scores to segment a monolingual training corpus and by analysing if it is better to add new translation phrases and/or to smooth the existing ones. We participated in the French-to-English BTEC task. Our primary and contrastive systems were two standard phrase-based SMT systems enriched with different novel segmentations.

This paper is organized as follows. Section 2 makes a brief description of some related work to the introduction of new segmentations in SMT. Section 3 describes the baseline system. Then, Section 4 reports different statistical criteria to segment monolingual data and Section 5 shows how to introduce the segmented data into the phrase-based system. As follows, Section 6 shows the experimental details of the system and the experiments performed with the novel technique. Finally, Section 7 presents the conclusions.

## 2. Related work

One of the main problems in the statistical machine translation approach is how to segment the bilingual corpus in order to build the most appropriate translation dictionary. Standard phrase-based SMT systems first align the parallel corpus at the word level by using IBM probabilities and then use standard constraints (see section 3) to extract final translation units [10]. Variations of this type of segmentation can be found in [8, 1, 6]. Other approaches consist in integrating the phrase segmentation and alignment, one example is in [14] where they use the point-wise mutual information between the source and target words to identify aligned phrase pairs. In [9] they use a greedy algorithm to compute recursive alignments from a bilingual parallel corpus.

Here, we propose to combine the standard phrase-based segmentation [10] with a complementary bilingual segmentation which is learned from a statistical collocation segmentation technique. This statistical collocation segmentation uses measures such as dice score to estimate segments of words. The benefit from this procedure is twofold: (1) it extracts new translation units and (2) it smooths the probability of existent translation units.

## 3. Phrase-based Baseline System

The basic idea of phrase-based translation is to segment the given source sentence into units (hereafter called phrases), then translate each phrase and finally compose the target sentence from these phrase translations.

Basically, a bilingual phrase is a pair of $m$ source words and $n$ target words. For extraction from a bilingual word aligned training corpus, two additional constraints are considered:

1. the words are consecutive, and,
2. they are consistent with the word alignment matrix.

Given the collected phrase pairs, the phrase translation probability distribution is commonly estimated by relative frequency in both directions.

The translation model is combined together with the following six additional feature models: the target language model, the word and the phrase bonus and the source-to-target and target-to-source lexicon model and the reordering model. These models are optimally weighted for decoding.

## 4. Collocation segmentation

A collocation segment is a piece of text between collocation segment boundaries. The collocation segmentation detects the boundaries of collocation segments within a text. First of all, for each paragraph in a text we calculate associativity values between adjacent tokens. Associativity values between the beginning and the first token and between the last token and the end of a paragraph are calculated also, i.e. the beginning and the end of a paragraph are treated as specific tokens. After this step, the sequences of associativity values are produced. They are long as much as the paragraphs plus one (minus one token and plus two for the beginning and the end). Next, the boundaries are set in two steps. At the beginning, the boundaries are set between two adjacent tokens when the associativity value is lower than a threshold. The threshold value is set for each paragraph separately. A threshold value for each paragraph or sequence of associativity values between adjacent tokens $\overline{w}$ is calculated by following formula:

$$threshold(\overline{w}) = avg(\overline{w}) - 0.95 * (avg(\overline{w}) - min(\overline{w}))$$

If a paragraph contains only two words then the boundary between these two tokens heavily depends on the associativity values between the beginning of a sentence and the first token and between the second token and the end of the paragraph. Such a dynamic assignment is necessary to produce similar threshold definition conditions for different associativity measures. Different associativity measures have different scale of values and it is difficult to set threshold manually. Threshold level is kept as low as possible. Higher threshold value makes shorter collocation segments and vice versa. Shorter collocation segments are more confident collocations and we may expect better translation results. Nevertheless, the results of [3] show that longer collocation segments are more preferable.

There are many associativity measures that could be used to calculate the associativity values between tokens (a more comprehensive list could be found in [11]). To explore different measures we included the six following metrics:

1. Mutual Information (MI):

$$MI(w_i, w_{i+1}) = \frac{N * f(w_i, w_{i+1})}{f(w_1) + f(w_{i+1})}$$

2. Dice:

$$dice(w_i, w_{i+1}) = \frac{2 * f(w_i, w_{i+1})}{f(w_i) + f(w_{i+1})}$$

3. Log-likelihood[1]:

$$likelihood_{Dunning}(w_i, w_{i+1}) = \\ = \begin{cases} 0 & : \; f(w) = f(w_i, w_{i+1}) = 1 \\ 0 & : \; f(w_{i+1}) = f(w_i, w_{i+1}) = 1 \\ L * R & : \; otherwise \end{cases}$$

, where

$$\begin{aligned} L \;=\; & 2 * ((f(w_i) - f(w_i, w_{i+1})) * \\ & log\left(\frac{f(w_i) - f(w_i, w_{i+1})}{N - f(w_i, w_{i+1})}\right) - \\ & f(w_i) * log\left(\frac{f(w_i)}{N}\right) + \\ & log\,((f(w_i, w_{i+1}) - f(w_i)) * \\ & log\,\left(\frac{f(w_i, w_{i+1})}{f(w_i)}\right)\right)\right) \end{aligned}$$

$$\begin{aligned} R \;=\; & 2 * ((f(w_{i+1}) - f(w_i, w_{i+1})) * \\ & log\left(\frac{f(w_{i+1}) - f(w_i, w_{i+1})}{N - f(w_i, w_{i+1})}\right) - \\ & f(w_{i+1}) * log\left(\frac{f(w_{i+1})}{N}\right) + \\ & log\,((f(w_i, w_{i+1}) - f(w_{i+1})) * \\ & log\,\left(\frac{f(w_i, w_{i+1})}{f(w_{i+1})}\right)\right)\right) \end{aligned}$$

[1] Log-likelihood evaluates the associativity strength from $w_i$ to $w_{i+1}$ only. Log-likelihood formula is modified to be symmetrical (strength from $w_i$ to $w_{i+1}$ multiplied by strength from $w_{i+1}$ to $w_i$). Original formula contains L part only.

4. Chi-squared (Chi2):

$$chi2(w_i, w_{i+1}) = \\ = \frac{N}{f(w_i) * f(w_{i+1})} \\ * \frac{N * f(w_i, w_{i+1}) - f(w_i) * f(w_{i+1})}{N - f(w_i) + f(w_i, w_{i+1})} \\ * \frac{N * f(w_i, w_{i+1}) - f(w_i) * f(w_{i+1})}{N - f(w_{i+1}) + f(w_i, w_{i+1})}$$

5. Gravity-Counts (GC)[5]:

$$gc(w_i, w_{i+1}) = \\ = log\left(\frac{f(w_i) * f(w_i, w_{i+1})}{n(w_i)}\right) \\ + log\left(\frac{f(w_{i+1}) * f(w_i, w_{i+1})}{n'(w_{i+1})}\right)$$

6. T-score:

$$tscore(w_i, w_{i+1}) = \frac{f(w_i, w_{i+1}) - \frac{f(w_i)f(w_{i+1})}{N}}{\sqrt{f(w_i, w_{i+1})}}$$

The next step after setting the associativity threshold boundaries is to apply an average minimum law (AML) as described in [3] and [4]. The average minimum law is applied to the three adjacent associativity values (i.e., four tokens). The boundary of a segment is set between adjacent tokens when the value of associativity between these two adjacent tokens is lower than the average of preceding and following associativity values. Some examples of segmentation of English and French sentences are presented in Table 1.

The result of collocation segmentation is a segmented text, no dictionaries are produces and no evaluation of segments is made. The segmented text could be used to create a dictionary of collocations. Such dictionary accepts all collocation segments. The main difference from Choueka [2] and Smadja[12] methods is that collocation segmentation accepts all collocations and no significance tests for collocations are performed. The main advantage of this segmentation is the ability to perform collocation segmentation of both small and large corpora, and no manually segmented corpora or other databases and language processing tools are required.

# 5. Introducing the collocation segmentation into a phrase-based system

In order to build the augmented phrase table with the technique mentioned in section 4, we segmented each language of the bilingual corpus independently and then, using the collocation segments as words, we aligned the corpus and extracted the phrases from it. Once the phrases were extracted, the segments of each phrase were split again in words to have standard phrases. Finally, we use the union of these phrases and the phrases extracted from the baseline system to compute the final phrase table. A diagram of the whole procedure can be seen in figure 1.

The objective of this integration is to add new phrases in the translation table and to enhance the relative frequency of the phrases that were extracted from both methods (hereinafter, collocation segmentation *both*). In order to analyse separately the improvement of each, we differentiate from new phrases (marked with $**$ in the figure) and phrases which do already exist in the baseline segmentation. Then, we integrate the baseline segmentation with the new phrases (hereinafter, collocation segmentation *new phrases*) or the baseline segmentation with the existing phrases (hereinafter, collocation segmentation *smooth).*

# 6. Experiments

We participated in the French-to-English BTEC task [13] in the correct recognition results. We build our baseline system using MOSES with the standard configuration *http://www.statmt.org/moses/.*

## 6.1. Data

We used the BTEC corpus provided in the evaluation without using out-of-domain additional corpus. The model weights were tuned with the development corpus named 1 (16 reference translations) and the development corpus named 3 was chosen as internal test set (16 reference translations), according to which we make a decision about better or worse system performance. All 16 references from the development corpus named 2 were added to the language model during tuning. The weights obtained in the optimization were used as well for the evaluation test. For translating the official test sets, we concatenated the training, development and test sets from Table 2 and we used the concatenation as training data. Because the three development sets included sixteen source sides and sixteen references for each sentence, we paired them one-to-one according with their ids in order to build a bigger parallel corpus before concatenating it with the training corpus. This full devset is also mentioned in Table 2. Finally, we also added all references from the three development sets to the language model corpus.

To build and tune the translation systems, we lowercased and tokenized all data using the standard Moses' tools. Once we get the final translation output, we recased and detokenized again following the standard Moses' procedure. A detailed explanation of this preprocess and postprocess can be found in parts III and VII of Moses' Step-by-Step Guide[2]

## 6.2. Automatic translation results

Here, we report the internal experimentation using different segmentations to enrich the phrase table. Using the approach described in section 5, we propose to study the effect of adding new translation units (new phrases), the effect of smoothing the translation units from the standard phrase-based system (smooth) and the effect of adding and smooth-

[2] http://www.statmt.org/moses_steps.html

| Measure | Collocation segmentation example |
|---|---|
| chi2 | they were listening to his speech with open ears . |
| likelihood | they were listening to his speech with open ears . |
| dice | they_were_listening to his_speech with open_ears . |
| MI | they_were_listening_to his_speech_with open_ears . |
| t-score | they_were listening_to his_speech_with open ears . |
| GC | they_were listening_to his_speech_with_open ears_. |
| chi2 | ils écoutaient attentivement son discours . |
| likelihood | ils écoutaient attentivement son discours . |
| dice | ils_écoutaient_attentivement son_discours . |
| MI | ils_écoutaient_attentivement son_discours . |
| t-score | ils écoutaient attentivement son discours . |
| GC | ils écoutaient_attentivement son_discours . |
| chi2 | could_you show me what places are worth seeing near here , please ? |
| likelihood | could_you show me what places are worth seeing near here ,_please ? |
| dice | could_you show_me what places are worth_seeing near_here ,_please ? |
| MI | could_you show_me what_places are_worth_seeing near_here , please ? |
| t-score | could_you show_me what_places are worth_seeing near_here ,_please ? |
| GC | could_you show_me what_places are_worth_seeing near_here ,_please ? |
| chi2 | pourriez-vous me montrer les endroits qui valent la peine d' être vus dans le voisinage , s'_il_vous_plaît ? |
| likelihood | pourriez-vous me montrer les endroits qui valent la peine d' être vus dans le voisinage , s' il vous plaît ? |
| dice | pourriez-vous_me_montrer les_endroits qui valent la peine d'_être vus dans_le voisinage ,_s'_il vous_plaît ? |
| MI | pourriez-vous_me_montrer les_endroits qui_valent la peine_d' être_vus dans le_voisinage , s'_il vous_plaît ? |
| t-score | pourriez-vous_me montrer les_endroits qui valent la peine d'_être vus dans_le_voisinage ,_s'_il vous_plaît ? |
| GC | pourriez-vous_me_montrer les_endroits qui valent la peine d'_être vus dans_le_voisinage ,_s' il_vous plaît_? |

Table 1: *Segmentation examples*

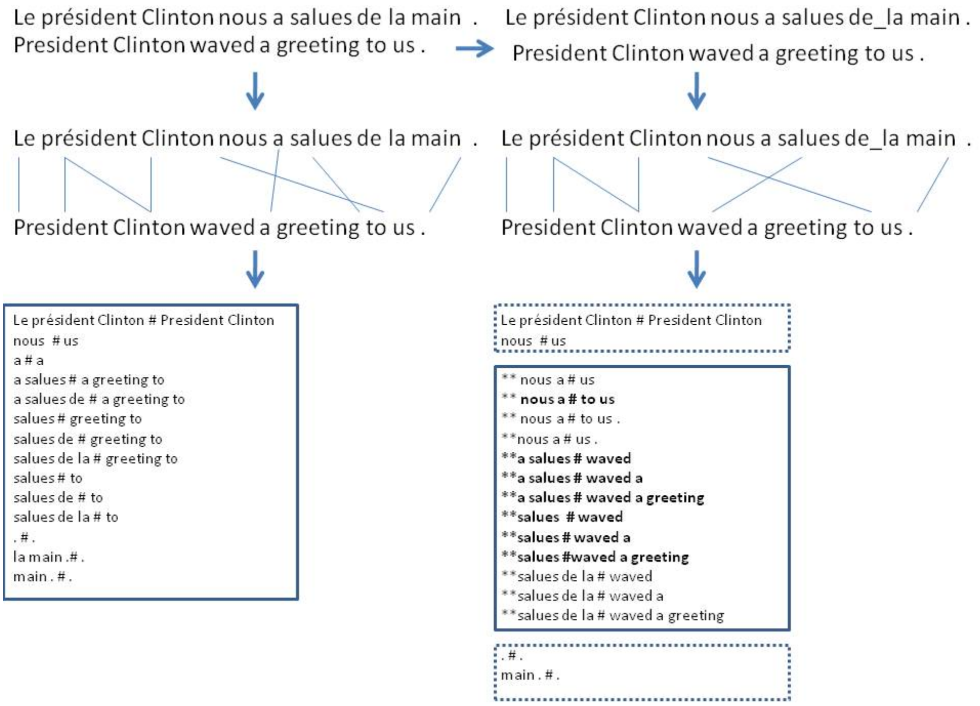
Le président Clinton nous a salues de la main .
President Clinton waved a greeting to us .
Le président Clinton nous a salues de_la main .
President Clinton waved a greeting to us .
Le président Clinton nous a salues de la main .
President Clinton waved a greeting to us .
Le président Clinton nous a salues de_la main .
President Clinton waved a greeting to us .
Le président Clinton # President Clinton
nous # us
a # a
a salues # a greeting to
a salues de # a greeting to
salues # greeting to
salues de # greeting to
salues de la # greeting to
salues # to
salues de # to
salues de la # to
. # .
la main .# .
main . # .
Le président Clinton # President Clinton
nous # us
** nous a # us
** nous a # to us
** nous a # to us .
**nous a # us .
**a salues # waved
**a salues # waved a
**a salues # waved a greeting
**salues # waved
**salues # waved a
**salues #waved a greeting
**salues de la # waved
**salues de la # waved a
**salues de la # waved a greeting
. # .
main . # .

Figure 1: Example of the expansion of the phrase table using collocation segmentation (in this particular case, likelihood). New phrases added by the collocation-based system are marked with a $**$. Most interesting new phrases are in bold. The union of both extractions is used to compute the final phrase table.

|  |  | Fr | En |
|---|---|---|---|
| Training | Num sentences | 19,972 |  |
|  | Words | 189k | 182k |
|  | Vocabulary | 10.7k | 8.3k |
| Development<br>(devset1) | Num sentences | 506 |  |
|  | Words | 3.8k |  |
|  | Vocabulary | 1077 |  |
| Test<br>(devset3) | Num sentences | 506 |  |
|  | Words | 3.8k |  |
|  | Vocabulary | 1103 |  |
| Evaluation<br>2010 testset | Num sentences | 464 |  |
|  | Words | 3.6k |  |
|  | Vocabulary | 1005 |  |
| Evaluation<br>2009 testset | Num sentences | 469 |  |
|  | Words | 3.6k |  |
|  | Vocabulary | 977 |  |
| devset2<br>Include in LM<br>during tuning | Num sentences |  | 8000 |
|  | Words |  | 55k |
|  | Vocabulary |  | 3.8k |
| All devsets<br>Pairing 16 sources<br>with 16 refs | Num sentences | 24,192 |  |
|  | Words | 171k | 166k |
|  | Vocabulary | 10.6k | 7.6k |

Table 2: *Corpus statistics*

ing (both). Table 3 show the results in the internal test set. We can divide the statistical measures in three groups:

1. Dice and Mutual Information
2. Chi-squared, Log-likelihood and T-Score
3. Gravity-Counts

This division is based on the fact that the first group outperforms the baseline system when adding new phrases, the second group outperforms the baseline system when smoothing the existing phrases and the third does not achieve the baseline results.

Initially, we did not expect any improvement with Chi-squared or Log-likelihood segmentation. For instance, Log-likelihood segmentations introduced very short lists of collocations (105 for English and 94 for French):

do you = 1299 / could you = 1237 / it 's = 1163 / 'd like = 1013 / would you = 963 / is the = 957 / 'd like to = 901 / how much = 642 / is it = 631 / , please = 621 / want to = 605 / to the = 589 / is there = 582 / do you have = 570 / tell me = 539 / in the = 511 / give me = 432 / like to = 378 / can you = 196 / you have = 196 / on the = 163 / to get = 158 / would like to = 77 / what 's = 76 / 's the = 72 / this is = 71 / at the = 52 / that 's = 47 / would like = 42 / thank you = 33 / would you have = 27 / for me = 26 / of the = 26 / want to get = 17 / it 's the = 14 / this is the = 13 / what 's the = 12 / are you = 10 / could you have = 7 / i 'd like = 7 / can 't = 6 / i 'd like to = 6 / what time = 6 / that 's the = 5 / want to go = 5 / where is = 5 / i 'd = 4 / to go = 4 / going to = 3 / i 'm = 3 / would you like = 3 / how many = 2 / like to get = 2 / no , = 2 / the number = 2 / where is the = 2 / you tell me = 2 / 're not = 1 / 's that ? i = 1 / 've got = 1 / , but = 1 / . what = 1 / Good morning = 1 / a nice = 1 / an account = 1 / an open = 1 / can you have = 1 / can you tell = 1 / could you tell = 1 / fifty minutes = 1 / five six seven = 1 / for the = 1 / get the = 1 / go to = 1 / have to = 1 / if you = 1 / in another = 1 / is this = 1 / it 's too = 1 / let 's = 1 / let me = 1 / like to go = 1 / like to the = 1 / long time = 1 / name is = 1 / no see = 1 / of cigarettes = 1 / that 's a = 1 / that is = 1 / the party = 1 / there 's no = 1 / this number = 1 / this train = 1 / those are = 1 / to go to = 1 / to leave = 1 / to take = 1 / two three four = 1 / we have = 1 / with us = 1 / would like to go = 1 / would you mind = 1 / would you tell = 1 / you can = 1 / you tell = 1

vous plaît = 3547 / s' il = 3036 / c' est = 2285 / j' ai = 1586 / est-ce que = 1534 / je voudrais = 1395 / je suis = 887 / j' aimerais = 822 / je peux = 783 / à l' = 754 / de la = 704 / , s' il = 603 / est-ce que vous = 593 / pourriez-vous me = 541 / quelque chose = 522 / est le = 508 / je ne = 488 / que vous = 400 / quelle heure = 347 / je veux = 336 / qu' il = 275 / est-ce que je = 273 / je vais = 242 / que je = 240 / à la = 229 / se trouve = 170 / je n' = 118 / je vous = 111 / de l' = 104 / est-ce qu' = 92 / au japon = 73 / c' est le = 71 / vous avez = 58 / que je veux = 35 / combien de = 34 / il vous = 30 / pourriez-vous me dire = 29 / que je ne = 28 / me dire = 24 / est-ce qu' il = 21 / n' est = 18 / que je suis = 18 / que je vais = 16 / que vous avez = 12 / que je peux = 10 / je n' ai = 7 / que je n' = 6 / que je voudrais = 6 / que je vous = 6 / quel est le = 6 / une chambre = 5 / qu' il vous = 4 / un peu = 4 / dans le = 3 / il vous plaît = 3 / je vous prie = 3 / n' est pas = 3 / quel est = 3 / de temps = 2 / est la = 2 / l' hôtel = 2 / où se trouve = 2 / pouvez-vous me = 2 / , merci = 1 / , s' il vous = 1 / . merci = 1 / belle boutique = 1 / bonjour . = 1 / c' est un = 1 / ce serait = 1 / dans cette = 1 / de le faire = 1 / depuis le temps = 1 / elle est = 1 / l' occasion = 1 / n' ai = 1 / non , je ne sais = 1 / nous avons = 1 / pas d' = 1 / pas de = 1 / pour le = 1 / pouvez-vous me dire = 1 / qu' il est = 1 / que c' est = 1 / que ce film = 1 / que j' = 1 / s' est pas = 1 / s' il vous = 1 / tout simplement = 1 / trop lourd = 1 / vous plaît . = 1 / à l' hôtel = 1 / ça fait = 1 / ça pèse = 1

A further analysis of segmentation phrase lists show that Chi-squared, Log-likelihood and T-score segmentations try to keep together high frequency words. Thus, phrases are very short. In opposite, Mutual Information and Dice try to keep together low frequency words and phrases are short enough on the average. Gravity-Counts try to keep together low and high frequency words that leads to long phrases. Thus, Mutual Information and Dice are good to capture collocations with low frequency words. This is acceptable in many situations from a lexical point of view. T-score, Log-likelihood and Chi-squared are good to capture collocations with high frequency words. Gravity counts takes care of low and high frequency words. So, the trend to get the best translation quality is to take care of either low frequency words or of high frequency words. This outcome is acceptable at least for small corpora. The main outcomes are:

1. It is important to keep very frequent words together before alignment. This allows to introduce new good translations.
2. The low frequency words allow to make the corrections to the probabilities of translations and do not introduce new good translation phrases.
3. It is very difficult to introduce good new entries and make smoothing of probabilities at the same time. Either smoothing or introduction of new translation phrases allow to achieve the best translation results, but not together.

The last row of Table 3 shows one last "both" translation system built with the smooth phrases extracted with the Dice collocation and the new phrases extracted with the Log-likelihood strategy. It can be seen that even though the internal test also outperforms the baseline system, it did not

| System | Internal |
|---|---|
| baseline | 60.88 |
| +dice smooth | **61.21** |
| +dice new phrases | 60.23 |
| +dice both | 60.28 |
| +mi smooth | **60.93** |
| +mi new phrases | 59.79 |
| +mi both | 60.10 |
| +chi2 smooth | 60.55 |
| +chi2 new phrases | **61.09** |
| +chi2 both | **61.11** |
| +likelihood smooth | **60.97** |
| +likelihood new phrases | **61.23** |
| +likelihood both | 60.61 |
| +t-score smooth | 60.79 |
| +t-score new phrases | **61.19** |
| +t-score both | **61.08** |
| +gc smooth | 60.58 |
| +gc new phrases | 60.47 |
| +gc both | 60.49 |
| +dice smooth +likelihood new phrases | **61.11** |

Table 3: *Translation results in terms of BLEU for the internal test set. Cases that ouperform the baseline system are in bold.*

| System | 2009 | 2010 |
|---|---|---|
| baseline | 60.93 | 52.61 |
| +likelihood new phrases | **62.00** | **53.27** |
| +dice smooth | 60.13 | **53.24** |

Table 4: *Translation results in terms of BLEU for the evaluation sets. Cases that ouperform the baseline system are in bold.*

achieved the level of "+dice smooth" nor "+likelihood new phrases".

Table 4 shows the results for the primary and contrastive systems presented in the evaluation compared to the baseline system. Our primary system was the "+likelihood new phrases" and the contrastive system was the "+dice smooth". Results over the 2010 test set are coherent with Table 3, nevertheless we are planning to study the difference between both test sets in order to explain why the contrastive system performed so poorly with the 2009 test set.

### 6.3. Manual analysis

We chose 100 random sentences from the evaluation set, and compared the performance of the baseline system against dice-smooth and likelihood-new-phrases approaches. We have observed that the new proposals are better or equal than the baseline. The main improvements are due to:

1. Better selection of translation units, which implies a better semantic preservation. For example: *My main matter is right* (baseline translation), and *My main matter is law* (dice-smooth and likelihood-new-phrases).

2. Better grammatical preservation. For example: *Can I bring a drink* (baseline translation), and *May I bring you a drink?* (dice-smooth and likelihood-new-phrases).

3. Better word order. For example: *How was on the paquebot life?* (baseline translation), and *How was life on the paquebot?* (dice-smooth and likelihood-new-phrases).

In the manual analysis we see that likelihood-new-phrases is making an indirect smoothing because adding new phrases affect the existing ones at least when hypotheses compete in decoding.

Additionally, the likelihood-new-phrases is able to reduce the number of unknown words. For example: *you should méfier of these people* (baseline translation), and *You should be careful of these people* (likelihood new phrases translation).

## 7. Conclusion

This paper describes the UPC-BMIC-VDU system for the French-to-English IWSLT 2010 task. The main contribution is the introduction of different collocation segmentations to enhance the phrase-based system. We have analysed whether the collocation segmentations benefit came from smoothing the existing baseline phrases or introducing new phrases. We can conclude that segmentations like Dice and Mutual Information help smoothing the existing baseline phrases and segmentations like Chi-squared, Log-likelihood and T-score help introducing new phrases. We evaluated the best proposed systems in 3 test sets and we obtained coherent improvements in all of them.

## 8. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762 (FAUST project, FP7-ICT-2009-4-247762), from the Spanish Ministry of Science and Innovation through the BUCEADOR project (TEC2009-14094-C04-01) and the Juan de la Cierva fellowship program. The authors also wants to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

## 9. References

[1] F. Casacuberta. Finite-state transducers for speech input translation. In *Proc. IEEE ASRU*, Madonna di Campiglio, Italy, 2001.

[2] Y. Choueka. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO*, pages 609–624, 1988.

[3] M.R. Costa-jussà, V. Daudaravicius, and R. E. Banchs. Integration of statistical collocation segmentations in a phrase-based statistical machine translation system. In *14th Annual meeting of the EAMT: European Association for Machine Translation*, Saint-Raphael, 2010.

[4] V. Daudaravicius. The influence of collocation segmentation and top 10 items to keyword assignment performance. In *11th International Conference on Intelligent Text Processing and Computational Linguistics, Springer Verlag, LNCS*, pages 648–660, Iasi, Romania, 2010.

[5] V. Daudaravicius and R Marcinkeviciene. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2):321–348, 2004.

[6] I. García-Varea, D. Ortiz, F. Nevado, P.A. Gómez, and F. Casacuberta. Automatic segmentation of bilingual corpora: A comparison of different techniques. In *Iberian Conference on Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 614–621. Springer-Verlag, Estoril (Portugal), June 2005.

[7] C. A. Henríquez Q., M. R. Costa-jussà, V. Daudaravicius, Rafael E. Banchs, and J. B. Mariño. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 104–108, Uppsala, Sweden, July 2010.

[8] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A.R. Fonollosa, and M. R. Costa-jussà. N-gram based machine translation. *Computational linguistics*, 32(4):527–549, 2006.

[9] F. Nevado and F. Casacuberta. Bilingual corpora segmentation using bilingual recursive alignments. In *Proceedings of the 3th Jornadas en Tecnología del Habla*, Valencia, 2004.

[10] F.J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449, December 2004.

[11] P. Pecina and P. Schlesinger. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[12] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.

[13] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceeding of LREC-2002: Third International Conference on Language Resources and Evaluation*, pages 147–152, Las Palmas, Spain, May 2002.

[14] Y. Zhang, S. Vogel, and A. Waibel. Automatic segmentation of bilingual corpora: A comparison of different techniques. In *Natural Language Processing and Knowledge Engineering, 2003. Proceedings*, pages 567–573. 2003.