# UPC-BMIC-VDU system description: testing several collocation segmentations in a PBSMT system*


Carlos A. Henríquez Q carlos.henriquez@upc.edu
Marta R. Costa-jussà marta.ruiz@barcelonamedia.org
Vidas Daudaravicius vidas@donelaitis.vdu.lt
Rafael E. Banchs rembanchs@i2r.a-star.edu.sg
José B. Mariño jose.marino@upc.edu



## Abstract

The UPC-BMIC-VMU system is a **standard phrase-based statistical machine translation (PBSMT) 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**


## Collocation segmentation (CS)

http://donelaitis.vdu.lt/~vidas/tools.htm

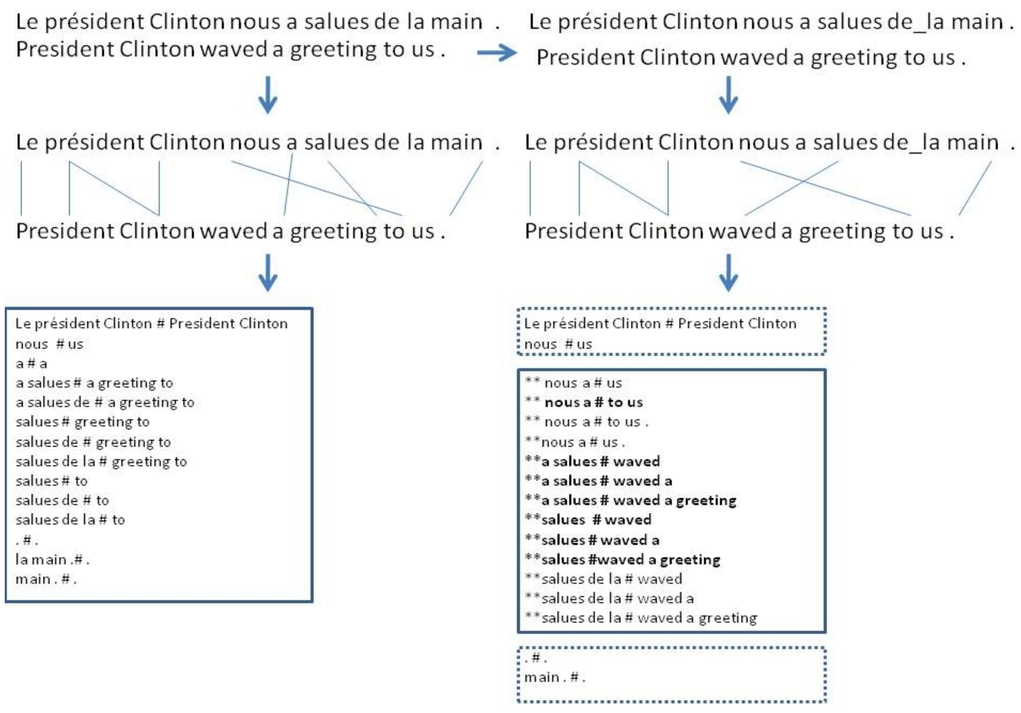
| Measure | Collocation segmentation example |
|---|---|
| chi2 | they were listening to his speech with open ears . |
| likelihood | they were listening to his speech with open ears . |
| dice | they_were_listening to his_speech with open_ears . |
| MI | they_were_listening_to his_speech_with open_ears . |
| 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_? |

## Translation results

| 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** |

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

## CS+PBSMT integration




## 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).

## Conclusions

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 system (likelihood new phrases)** in 3 test sets and we obtained coherent improvements in all of them.


**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.*