# Improving Back-off Models with Bag of Words and Hollow-grams

Benjamin Lecouteux, Raphaël Rubino, Georges Linarès

# Improving Back-off Models with Bag of Words and Hollow-grams

Benjamin Lecouteux, Raphaël Rubino, Georges Linarès
Laboratoire Informatique d'Avignon (LIA), University of Avignon , France
{benjamin.lecouteux, raphael.rubino, georges.linares}@univ-avignon.fr

## Introduction

- We propose a simple and efficient model based on **word co-occurrences** and a **new hollow-gram model**.
- Our approaches are applied to traditional modified Kneser-Ney back-off language models. []
- We decide to take into account the short context around $n$-grams to tackle the issue of **sparse training data**.
- A slight improvement on word error rate (WER) is reached, with and without acoustic adaptation.

## Proposed Approaches

### Hollow-gram model

- We propose an alternative to the indepedence assumption between two distant words in modified Kneser-ney back-off models.
- A 2-gram model based on the pairs (start, end) of each 3-gram, allowing relationship between short distant words.
- Can be assimilated to a regular expression $(w1, *, w3)$, where $*$ is an unobserved word.
- In the case of un unseen event, the hollow-gram backs off using the equation:
$\tilde{p}(w_3|w_1, w_2) = \alpha(w_1, w_2)^{1-\beta} p_\phi(w_3|w_1)^\beta p(w_3|w_2)$
where $\tilde{p}(w_3|w_1, w_2)$ : resulting updated probability of the 3-gram, $\alpha(w_1, w_2)$ : initial back-off value of the word pair $(w_1, w_2)$, $p_\phi(w_3|w_1)$ : hollow-gram probability, $\beta$ : empirical fudge factor.

### Word co-occurence model

- We combine a word association score [**?**] to the back-off based on classical n-gram language model.
- It eliminates word frequency effects and emphasizes relations between significant word co-occurrences.
- Our model can be interpolated with an initial modified Kneser-ney back-off model, on 2-gram and 1-gram.
- In the case of a full back-off behavior,
$\tilde{p}(w_3|w_1, w_2) = \alpha(w_1, w_2)^{1-\beta} p_\psi(w_1, w_3)^\beta \alpha(w_2)^{1-\beta} p_\psi(w_2, w_3)^\beta p(w_3)$
where $\alpha(w_1, w_2)$ : initial back-off value of the words, $\beta$ : empirical fudge factor, $p_\psi$ : back-off smoothing function based on word co-occurrences.

## Experimental Framework

- Experiments are carried out by using the **LIA broadcast news system** [**?**], which relies on the HMM-decoder **LIA SPEERAL**. []
- A **classical 3-gram model** built on *Le Monde* French Newspaper and Gigaword corpus (1.3G words).
- Training and development parts of the data set are based on the **ESTER-2** [**?**] evaluation campagn corpus (100h).

| Shows | baseline WER | baseline SER | baseline CWR |
|---|---|---|---|
| Inter (4h) | 33.1 | 75.0 | 69.7 |
| TVME (1h) | 31.3 | 67.6 | 71.2 |
| RFI (1h) | 18.7 | 66.6 | 84 |

Table: Baseline with the experimental framework

## Experiments

- Four series of experiments: for each approach separately, for the combination and for a compact model.
- For each experiment, the initial back-off value is re-estimated according to: $\tilde{\alpha}(w_{i-n}, .., w_i) = \alpha(w_{i-n}, .., w_{i-1})^{1-\beta} p_\phi(w_{i-n}, w_i)^\beta$ with $\tilde{\alpha}$ : updated back-off value, $\alpha$ : initial back-off value, $p_\phi(w_{in}, w_i)$ : smoothing function
- The **back-off based hollow grams** has two advantages: it is a regular expression based model, it can capture hollow-grams events into the training corpora.
- For the **back-off based on word co-occurrences**, a word co-occurence symmetric matrix is built on the whole training corpora, counting word pairs with a window size of five words.
- The **compact model** depicts the binary possibility of a back-off. This binary possibility is computed from the co-occurrence matrix: if the value is not null, we consider as true the possibility of the combination.

## Results

### Hollow-gram model

| Shows | WER | SER | CWR |
|---|---|---|---|
| Inter (4h) | 32.8 (-0.3) | 74.5 (-0.5) | 70.5 (+0.8) |
| TVME (1h) | 31.3 (0.0) | 67.0 (-0.3) | 71.7 (+0.5) |
| RFI (1h) | 18.5 (-0.2) | 65.7 (-0.9) | 84.4 (+0.4) |
| GLOBAL | -0.2 | -0.5 | +0.7 |

Table: WER, SER, CWR using the hollow-gram based backoff model

### Word co-occurence model

| Shows | WER | SER | CWR |
|---|---|---|---|
| Inter (4h) | 32.7 (-0.4) | 74.5 (-0.5) | 70.5 (+0.8) |
| TVME (1h) | 31.0 (-0.3) | 66.8 (-0.8) | 71.8 (+0.6) |
| RFI (1h) | 18.4 (-0.3) | 65.6 (-1) | 84.5 (+0.5) |
| GLOBAL | -0.4 | -0.6 | +0.7 |

Table: WER, SER, CWR using the cooccurrences based backoff model

### Combination

| Shows | WER | SER | CWR |
|---|---|---|---|
| Inter (4h) | 32.4 (-0.7) | 74.6 (-.4) | 70.8 (+1.1) |
| TVME (1h) | 30.9 (-0.4) | 67.2 (-.4) | 72.0 (+0.8) |
| RFI (1h) | 18.4 (-0.3) | 65.6 (-1.0) | 84.5 (+0.5) |
| GLOBAL | -0.6 | -0.5 | +0.9 |

Table: WER, SER, CWR using combination between cooccurrences based backoff and hollow-gram models

### Compact model

| Shows | WER | SER | CWR |
|---|---|---|---|
| Inter (4h) | 32.7(-0.3) | 74.8(-0.2) | 70.3(+0.6) |
| TVME (1h) | 31.1(-0.2) | 67.5(-0.1) | 71.6(+0.3) |
| RFI (1h) | 18.6(-0.1) | 65.7(-0.9) | 84.2(+0.2) |
| GLOBAL | -0.25 | -0.3 | +0.5 |

Table: WER, SER, CWR using combination between cooccurrences based backoff and hollow-gram models

## Conclusion

- A simple back-off values reordering can improve a Kneser-Ney based model for a **0.6% absolute gain of WER and 0.9% of CWR**.
- Our compact version leads to a slight improvement of the classical back-off model, with a very **low memory consumption**.
- The **hollow-gram model** can be extended to $n$-grams.
- We plan to extend the **co-occurence model** to more sophisticated heuristics and algorithms, smoothing values with word distances.

## Ref