# Textual Data Selection for Language Modelling in the Scope of Automatic Speech Recognition

F. Mezzoudj[1], D. Langlois [2], D. Jouvet[2], A. Benyettou[1]

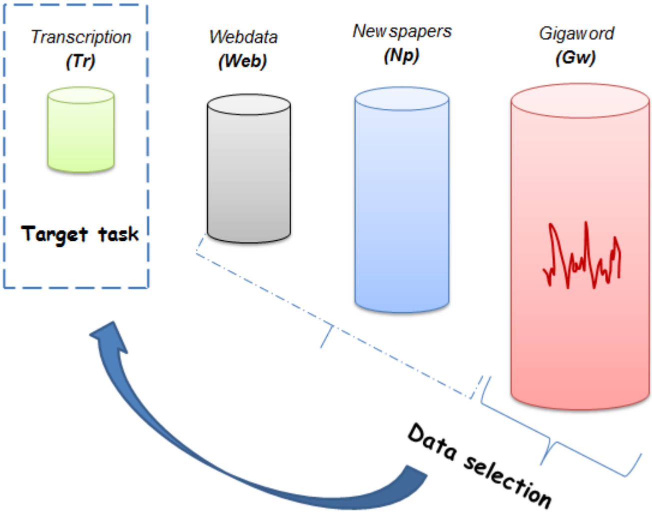
[1] USTO/UHBC-Algeria, [2] INRIA, LORIA, France

ICNLSP, Algiers, October $18^{th}$, 2015

# Introduction

- **Language model** (LM) is an important module in ASR systems.
- Learning LM requires a large amount of textual data.
- A high-performance LM is trained using a small corpus close to the target task (in-domain) and a huge corpus not close to this task (non-domain).
- We investigate selection of French textual data in order to improve LMs for automatic speech transcription of broadcast news & TV shows.

# Introduction



# Outline

# Data selection / Related work

**Application on 2 sources :**

- **Klakow (2000)** used a log-likelihood criterion to select newspaper articles.
- **Wang et al.(2002)** selected text units from the non-domain with lowest perplexity according to the in-domain LM.
- **Moore et al. (2010)** selected sentences from the non-domain with lowest difference cross- entropy according to 2 LMs with equal size, representing the in-domain LM and a non-domain LM).

# Data selection / New situation

We face a completely different situation :

- We use **4** corpora corresponding to different sources :
  - manual transcriptions of broadcast news & TV shows ;
  - Webdata (from Web sites : Magazines, TV) ;
  - Newspapers (Le Monde & L'Humanité) ;
  - Gigaword corpus $2^{nd}$ edition.
- Each corpus contributes differently to the final LM's training.
- Corpora may be noisy because of the variable quality of the sources.

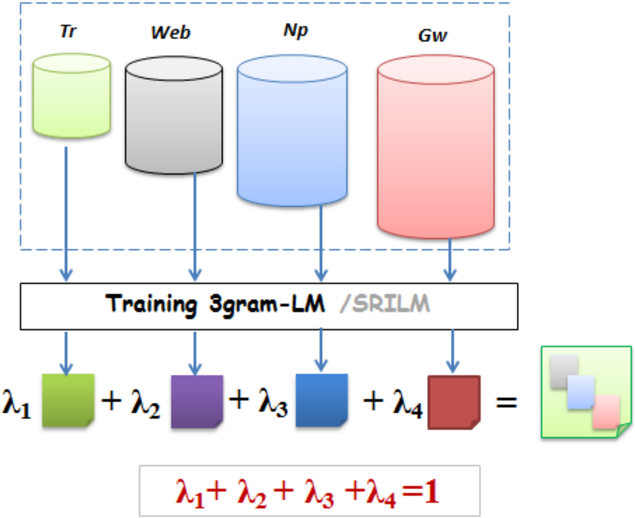
# Experimental setup

- **Training corpora** Sources & # words [M]

| | |
|---|---|
| *Tr* (radio broadcast transcriptions) | 114 |
| *Web* (web data) | 334 |
| *Np* (newspapers) | 526 |
| *Gw* (gigaword corpus) | 783 |
| $Tr + Web + Np + Gw$ | **1 757** |

- **Validation Corpus**, $DevLM \simeq 300$ K words
- **Test Corpus**, $TestLM \simeq 90$ K words
- **Vocabulary** $\simeq$ 100K words

# Experimental setup / baseline LM



# Experimental setup / baseline LM

Table : **Baseline LM**, interpolated from the individual LMs.

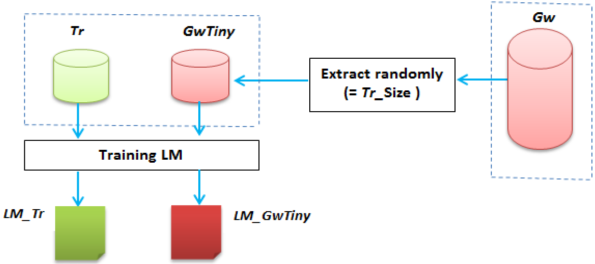
| Sources | Interpolated LM | | |
|---|---|---|---|
| | weights | ppl *DevLM* | ppl *TestLM* |
| *Tr* | 0.685 | 185.7 | **218.9** |
| *Web* | 0.246 | | |
| *Np* | 0.062 | | |
| GW | 0.007 | | |

- large difference in the weight of the individual LMs ;
- *Gw* brings small contribution in the final LM.

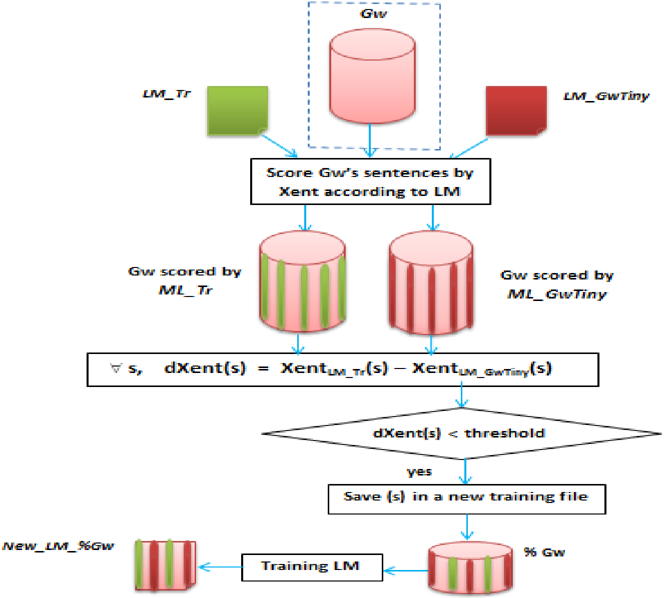
# Data selection strategy

## Step 1

- # source used : 2
- data selection on : ***Gw***
- "in-domain" : *LM_Tr*
- "non-domain" : *LM_GwTiny*



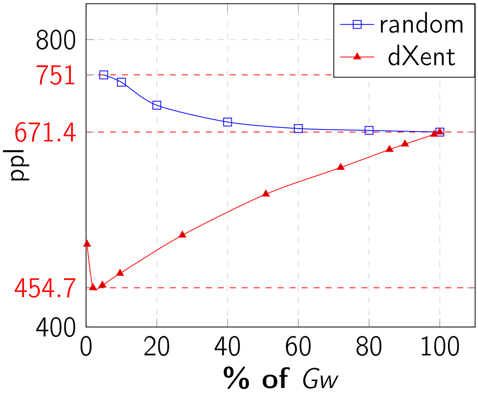
# Step 2



Gw
LM_Tr
LM_GwTiny
Score Gw's sentences by Xent according to LM
Gw scored by ML_Tr
Gw scored by ML_GwTiny
∀ s, dXent(s) = Xent$_{LM_Tr}$(s) − Xent$_{LM_GwTiny}$(s)
dXent(s) < threshold
yes
Save (s) in a new training file
New_LM_%Gw
Training LM
% Gw

# Data selection strategy

- The LM's ppl obtained by using the whole *Gw* corpus is 671.4 ;
- Small subsets selected **randomly** on the *Gw* data degrades the ppl ;
- The difference cross-entropy (**dXent**) selection data on the Gw corpus improves the ppl.

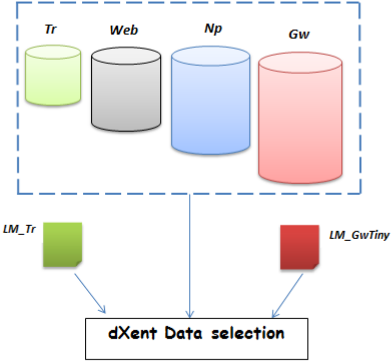


# Multisource-LM / approach 1

## approach 1

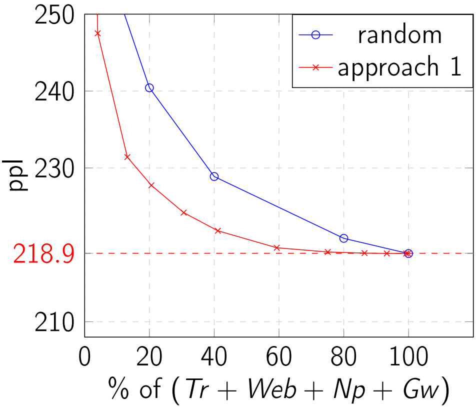
- # source used : 4 ;
- selected data :

  ***(Tr+ Web+ Np+ GW)*** ;
- "in-domain" : *LM _ Tr* ;
- "non-domain" : *LM _ GwTiny* ;



# Multisource-LM / approach 1

- The LM's ppl obtained by using the 4 corpora is 218.9
- The selection applied on data with a **random** process degrades the ppl ;
- The selection applied on data with the ***dXent*** computed with $LM\_(Tr)$ and $LM\_GwTiny$ (approach 1) doesn't improve the ppl.



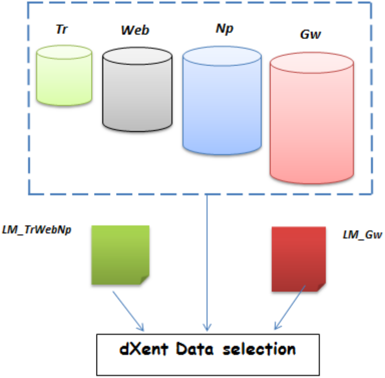
# Multisource-LM / approach 2

## approach 2

- # source used : 4 ;
- selected data :

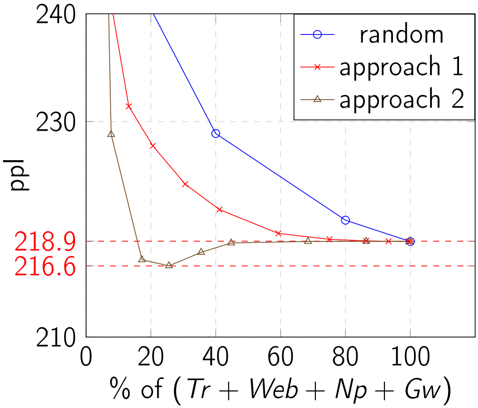
  **(*Tr*+ *Web*+ *Np*+ *GW*) ;**

- "in-domain" : *LM _ TrWebNp* ;
- "non-domain" : *LM _ Gw*.



# Multisource-LM / approach 2

- The selection (approach 2) applied on $(Tr, Web, Np, Gw)$ data with the ***dXent*** [computed with $LM\_(TrWebNp)$ and $LM\_Gw$] improves the ppl.



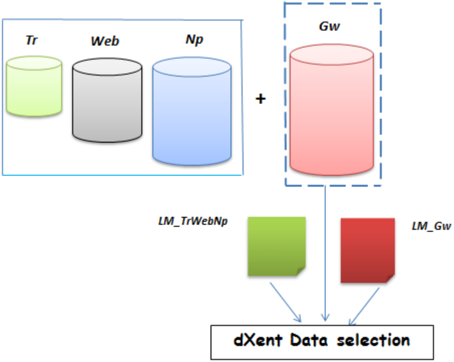
# Multisource-LM / approach 2

Table : Best LM, interpolated from the individual source LMs, after data selection using **approach 2**.

| Sources | Interpolated LM | | |
|---|---|---|---|
| | weights | ppl *DevLM* | ppl *TestLM* |
| *Tr* (88%) | 0.608 | 185.1 | **216.6** |
| *Web* (62%) | 0.234 | | |
| *Np* (26%) | 0.062 | | |
| *Gw* (0.2%) | **0.096** | | |

## approach 3

- # source used : 4 ;
- selected data : ***GW*** ;
- "in-domain" : *LM _ TrWebNp* ;
- "non-domain" : *LM _ Gw*.

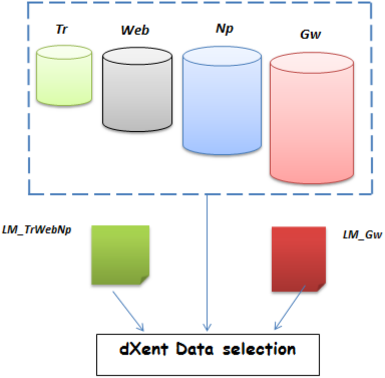


# vs approach 2

## approach 2

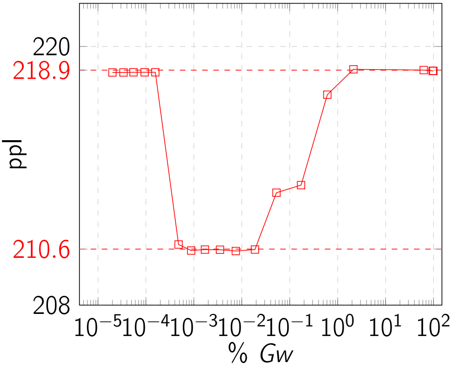
- # source used : 4 ;
- selected data :

  ***(Tr+ Web+ Np+ GW);***
- "in-domain" : *LM_TrWebNp* ;
- "non-domain" : *LM_Gw*.



# Multisource-LM / approach 3

- The selection applied on *Gw* data [with the ***dXent*** scored by *LM_(TrWebNp)* and *LM_Gw*] added to *Tr*, *Web* and *Np* data (approach 3) improves the ppl.



# Multisource-LM / approach 3

Table : Best LM, interpolated from the individual source LMs, after data selection using **approach 3**.

| Sources | Interpolated LM | | |
|---|---|---|---|
| | weights | ppl *DevLM* | ppl *TestLM* |
| *Tr*(100%) | 0.660 | 179.9 | **210.6** |
| *Web*(100%) | 0.240 | | |
| *Np*(100%) | 0.065 | | |
| *Gw*(0.05%) | **0.054** | | |

# Transcription experiments

- The speech corpora used come from ESTER2, ETAPE evaluation compaigns and the EPAC project.
- The speech transcription system relies on a diarization step and on the Sphinx toolkit.
- 39 HTK MFCC features are used (+ $1^{st}$ & $2^{nd}$ temporal derivatives).

# Transcription experiments

| LM | Size (gz file) | Etape Dev corpus | |
|---|---|---|---|
| | | ppl | WER[%] |
| $(Tr + Web + Np + Gw)$ | 1.2 Gb | 218.9 | 27.84 |
| $(Tr + Web + Np)$ | 809.8 Mb | 218.9 | 27.82 |
| LM(app.2, thresh. -0.3) | 391.3 Mb | 217.2 | 28.07 |
| LM(app.2, thresh. -0.2) | 501.6 Mb | 216.6 | 27.89 |
| LM(app.3, thresh. -0.6) | 809.3 Mb | **210.6** | **27.68** |

- The best LM is trained with 55.4% of (Tr,Web,Np,Gw).

# Conclusion

- The choice of the LMs that represent in-domain and non-domain is important for data selection.

- Keeping the 3 data sources (*Tr*, *Web*, *Np*) and selecting data from the *Gw* corpus with the *dXent* leads to better results than when selecting data from the whole corpora (*Tr*, *Web*, *Np*, *Gw*).

- We obtain competitive results in WER with reducing strongly the size of training corpus for LMs.

# Perspective

- We have to explore other ways to select data from the huge Gw in order to improve the LM performance.
- The vocabulary is a crucial module for transcription, so we take into account the time period of data sub-parts.

# For Further Reading I


R.C. Moore and W. Lewis.
"Intelligent selection of language model training data". In Proceedings of the ACL 2010 Conference Short Papers. pp. 220-224

G. Gravier, et al.
*"The etape corpus for evaluation of speech-based tv content processing in the french language", in LREC-Eighth ICLRE, 2012, p. na.*

# Thank you

for your attention !