2012
AMTA
20 Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Practical Domain Adaptation

Marcello Federico
Fondazione Bruno Kessler
Nicola Bertoldi
Fondazione Bruno Kessler

SAN DIEGO, CA
OCTOBER 28- NOVEMBER 1, 2012

Several studies have recently reported significant productivity gains by human translators when besides translation memory (TM) matches they do also receive suggestions from a statistical machine translation (SMT) engine. In fact, an increasing number of language service providers and in-house translation services of large companies is nowadays integrating SMT in their workflow. The technology transfer of state-of-the-art SMT technology from research to industry has been relatively fast and simple also thanks to development of open source software, such as MOSES, GIZA++, and IRSTLM.

While a translator is working on a specific translation project, she evaluates the utility of translating versus post-editing a segment based on the adequacy and fluency provided by the SMT engine, which in turn depends on the considered language pair, linguistic domain of the task, and the amount of available training data.

Statistical models, like those employed in SMT, rely on a simple assumption: data used to train and tune the models represent the target translation task. Unfortunately, this assumption cannot be satisfied for most of the real application cases, simply because for most of the language pairs and domains there is no sufficient data to adequately train an SMT system. Hence, common practice is to train SMT systems by merging together parallel and monolingual data from the target domain with as much as possible data from any other available source. This workaround is simple and gives practical benefits but is often not the best way to exploit the available data. This tutorial copes with the optimal use of in-domain and out-of-domain data to achieve better SMT performance on a given application domain.

Domain adaptation, in general, refers to statistical modeling and machine learning techniques that try to cope with the unavoidable mismatch between training and task data that typically occurs in real life applications. Our tutorial will survey several application cases in which domain adaptation can be applied, and presents adaptation techniques that best fit each case. In particular, we will cover adaptation methods for n-gram language models and translation models in phrase-based SMT. The tutorial will provide some high-level theoretical background in domain adaptation, it will discuss practical application cases, and finally show how the presented methods can be applied with two widely used software tools: Moses and IRSTLM.

The tutorial is suited for any practitioner of statistical machine translation. No particular programming or mathematical background is required.

Presenters

- Marcello Federico, Co-Director of the Human Language Technology Research Unit at Fondazione Bruno Kessler (FBK-irst), Trento, Italy.
- Nicola Bertoldi, PhD, Researcher for the Human Language Technology Research Unit at Fondazione Bruno Kessler (FBK-irst), Trento, Italy.

FONDAZIONE
BRUNO KESSLER

Practical
Domain Adaptation in SMT

Nicola Bertoldi
Marcello Federico
FBK, Trento, Italy

AMTA Tutorial, San Diego, 1 November 2012

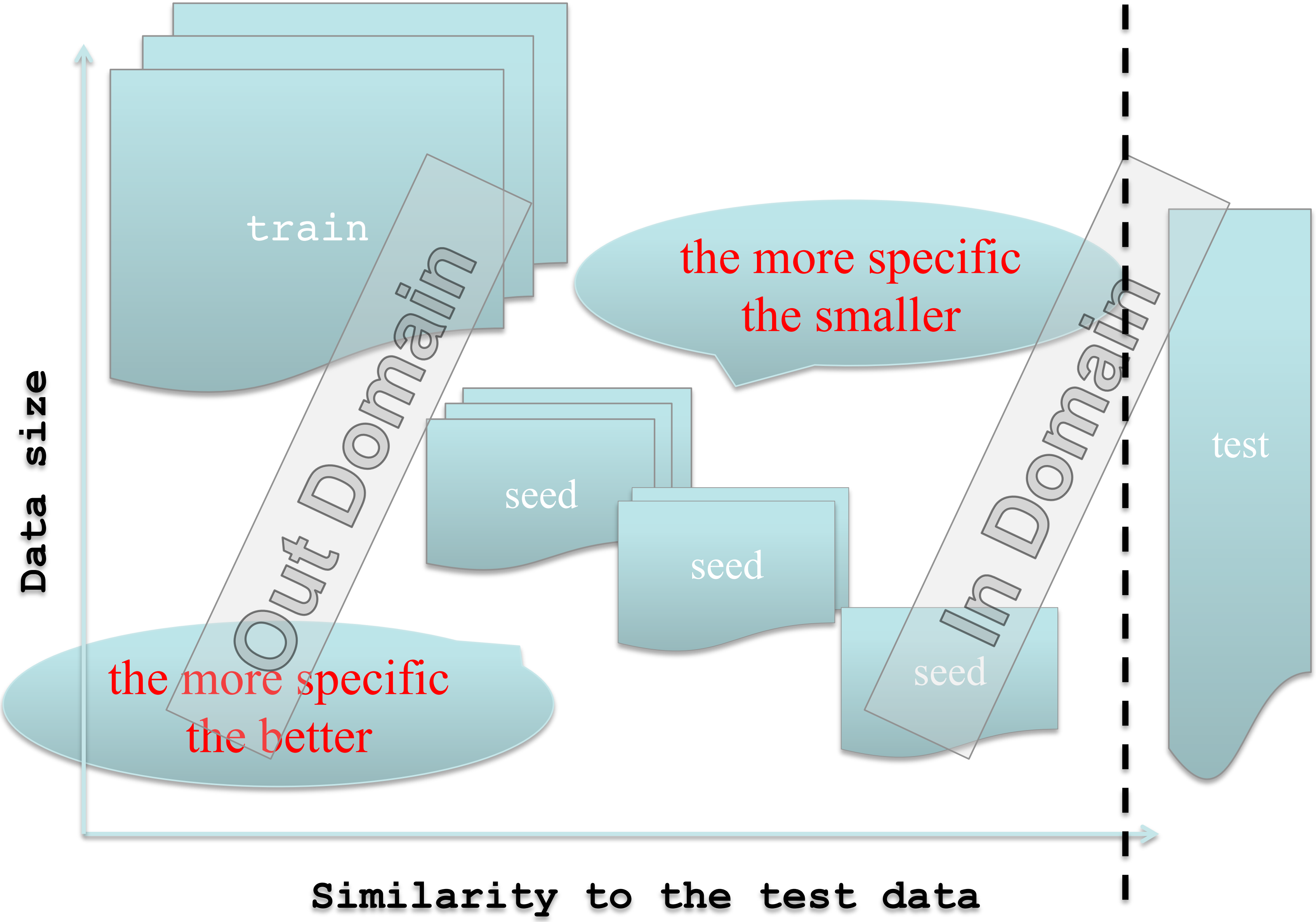
Outline - Practice

- case study
  - MateCat scenario
- data selection
- adaptation with IRSTLM and Moses
  - LM adaptation
  - TM adaptation
  - tuning
  - experimental comparisons
- guidelines

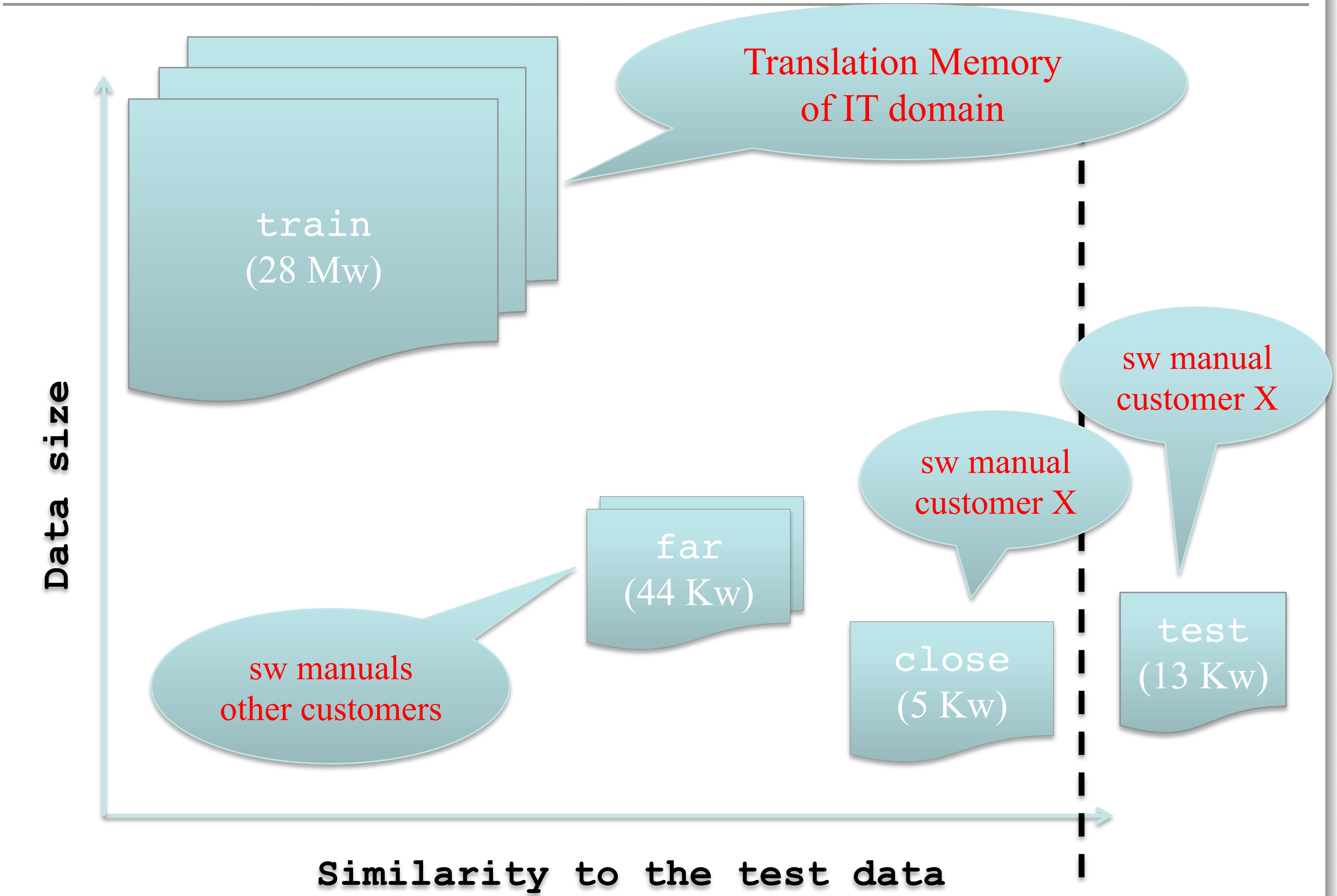
Outline - Practice

- **case study**
  - MateCat scenario
- data selection
- adaptation with IRSTLM and Moses
  - LM adaptation
  - TM adaptation
  - tuning
  - experimental comparisons
- guidelines

General scenario



Matecat – case study



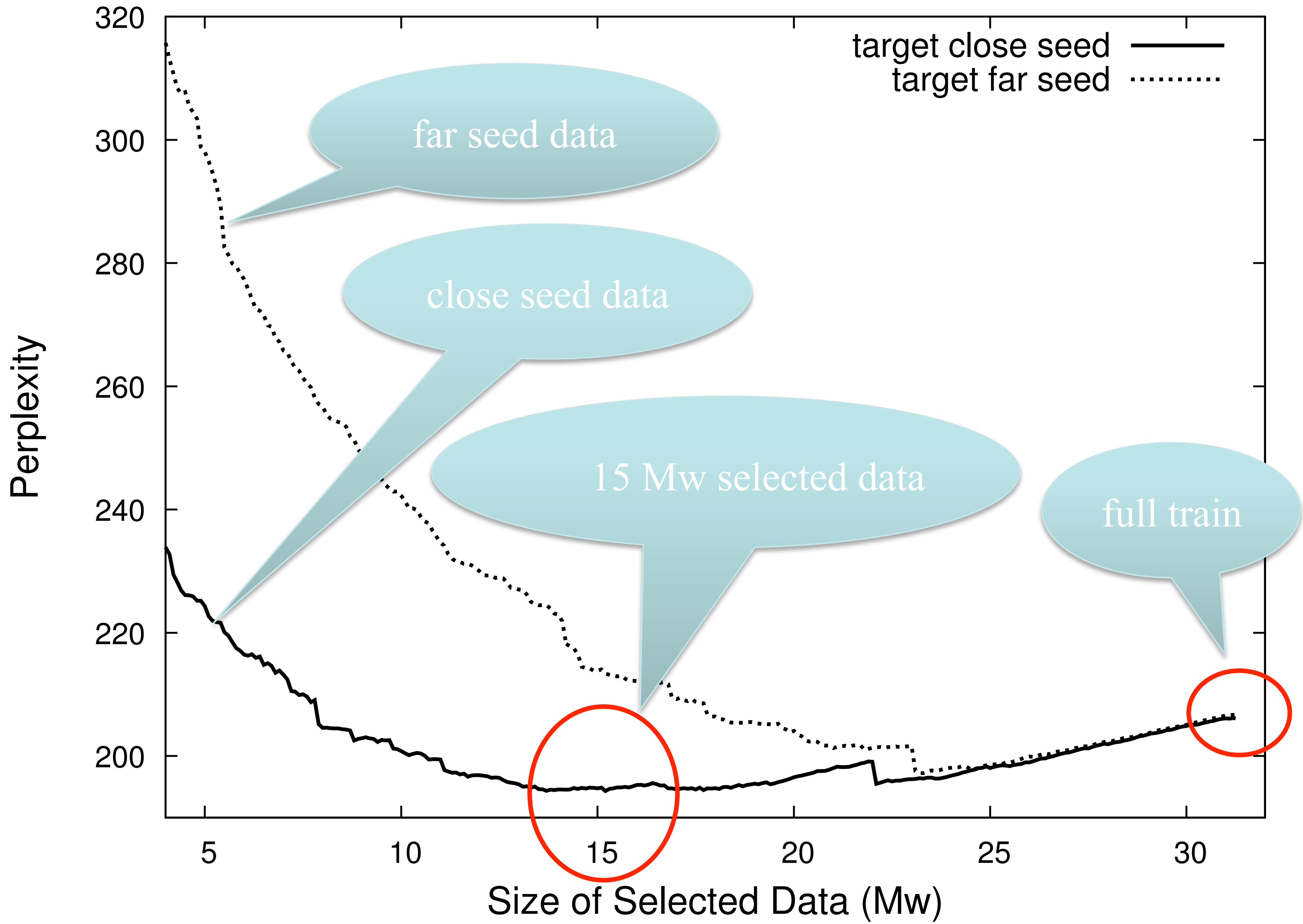
Matecat – case study

- test data:
    - `test`: software manual of a specific customer (13 Kw)
- training data:
    - `train`: Translation Memory of IT domain (28 Mw)
- seed data for adaptation:
    - `far`: software manuals of different customers (44 Kw)
    - `close`: software manual of the customer (5 Kw)
- results in terms of:
    - PP
    - BLEU

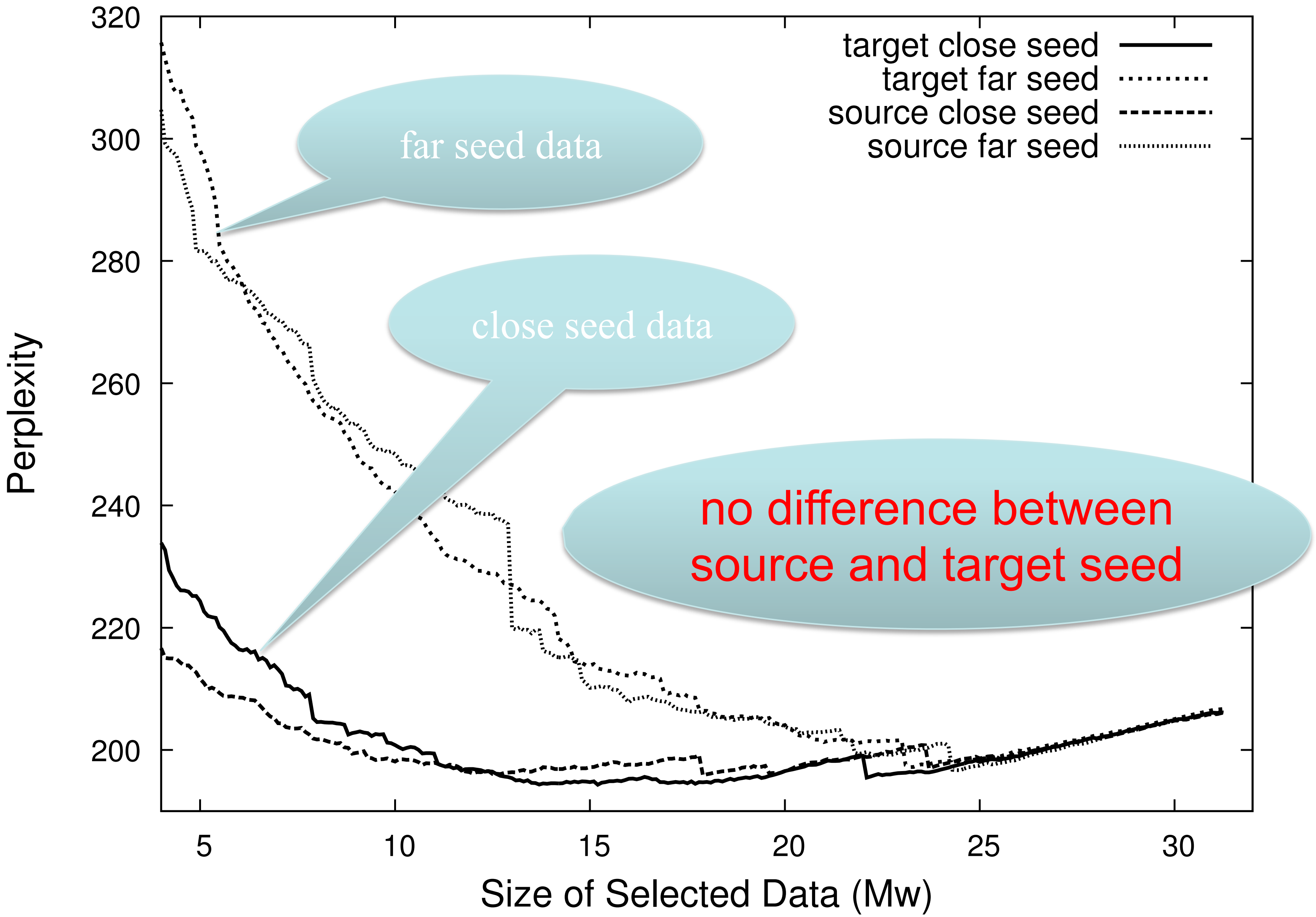
Outline - Practice

- case study
  - MateCat scenario
- **data selection**
- adaptation with IRSTLM and Moses
  - LM adaptation
  - TM adaptation
  - tuning
  - experimental comparisons
- guidelines

Data selection



Data selection



far seed data

close seed data

no difference between source and target seed

IRSTLM – dtsel

performs data selection

```
dtsel -i=seed -o=train -s=scores -x=1
```

seed

<s> A design element required for a timely , effective and efficient understanding of the risks involved in carrying out the activities . </s>
<s> Other graphic notations allow the reader to quickly recognise the various aspects of the proposed solutions and , in particular : </s>
…..

ordered scores

scores

-10.0464 605532 <s> Purpose </s>
-8.93316 25078 <s> Severity </s>
-8.80525 71650 <s> 7.9 </s>
…
-1.80406 674258 <s> Identification and evaluation of project risks ; </s>
-1.80395 365751 <s> IT Regulatory and Corporate Compliance </s>
-1.80151 258035 <s> Rational Build Forge Customer Benefits </s>
…

train

…
73757 <s> Depending on the sound card driver implementation th Device control may contain the list of installed sound car only while the Input control will hold the list of available ts for the chosen sound card including </s>
73758 <s> This s that you are interested in communic
7375
chatting an us
offline to all your contacts . </s>
73760 <s> Following service your replacement iPod touch may have a newer version of the OS . </s>
73761 <s> Lets you know if your browser supports CSS files . </s>
…

start and end symbols

segment index

IRSTLM – dtsel

computes PP with increasing selected data

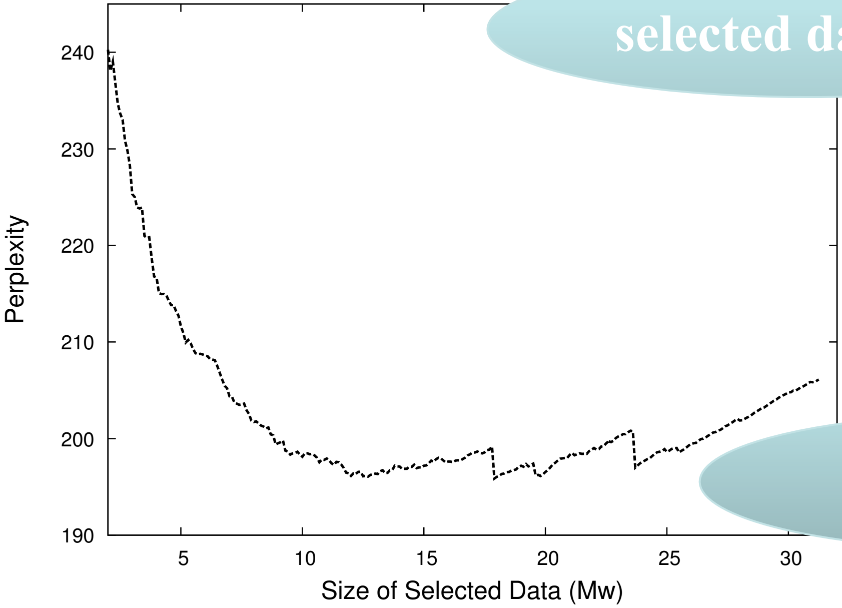
```
dtsel -test=test -s=scores -n=5 -x=1 > PP
```

test

```
<s> The duration that was entered is valid . </s>
<s> No action is required . </s>
<s> The date specified for the repeat until field is not valid . </s>
```

scores

```
-10.0464 605532 <s> Purpose </s>
-8.93316 25078 <s> Severity </s>
-8.80525 71650 <s> 7.9 </s>
…
-1.80406 674258 <s> Identification and evaluation of project risks ; </s>
-1.80395 365751 <s> IT Regulatory and Corporate Compliance </s>
-1.80151 258035 <s> Rational Build Forge Customer
</s>
```



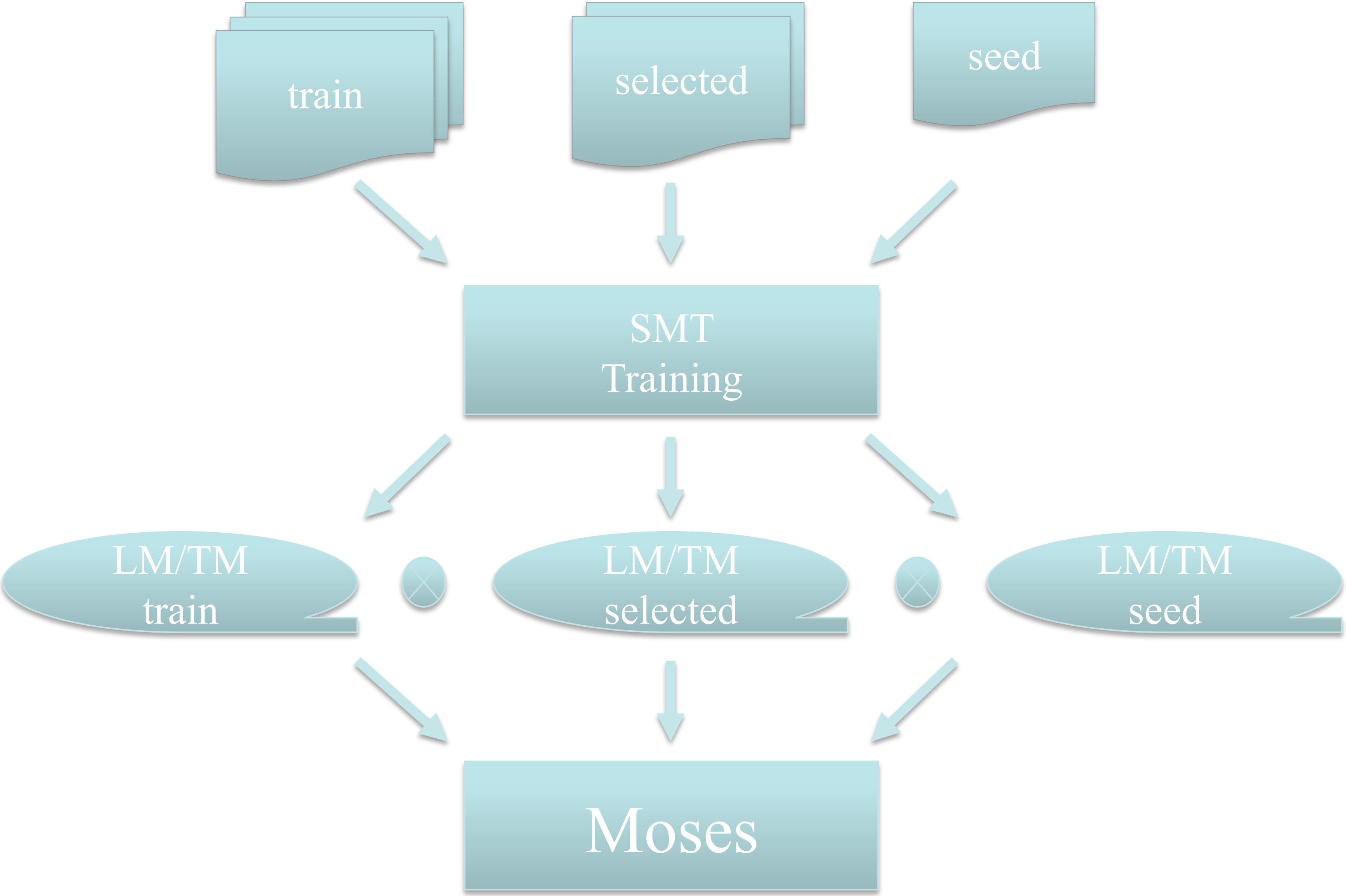
selected data size

PP

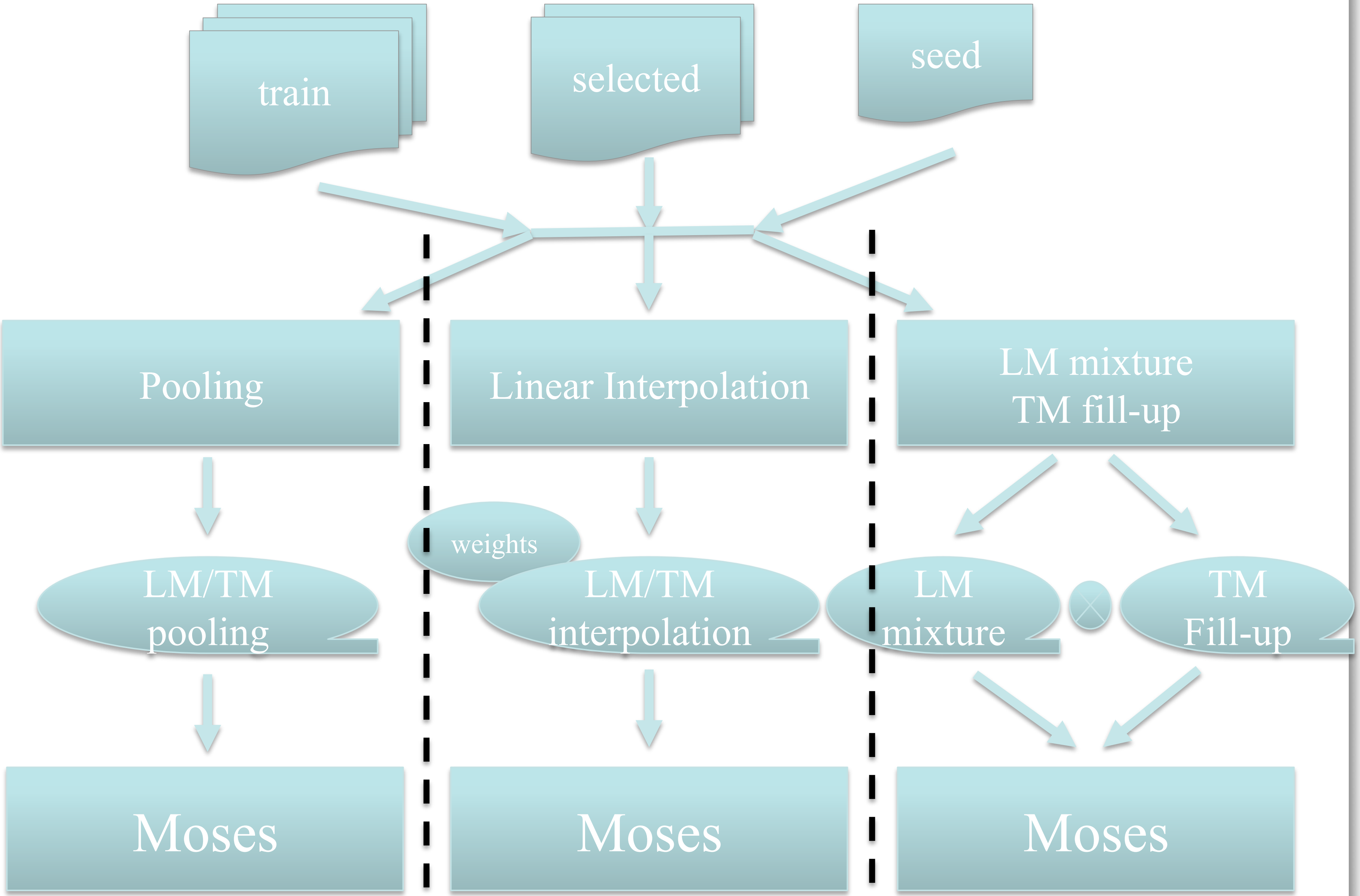
```
100001 1553.76
200004 843.818
300001 709.556
400023 539.762
500007 468.384
….
 31000020 206.088
31100051 206.079
31200004 206.148
31250910 206.09
```

perplexity

Using selected data



Using selected data


train

selected

seed

Pooling

Linear Interpolation

LM mixture
TM fill-up

weights

LM/TM
pooling

LM/TM
interpolation

LM
mixture

TM
Fill-up

Moses

Moses

Moses

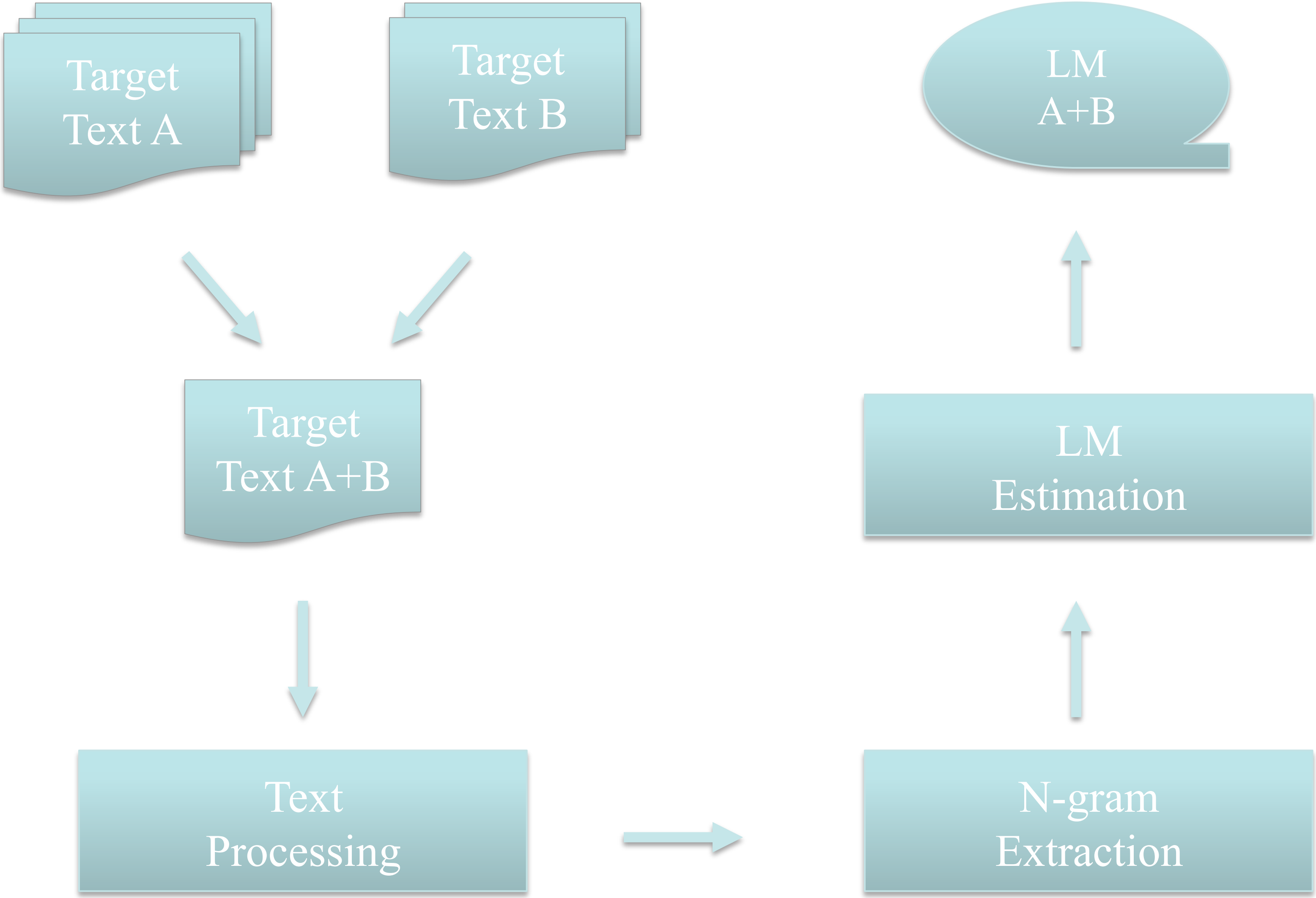
Outline - Practice

- case study
  - MateCat scenario
- data selection
- **adaptation with IRSTLM and Moses**
  - LM adaptation
  - TM adaptation
  - tuning
  - experimental comparisons
- guidelines

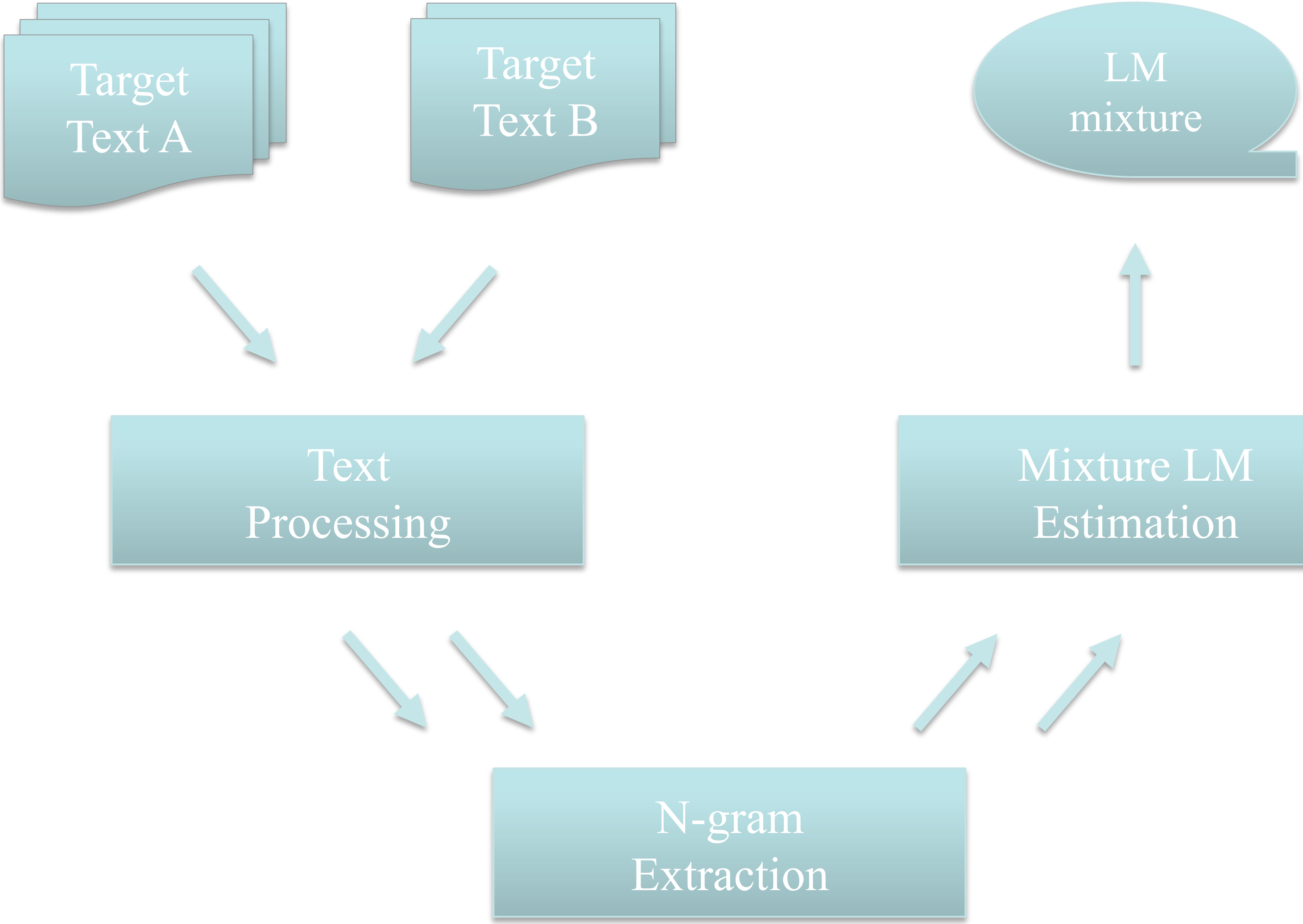
Outline - Practice

- case study
  - MateCat scenario
- data selection
- adaptation with IRSTLM and Moses
  - **LM adaptation**
  - TM adaptation
  - tuning
  - experimental comparisons
- guidelines

LM adaptation - pooling



LM adaptation - mixture



Target Text A

Target Text B

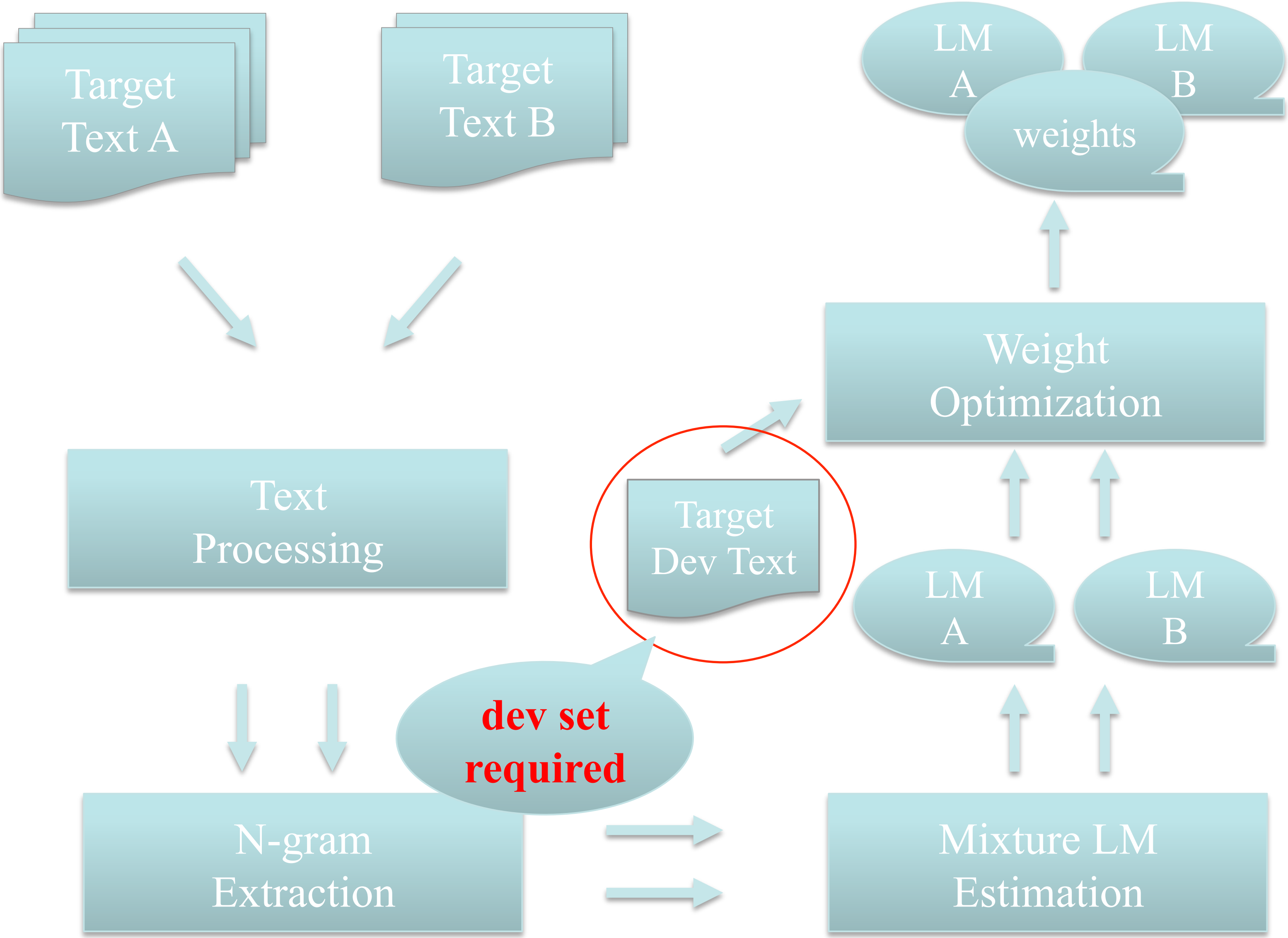
LM mixture

Text Processing

Mixture LM Estimation

N-gram Extraction

LM adaptation – linear interpolation



Target Text A

Target Text B

Text Processing

N-gram Extraction

Mixture LM Estimation

LM A

LM B

Target Dev Text

Weight Optimization

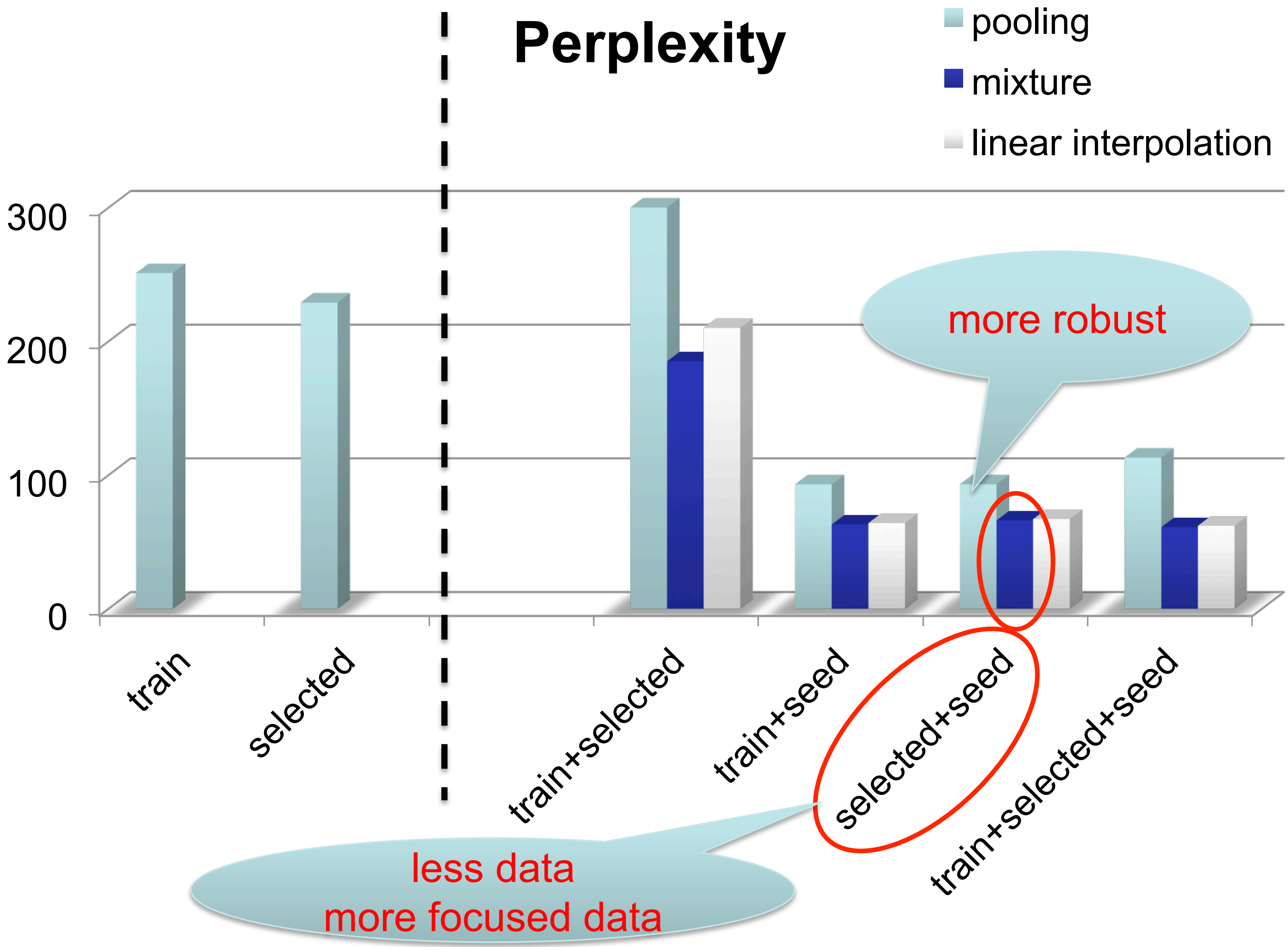
LM A

LM B

weights

dev set required

LM adaptation – comparison

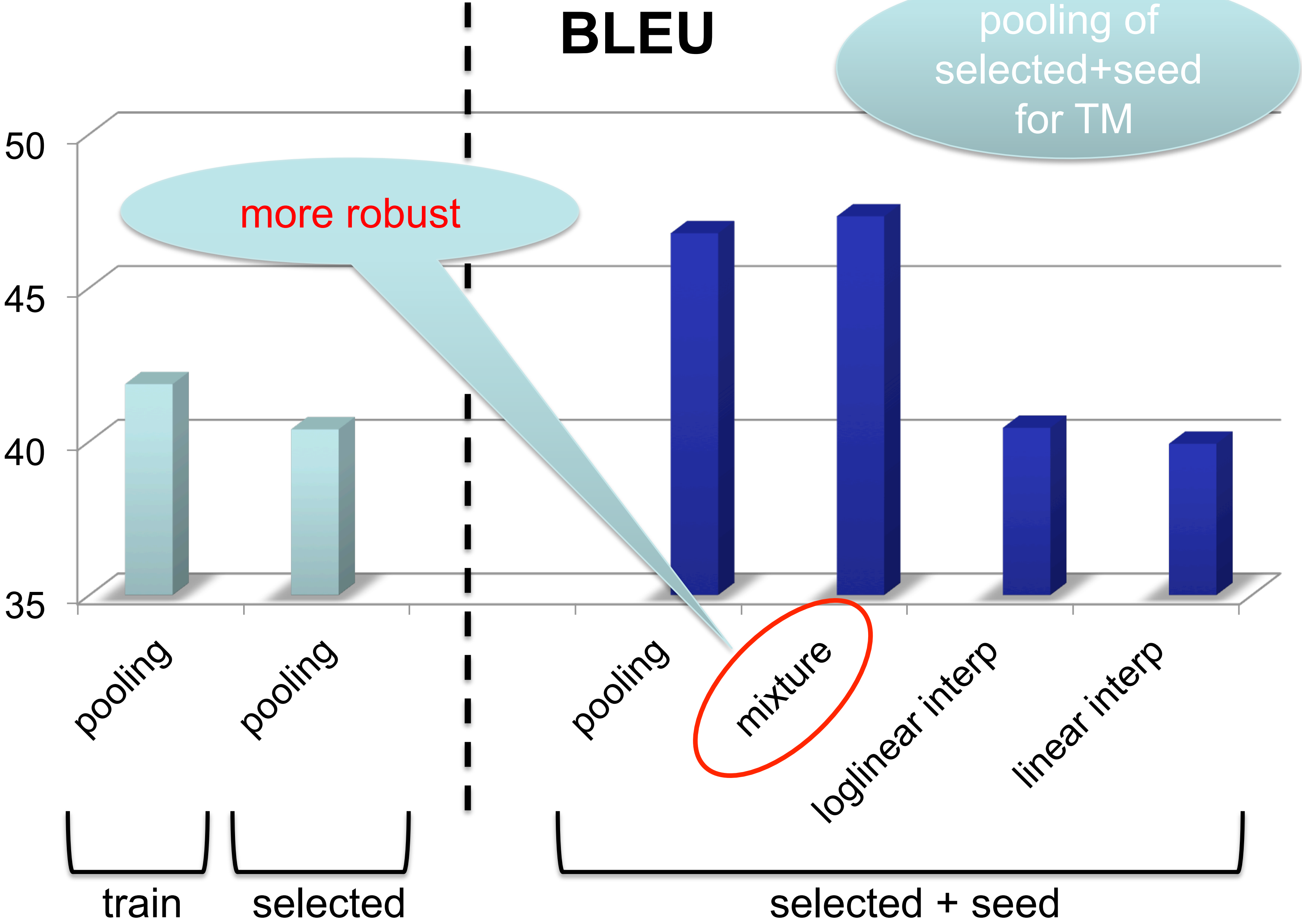


Perplexity

more robust

less data
more focused data

LM adaptation – comparison



IRSTLM – add-start-end.sh

adds start and end symbols

```
add-start-end.sh < train.txt > train.txt.se
```

```
train.txt
```

solemn ceremony marks handover
a solemn , historic ceremony has marked the resumption of the exercise of sovereignty over hong_kong by the people 's republic of china .
his royal highness the prince of wales and the president of the people 's republic of china (prc) he mr jiang zemin both spoke at the ceremony , which straddled midnight of june 30 and july 1 . the ceremony was telecast live around the world .
the ceremony took place in the grand hall of the hong_kong convention and exhibition centre (hkcec) extension and was attended by some 4,000 guests , including foreign ministers and dignitaries from more than 40 countries and international organisations ,
….

```
train.txt.se
```

end symbol

<s> solemn ceremony marks handov
<s> a solemn , historic ceremony has marked the resumption of the exercise of sovereignty over hong_kong by the
people republic of china . </s>
<s> his the prince of wales and the president of the people 's republic of china (prc) he mr jiang zemin
both y , which straddled midnight of june 30 and july 1 . the ceremony was telecast live around the
worl
<s> the ceremony took place in the grand hall of the hong_kong convention and exhibition centre (hkcec) extension and was attended by some 4,000 guests , including foreign ministers and dignitaries from more than 40 countries and international organisations , </s>
….

start symbol

IRSTLM - ngt

computes
n-gram statistics

```
ngt -i=train.txt -o=train.www.txt -n=3
ngt -i=train.txt -o=train.www.txt -n=3 -b=y
```

train.txt

```
<s> solemn ceremony marks handover  </s>
<s> a solemn , historic ceremony has marked the resumption of the exercise of sovereignty over hong_kong by the people 's republic of china .  </s>
<s> his royal highness the prince of wales and the president of the people 's republic of china ( prc ) he mr jiang zemin both spoke at the ceremony , which straddled midnight of june 30 and july 1 . the ceremony was telecast live around the world .  </s>….
```

train.www.txt

```
nGrAm 3 296851 ngram
15058
<s> 25000
solemn 12
ceremony 163
….
<s> <s> <s>     2
<s> <s> solemn  1
<s> solemn ceremony     1
<s> a solemn    1
<s> a ceremony  2
<s> a hong_kong 5
….
```

dictionary size

dictionary

n-gram statistics

train.www.bin

```
NgRaM 3 296851 ngram
15058
<s> 25000
solemn 12
ceremony 163
….

<binary_data>
```

different header

binary n-gram statistics

IRSTLM - tlm

estimates a language model

```
tlm -tr=train.txt -oarpa=train.msb.lm -n=3 -lm=msb -dub=1000000
tlm -tr=train.www.txt -oarpa=train.msb.lm -n=3 -lm=msb -dub=1000000
tlm -tr=train.www.bin -obin=train.msb.blm -n=3 -lm=msb -dub=1000000
```

train.txt

```
<s> solemn ceremony marks handover  </s>
<s> a solemn , historic ceremony has marked the
resumption of the exercise of sovereignty over
hong_kong by the people 's republic of china .
….
```

train.www.txt

```
nGrAm 3 296851 ngram
15058
<s> 25000
solemn 12
….
<s> <s> <s>     2
<s> <s> solemn  1
….
```

train.www.bin

```
NgRaM 3 296851 ngram
15058
<s> 25000
solemn 12
….

<binary_data>
```

train.msb.lm

```
\data\
ngram  1=     15059
ngram  2=   142684
ngram  3=     67566

\1-grams:
-5.20   <s>  -1.02
-4.29   solemn  -0.18
….
\3-grams:
-0.62   <s> <s> <s>
-2.75   <s> a ceremony
….
/end
```

n-gram size

logB(solemn)

logP(<s> a ceremony)

train.msb.blm

```
blmt 3     15059     142684     67566
15059
<s> 1
solemn 12
ceremony 163
….

<binary_data>
```

different header

dictionary

binary probs and backoff

IRSTLM – tlm for mixture

estimates a mixture LM

```
tlm -slmi=sublm -oarpa=train.mix.blm -n=3 -lm=mix -dub=1000000
tlm -slmi=sublm -obin=train.mix.blm -n=3 -lm=mix -dub=1000000
```

sublm

```
2
-slm=msb -str=adapt.www.bin -sp=0
-slm=msb -str=train.www.bin -sp=0
```

adapt.www.bin

```
NgRaM 3 56697 ngram
6208
<s> 2500
we 794
need 72
….

<binary_data>
```

train.www.bin

```
NgRaM 3 296851 ngram
15058
<s> 25000
solemn 12
ceremony 163
….

<binary_data>
```

train.mix.lm

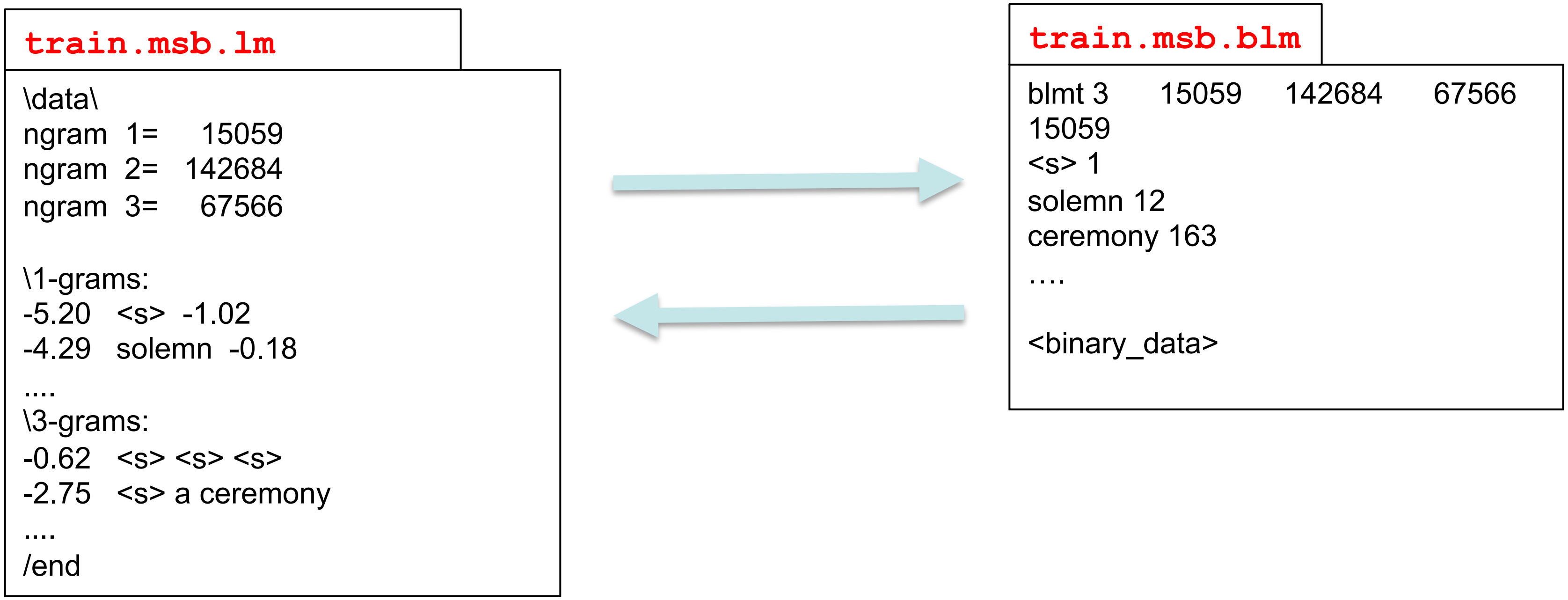
```
\data\
ngram  1=     16952
ngram  2=    163977
ngram  3=     71823
\1-grams:
-4.74   <unk>
-4.63   <s>  -0.99
-2.63   we
….
\3-grams:
-0.61   <s> <s> <s>
-1.76   <s> we need
-1.52   <s> we also
….
/end
```

mixture model can combine any number of language models

IRSTLM – compile-lm

transforms a LM
from txt to bin

```
compile-lm train.msb.lm train.msb.blm
compile-lm train.msb.blm train.msb.lm -t=y
```



train.msb.lm

```
\data\
ngram  1=     15059
ngram  2=   142684
ngram  3=     67566

\1-grams:
-5.20   <s>  -1.02
-4.29   solemn  -0.18
....
\3-grams:
-0.62   <s> <s> <s>
-2.75   <s> a ceremony
....
/end
```

train.msb.blm

```
blmt 3      15059     142684      67566
15059
<s> 1
solemn 12
ceremony 163
….

<binary_data>
```

IRSTLM – interpolate-lm

estimates the weights
of a interpolated LM

```
interpolate-lm config.in config.out -learn=test
```

config.in

```
LMINTERPOLATION 2
0.3 adapt.wb.blm
0.7 train.msb.blm
```

test

```
<s> debates of the senate ( hansard ) </s>
<s> 2 nd session , 36 th parliament , </s>
<s> volume 138 , issue 42 </s>
<s> tuesday , april 4 , 2000 </s>
….
```

config.out

```
LMINTERPOLATION 2
0.44589 adapt.wb.blm
0.55411 train.msb.blm
```

adapt.wb.blm

```
NgRaM 3 56697 ngram
6208
<s> 2500
we 794
need 72
….

<binary_data>
```

train.msb.blm

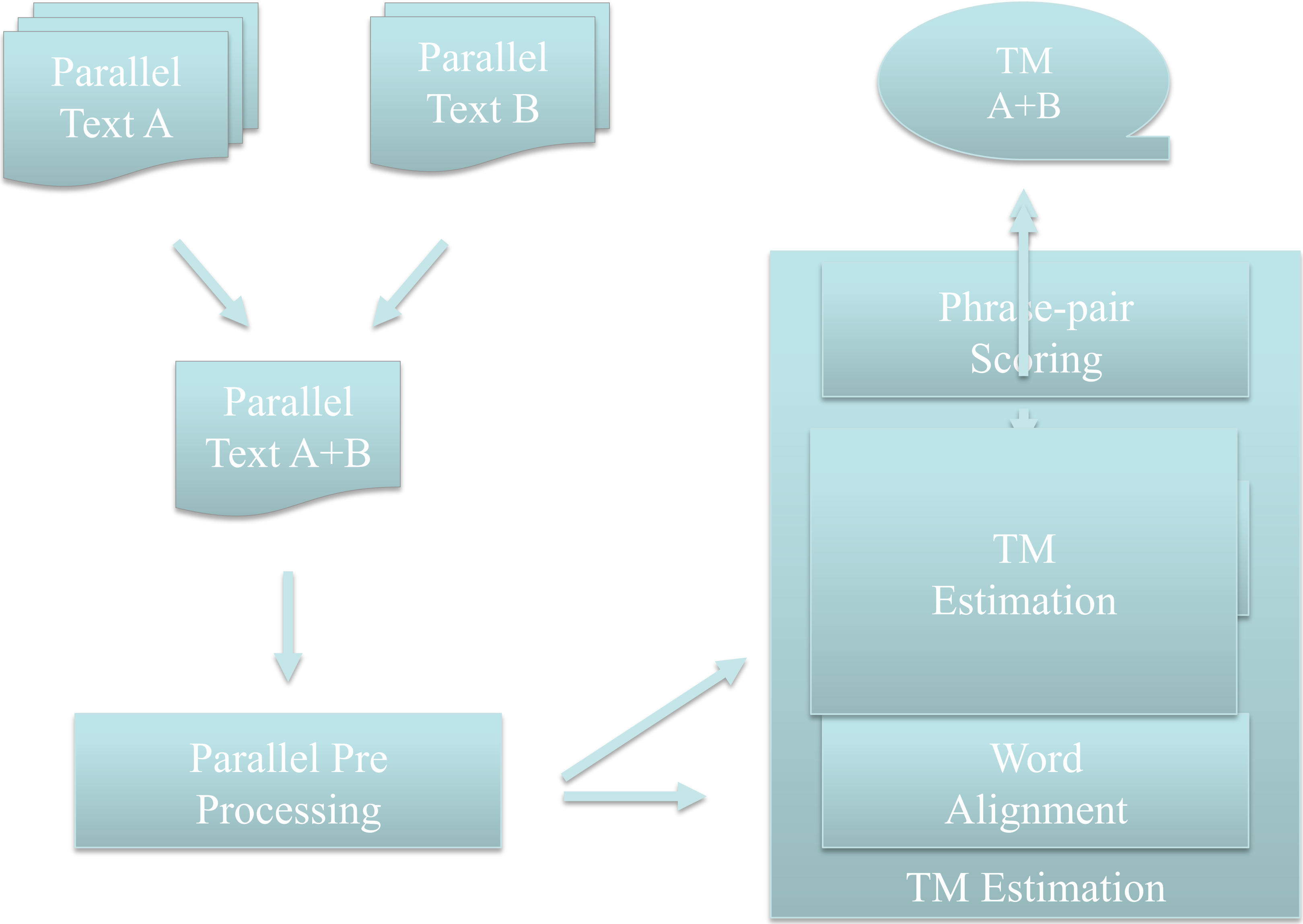
```
NgRaM 3 296851 ngram
15058
<s> 25000
solemn 12
ceremony 163
….

<binary_data>
```

Interpolated LM can combine
any number of language models
of any type

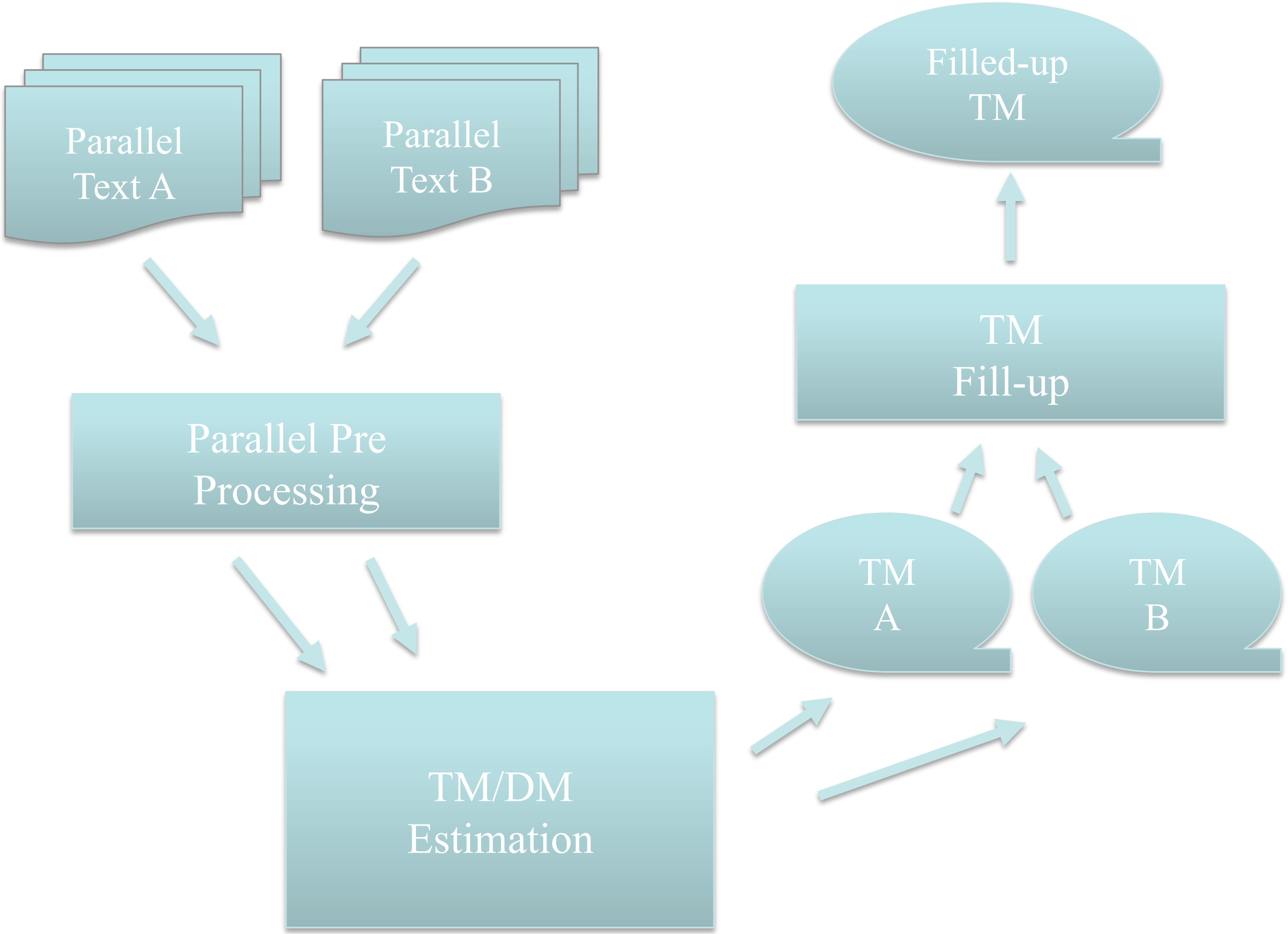
Outline - Practice

- case study
  - MateCat scenario
- data selection
- adaptation with IRSTLM and Moses
  - LM adaptation
  - **TM adaptation**
  - tuning
  - experimental comparisons
- guidelines

TM adaptation - pooling



TM adaptation – fill-up



Parallel Text A

Parallel Text B

Parallel Pre Processing

TM/DM Estimation

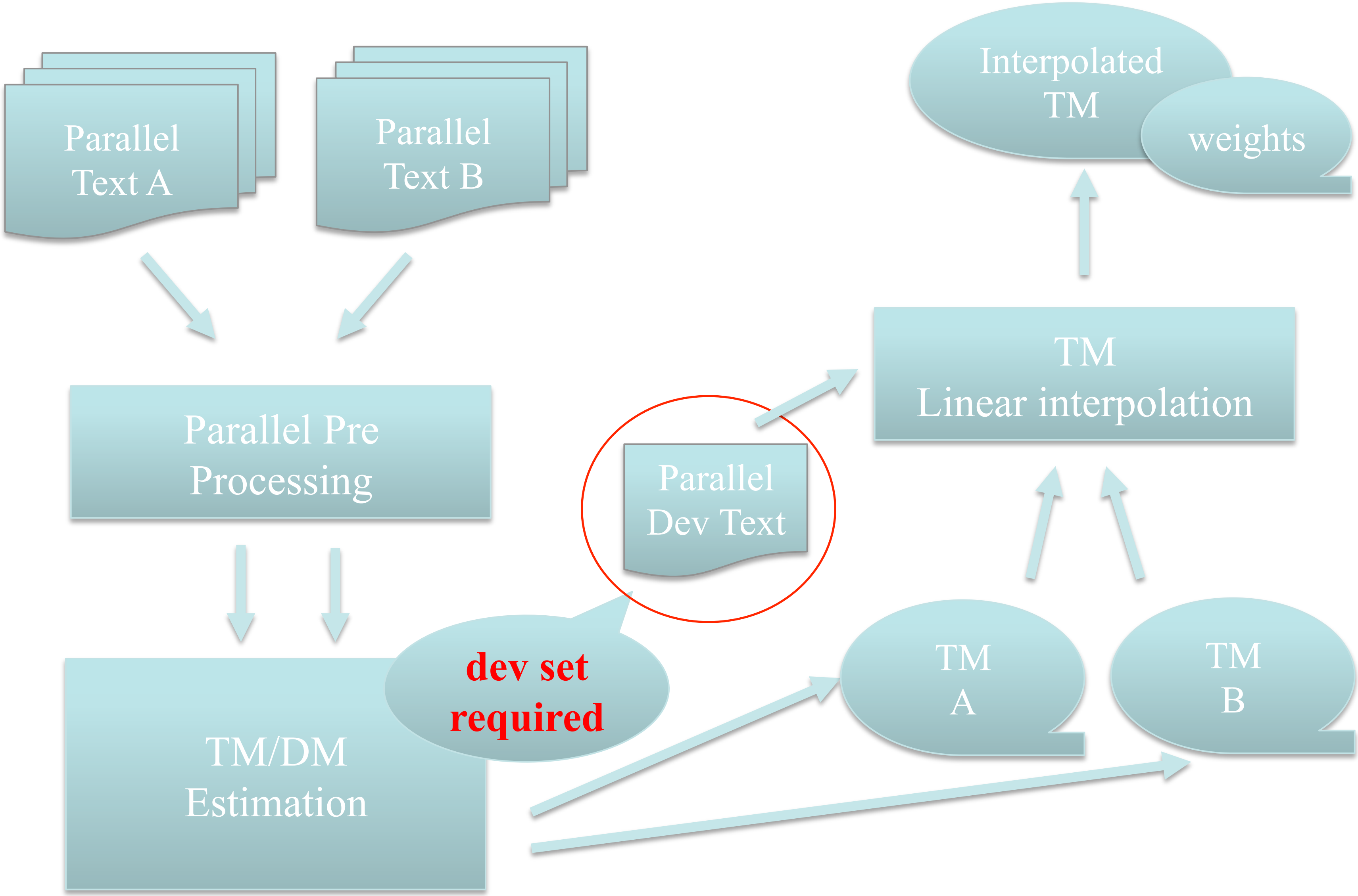
TM A

TM B

TM Fill-up

Filled-up TM

TM adaptation – linear interpolation



Parallel Text A

Parallel Text B

Parallel Pre Processing

TM/DM Estimation

dev set required

Parallel Dev Text

TM A

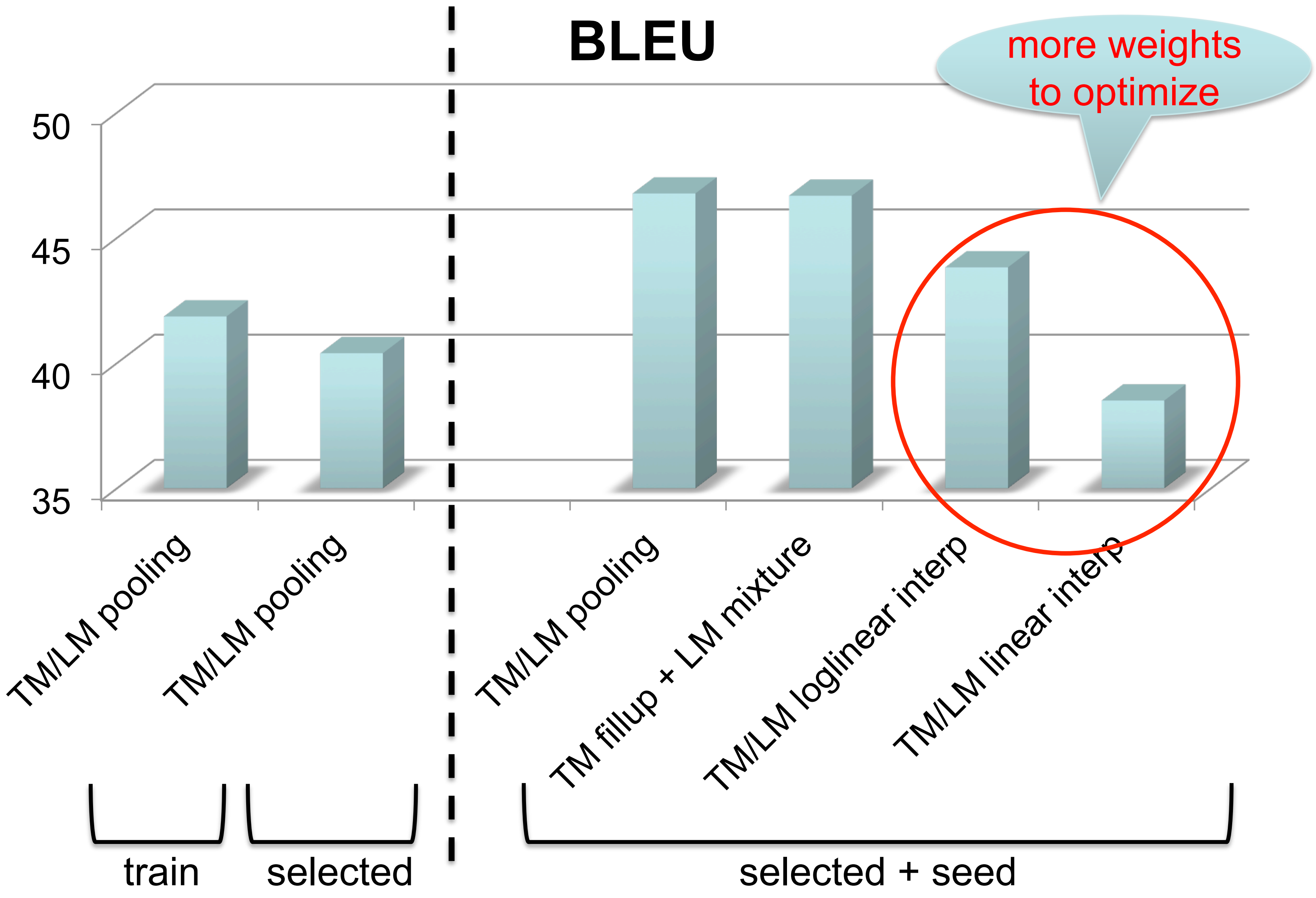
TM B

TM Linear interpolation

Interpolated TM

weights

TM/LM adaptation – comparison



more weights to optimize

Moses – clean-corpus.perl

cleans data

```
clean-corpus.perl –i train.clean –f en –e it
```

train.en

…
Accept the password if unable to check it
Access Allowed , Access Denied , Audit
...
Perform the following optional tasks to complete initial setup and prepare PRODUCT_TRADEMARK for production.
…
Folder Access Error
Fix Access Error .
…

train.it

…
Accetta la password se non è possibile verificarla
Accesso consentito , Accesso negato , Controllo
…
Effettuare le seguenti attività facoltative per completare l'impostazione iniziale e preparare PRODUCT_TRADEMARK per la produzione.
…
Errore di accesso alla cartella
Correggere l' errore di accesso .
…

train.clean.en

…
Accept the password if unable to check it
Access Allowed , Access Denied , Audit
...
~~Perform the following optional tasks to complete initial setup and prepare PRODUCT_TRADEMARK for production.~~
…
Folder Access Error
Fix Access Error .
…

train.clean.it

…
Accetta la password se non è possibile verificarla
Accesso consentito , Accesso negato , Controllo
…
~~Effettuare le seguenti attività facoltative per completare l'impostazione iniziale e preparare PRODUCT_TRADEMARK per la produzione.~~
…
Errore di accesso alla cartella
Correggere l' errore di accesso .
…

too long

Moses – train-perl

estimates
TM and DM

```
train-perl –i train.clean –f en –e it
```

train.clean.en

```
…
Accept the password if unable to check it
Access Allowed , Access Denied , Audit
 …
Folder Access Error
Fix Access Error .
…
```

train.clean.it

```
…
Accetta la password se non è possibile verificarla
Accesso consentito , Accesso negato , Controllo
…
Errore di accesso alla cartella
Correggere l' errore di accesso .
…
```

phrase-table

```
….
Accept the password if ||| Accetta la password se ||| 1.0  1.8e-1  1.0  6.2e-1  2.7
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.1e-2  2.7
Accept the ||| Accetta la ||| 1.0  3.5e-2  1.0  1.0e-1  2.7
Accept ||| Accetta ||| 1.0  1.0  1.0  1.0  2.7
Access Error . ||| errore di accesso . ||| 1.0  6.3e-2  1 2.0e-2 2.7
Access Error ||| errore di accesso ||| 1.0  6.3e-2 1 2.0e-2 2.7
Access State ||| Access State ||| 1.0  1.0  6.6e-1 2.0e-2 2.7
Access State ||| Stato di accesso ||| 1.0 1.0e-2  3.3e-1 1.5e-2  2.7
….
```

lexicographically
sorted

config

```
[ttable-file]
0 0 0 5 phrase-table.gz

[weight-t]
0.2
0.2
0.2
0.2
0.2

[lmodel-file]
1 0 5 train.blm

[weight-l]
0.5
```

Moses – fill-up

computes filled-up TM

```
combine-ptables.pl  -mode fillup  pt-1   pt-2   >   pt-fillup
```

pt-1

```
….
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.0e-2  2.7
…
Access Error . ||| errore di accesso . ||| 1.0  6.2e-2  1.0  2.1e-2  2.7
Access Error ||| errore di accesso ||| 1.0  6.7e-2  1.0  2.3e-2  2.7
….
```

pt-2

```
….
Access Allowed , ||| Accesso consentito , ||| 1.0  5.3e-3  1.0  2.3e- 2  2.7
Access Allowed ||| Accesso consentito ||| 1.5e-1  6.3e-3 1.0  3.e-2  2.7
….
Access error ||| Errore Accesso ||| 1.0  1.1e-1  1.0  2.9e-2  2.7
….
```

pt-fillup

```
….
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.0e-2  2.7   1
….
Access Allowed , ||| Accesso consentito , ||| 1.0  5.3e-3  1.0  2.3°-2  2.7   2.7
Access Allowed ||| Accesso consentito ||| 1.5e-1  6.3e-3 1.0  3.4e-2  2.7   2.7
….
Access Error . ||| errore di accesso . ||| 1.0  6.2e-2  1.0  2.1e-2  2.7   1
Access Error ||| errore di accesso ||| 1.0  6.7e-2  1.0  2.3e-2  2.7   1
….
```

belongs to pt-1

belongs to pt-2

Moses – linear interpolation

computes interpolated TM

```
combine-ptables.pl  -mode interp   pt-1  pt-2   >   pt-interp
```

pt-1

```
….
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.0e-2  2.7
…
Access Error . ||| errore di accesso . ||| 1.0  6.2e-2  1.0  2.1e-2  2.7
Access Error ||| errore di accesso ||| 1.0  6.7e-2  1.0  2.3e-2  2.7
….
```

pt-2

```
….
Access Allowed , ||| Accesso consentito , ||| 1.0  5.3e-3  1.0  2.3e- 2  2.7
Access Allowed ||| Accesso consentito ||| 1.5e-1  6.3e-3 1.0  3.e-2  2.7
….
Access error ||| Errore Accesso ||| 1.0  1.1e-1  1.0  2.9e-2  2.7
….
```

pt-interp

```
….
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.0e-2  2.7
….
Access Allowed , ||| Accesso consentito , ||| 5.0e-1  2.6e-3  5.0e-1  1.1°-2  1.3
Access Allowed ||| Accesso consentito ||| 7.6e-2  3.1e-3  5.0e-1  1.5°-2  1.3
….
Access Error . ||| errore di accesso . ||| 5.0 e-1  3.4e-2  5.0e-1  1.3e-2  1.
Access Error ||| errore di accesso ||| 5.0 e-1  3.4e-2  5.0e-1  1.3e-2
….
```

interpolation and fill-up
can combine
any number of phrase tables

Moses – compression

trasforms into binary TM

```
processPhraseTable -ttable 0 0 phrase-table -out phrase table -nscores 5
```

actual number of scores

phrase-table

```
….
Accept the password if ||| Accetta la password se ||| 1.0  1.8e-1  1.0  6.2e-1  2.7
Accept the password ||| Accetta la password ||| 1.0  1.8e-1  1.0  9.1e-2  2.7
Accept the ||| Accetta la ||| 1.0  3.5e-2  1.0  1.0e-1  2.7
Accept ||| Accetta ||| 1.0  1.0  1.0  1.0  2.7
Access Error . ||| errore di accesso . ||| 1.0  6.3e-2  1 2.0e-2 2.7
Access Error ||| errore di accesso ||| 1.0  6.3e-2 1 2.0e-2 2.7
Access State ||| Access State ||| 1.0  1.0  6.6e-1 2.0e-2 2.7
Access State ||| Stato di accesso ||| 1.0 1.0e-2  3.3e-1 1.5e-2  2.7
….
```

```
phrase-table.binphr.binphr.srctree
phrase-table.binphr.binphr.srcvoc
phrase-table.binphr.binphr.tgtdata
phrase-table.binphr.binphr.tgtvoc
phrase-table.binphr.binphr.idx
```

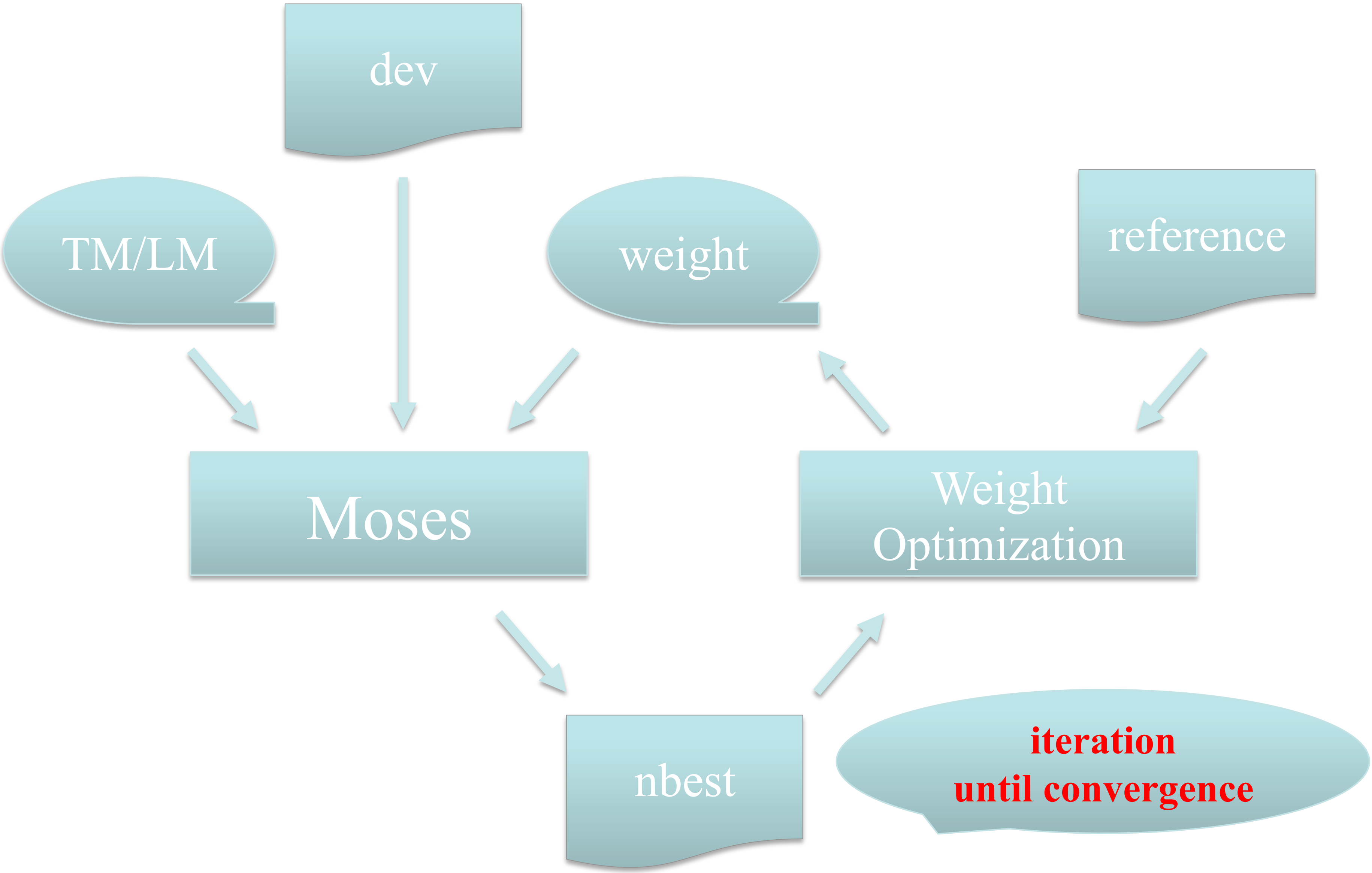
few binary files with this prefix

Outline - Practice

- case study
  - MateCat scenario
- data selection
- adaptation with IRSTLM and Moses
  - LM adaptation
  - TM adaptation
  - **tuning**
  - experimental comparisons
- guidelines

Tuning

optimizes Moses weights through MERT



dev

TM/LM

weight

reference

Moses

Weight Optimization

nbest

iteration until convergence

Moses – mert-moses.pl

estimates
TM and DM

```
mert-moses.pl dev.en dev.it moses-cmd config.baseline
```

actual
Moses decoder

dev.en

Perform the following initial setup tasks to set up PRODUCT_TRADEMARK for the first time.
PRODUCT_TRADEMARK contains the following plug-ins.
….

dev.it

Eseguire le attività di impostazione iniziale per configurare PRODUCT_TRADEMARK per la prima volta.
PRODUCT_TRADEMARK contiene i seguenti plug-in.
….

config.baseline

```
[ttable-file]
0 0 0 5 phrase-table.gz

[weight-t]
0.2
0.2
0.2
0.2
0.2

[lmodel-file]
1 0 5 train.blm

[weight-l]
0.5
```

actual
config file

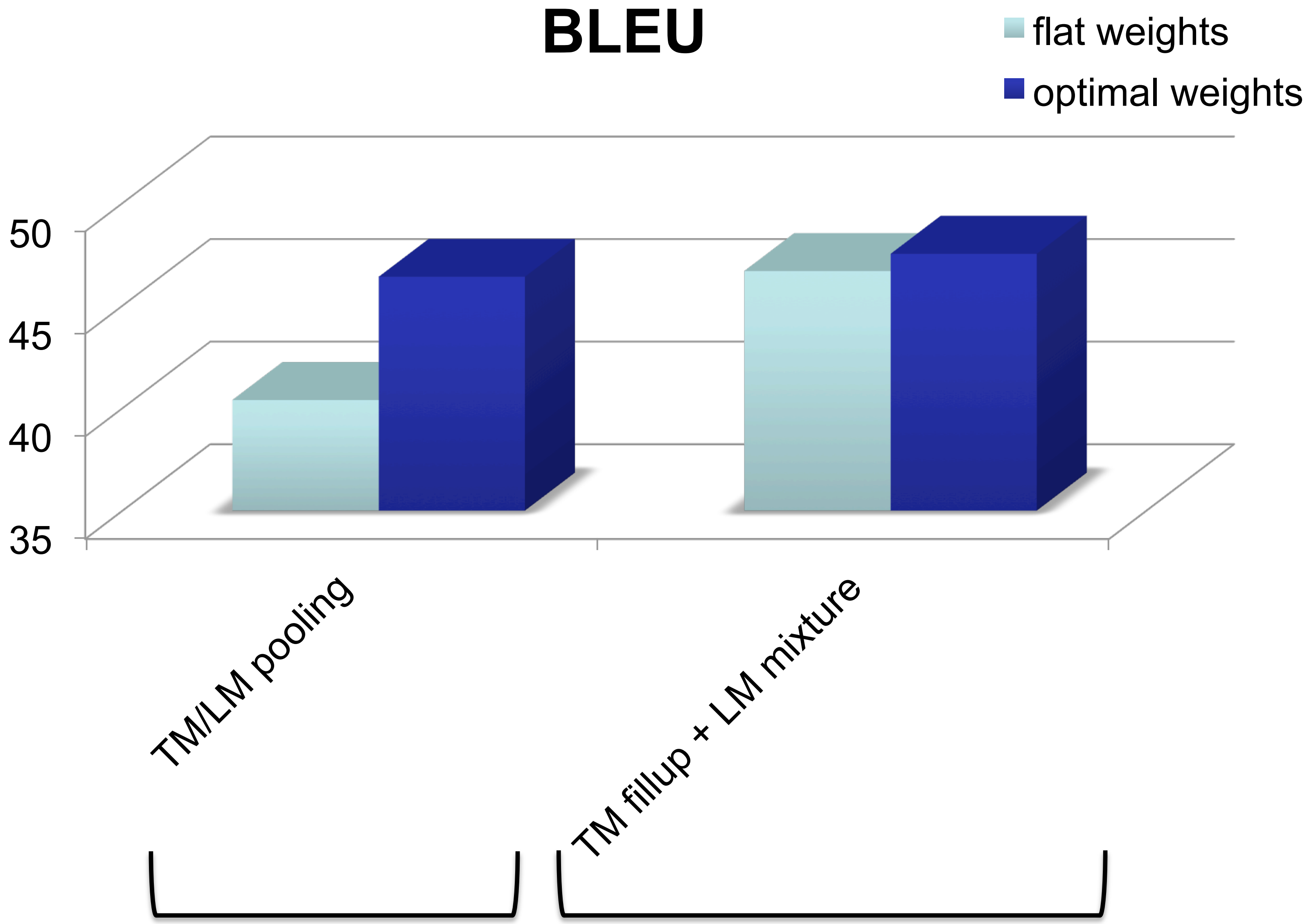
config.optimized

```
[ttable-file]
0 0 0 5 phrase-table.gz

[weight-t]
0.039
0.014
0.157
0.137
-0.030

[lmodel-file]
1 0 5 train.blm

[weight-l]
0.097
```

created in a
working dir

optimal
weights

Tuning



selected

selected + seed

Practical recipe

- **data selection**
  - use source text of seed data
  - get seed data as large and as close to test as possible
  - select data until perplexity improves
- **TM/LM adaptation**
  - use mixture LM and filled-up TM
    - more robust
    - fewer weights to optimize
- **tuning**
  - select a dev set as close to test as possible
  - use about 20K words

Software

- IRSTLM
  - www.fbk.eu/irstlm
  - www.sourceforge.net/projects/irstlm/

- MOSES
  - www.statmt.org/moses
  - www.github.com/moses-smt/mosesdecoder

- MATECAT project
  - www.matecat.com