A Greedy Decoder for Phrase-Based Statistical Machine Translation

Philippe Langlais, Alexandre Patry and Fabrizio Gotti

Dept. I.R.O.
Université de Montréal, Québec, Canada
{felipe,patryale,gottif}@iro.umontreal.ca

TMI, Skövde, September 7-9, 2007

A bit of context : WMT'06 1/3

SRC	les avantages sont déjà présents , il sont visibles et ils profitent à tous .
REF	the advantages are already there ; they are visible and everyone stands to gain .
cmu	the advantages are already present , it is visible and they benefit to all .
lcc	the benefits are already there , it is visible and they should benefit everyone .
nrc	the benefits are already present , there are already visible and they should benefit everyone .
nrc	the benefits are already present , there are already visible and they should benefit everyone .
nrc	the benefits are already present , there are visible and they benefit to all .
ntt	the advantages are already present , there are clear and they should benefit everyone .
ntt	the advantages are already present , there are visible and they benefit to all .
rali	**the advantages are already there , it is visible and they will benefit at all .**
systran	the advantages are already present , it are visible and they benefit all .
uedin	the advantages are already there , they are visible and they benefit all .
upc	the advantages are already present , are visible and they benefit everyone .
upc	the advantages are already present , it is visible and they benefit everyone .
upv	the benefits , there are clear and make use of all .
utd	the advantages are present , there are already visible and they should benefit everyone .

A bit of context : WMT'06 1/3

SRC	les avantages sont déjà présents , il sont visibles et ils profitent à tous .
REF	the advantages are already there ; they are visible and everyone stands to gain .
cmu	the advantages are already present , it is visible and they benefit to all .
lcc	the benefits are already there , it is visible and they should benefit everyone .
nrc	the benefits are already present , there are already visible and they should benefit everyone .
nrc	the benefits are already present , there are already visible and they should benefit everyone .
nrc	the benefits are already present , there are visible and they benefit to all .
ntt	the advantages are already present , there are clear and they should benefit everyone .
ntt	the advantages are already present , there are visible and they benefit to all .
rali	**the advantages are already there , it is visible and they will benefit at all .**
systran	the advantages are already present , it are visible and they benefit all .
uedin	the advantages are already there , they are visible and they benefit all .
upc	the advantages are already present , are visible and they benefit everyone .
upc	the advantages are already present , it is visible and they benefit everyone .
upv	the benefits , there are clear and make use of all .
utd	the advantages are present , there are already visible and they should benefit everyone .

A bit of context : WMT'06 2/3

SRC	ce n ' est pas seulement une question de précaution : c ' est du simple bon sens .
REF	that is not just a precaution , it is common sense .
cmu	it is not just a precautionary issue : it is of simple common sense .
lcc	it is not just a question precautionary : it is simply the right direction .
nrc	it is not just a question of caution : that of simple common sense .
nrc	it is not just a question of caution : this is the simple good sense .
nrc	it is not just a question of caution : this is the simple good sense .
ntt	this is not just a question of precautionary : it is simple common sense .
ntt	this is not just a question of precautionary : it is simply common sense .
rali	**this is not just a question of precaution is the simple good sense .**
systran	it is not only one question of precaution : it is simple good direction .
uedin	this is not only a question of caution : that is the simple good sense .
upc	this is not only a question of caution : it is a simple common sense .
upc	this is not just a question of precaution is the simple common sense .
upv	this is not just a question of caution : it is , of simple common sense .
utd	this is not just a question precautionary : it is , of simple common sense .

A bit of context : WMT'06 2/3

SRC	ce n ' est pas seulement une question de précaution : c ' est du simple bon sens .
REF	that is not just a precaution , it is common sense .
cmu	it is not just a precautionary issue : it is of simple common sense .
lcc	it is not just a question precautionary : it is simply the right direction .
nrc	it is not just a question of caution : that of simple common sense .
nrc	it is not just a question of caution : this is the simple good sense .
nrc	it is not just a question of caution : this is the simple good sense .
ntt	this is not just a question of precautionary : it is simple common sense .
ntt	this is not just a question of precautionary : it is simply common sense .
rali	**this is not just a question of precaution is the simple good sense .**
systran	it is not only one question of precaution : it is simple good direction .
uedin	this is not only a question of caution : that is the simple good sense .
upc	this is not only a question of caution : it is a simple common sense .
upc	this is not just a question of precaution is the simple common sense .
upv	this is not just a question of caution : it is , of simple common sense .
utd	this is not just a question precautionary : it is , of simple common sense .

A bit of context : WMT'06 3/3

SRC	il est certain que la déclaration complète implique à nouveau **des coûts** . c ' est l ' agriculteur qui doit , en fin de compte , supporter **les coûts** .
REF	the full declaration certainly **costs money** , and the farmer ultimately has to **foot the bill** .
cmu	there is no doubt that the full statement involves **costs** again . that is the farmer which must , at the end of the day bear the **costs** .
ntt	it is true that the statement that is the farmer who must , in the end , bear the **costs** . full means to new **costs** .
rali	it is true that the full statement implies again this is the farmer who must , ultimately , bear the **costs** . **costs**.

Several solutions

- better models

  (of course. . .)

- monotone decoding

  (faster, sometimes improves)

- enlarging the search space

  (we do not care about speed, do we ?)

The solution we considered

greedy search

Hill-climbing a given translation

Pros :

- easy, memory efficient, and often successful in search problems
- operations can be customized *(post-processing)*
- greedy search has never been evaluated within a phrase-based paradigm [Germann et al. , 2001]

Con : search space visited usually small

Algorithm

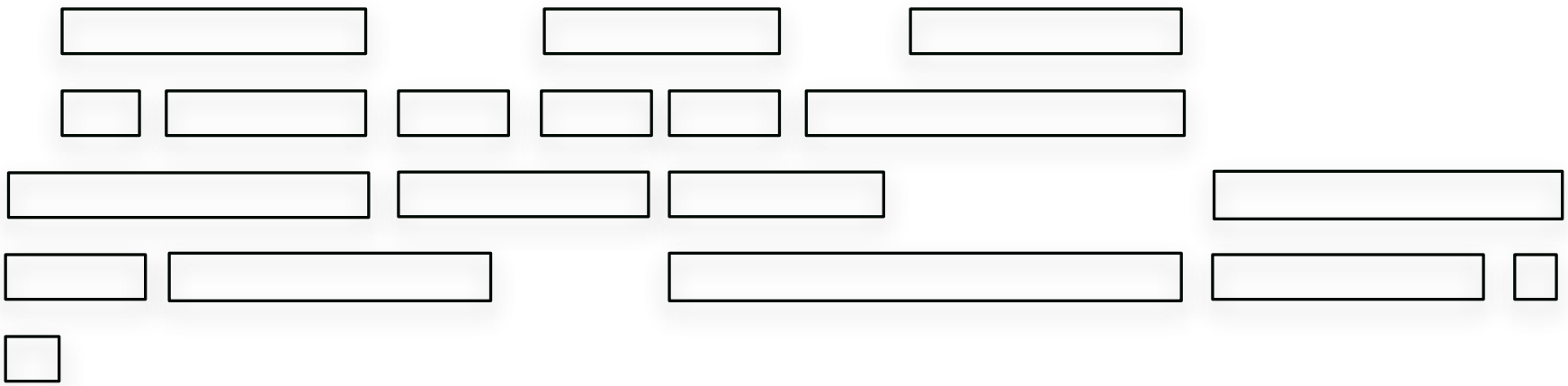
Require: *source* a sentence to translate
  current $\leftarrow$ SEED(*source*)
  loop
    s_current $\leftarrow$ SCORE(*current*)
    s $\leftarrow$ *s_current*
    for all *h* $\in$ NEIGHBORHOOD(*current*) **do**
      c $\leftarrow$ SCORE(*h*)
      if *c* $>$ *s* **then**
        s $\leftarrow$ *c*
        best $\leftarrow$ *h*
    if *s* $=$ *s_current* **then**
      return *current*
    else
      current $\leftarrow$ *best*

The Seed function

Seed the engine with either the output of :

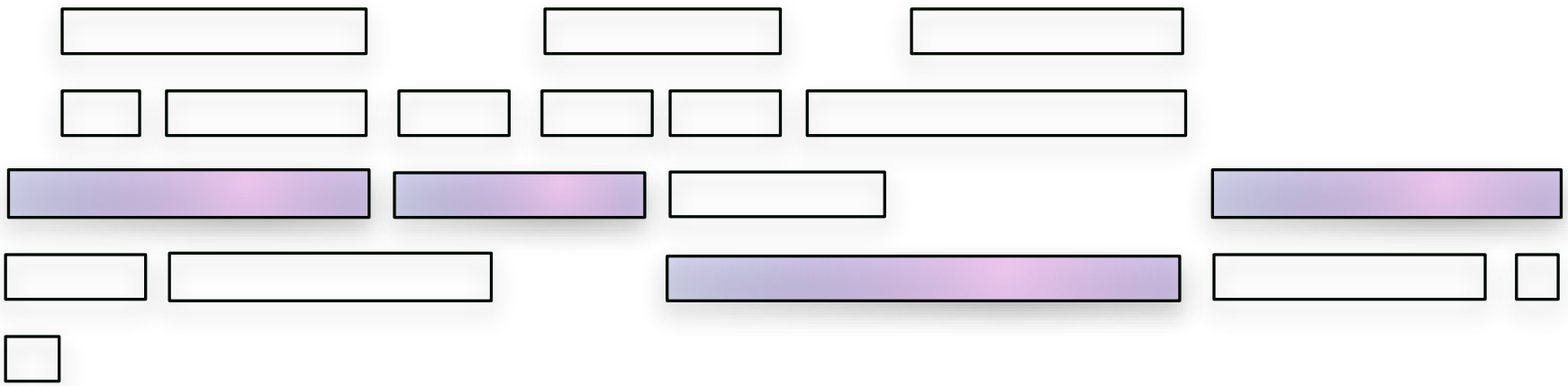
1. a DP-algorithm which selects the minimum number of phrases covering the source sentence (G-GLOSS)
2. another phrase-based engine (G-PHARAOH)

Seeding with DP-segmentation (1/3)



je les remercie tous deux pour leur formidable engagement .

Seeding with DP-segmentation (1/3)



je les remercie tous deux pour leur formidable engagement .

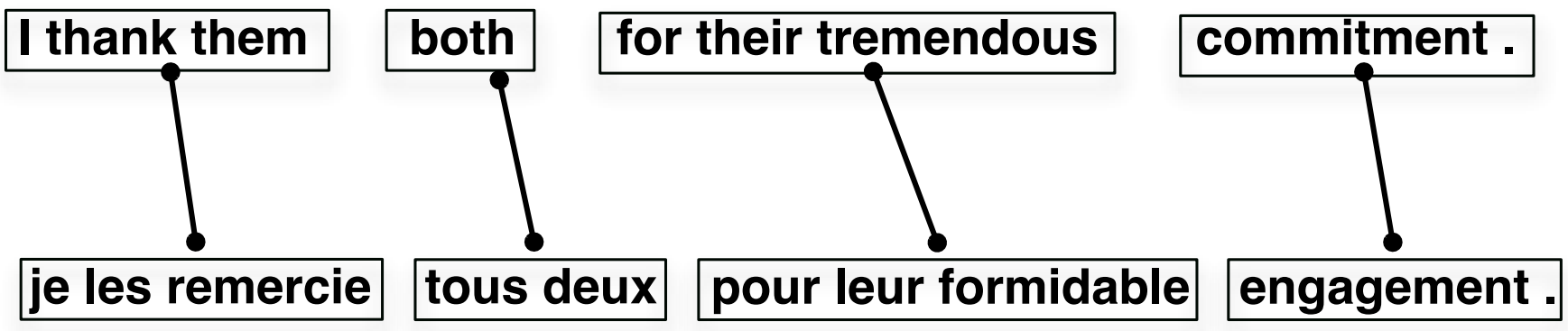
Seeding with DP-segmentation (2/3)

je les remercie → i thank them (-1.03) , i thank them (-1.5) i wish to thank them (-2.0) i would like to thank them (-2.2) i congratulate them (-2.4) i should also like to thank them (-2.6) i wish to thank (-2.7) i offer them my thanks (-2.7) i would like to thank parliament (-3.2)
tous deux → both (-1.4) both of (-1.9) , both (-2.2) both will (-2.2) , both of (-2.2) both to (-2.3) both to be (-2.3) which both (-2.3) both of which (-2.4) they both (-2.4)
pour leur formidable → for their tremendous (-1.33) on their comprehensive (-2.6) them on their comprehensive (-2.9)
engagement . → commitment . (-0.3) engagement . (-1.1) undertaking . (-1.2) involvement . (-1.4) pledge . (-1.5) dedication . (-1.5) commitments . (-1.5) committed . (-1.7) promise . (-1.8) obligation . (-2.0)

Seeding with DP-segmentation (2/3)

je les remercie → i thank them (-1.03) , i thank them (-1.5) i wish to thank them (-2.0) i would like to thank them (-2.2) i congratulate them (-2.4) i should also like to thank them (-2.6) i wish to thank (-2.7) i offer them my thanks (-2.7) i would like to thank parliament (-3.2)
tous deux → both (-1.4) both of (-1.9) , both (-2.2) both will (-2.2) , both of (-2.2) both to (-2.3) both to be (-2.3) which both (-2.3) both of which (-2.4) they both (-2.4)
pour leur formidable → for their tremendous (-1.33) on their comprehensive (-2.6) them on their comprehensive (-2.9)
engagement . → commitment . (-0.3) engagement . (-1.1) undertaking . (-1.2) involvement . (-1.4) pledge . (-1.5) dedication . (-1.5) commitments . (-1.5) committed . (-1.7) promise . (-1.8) obligation . (-2.0)

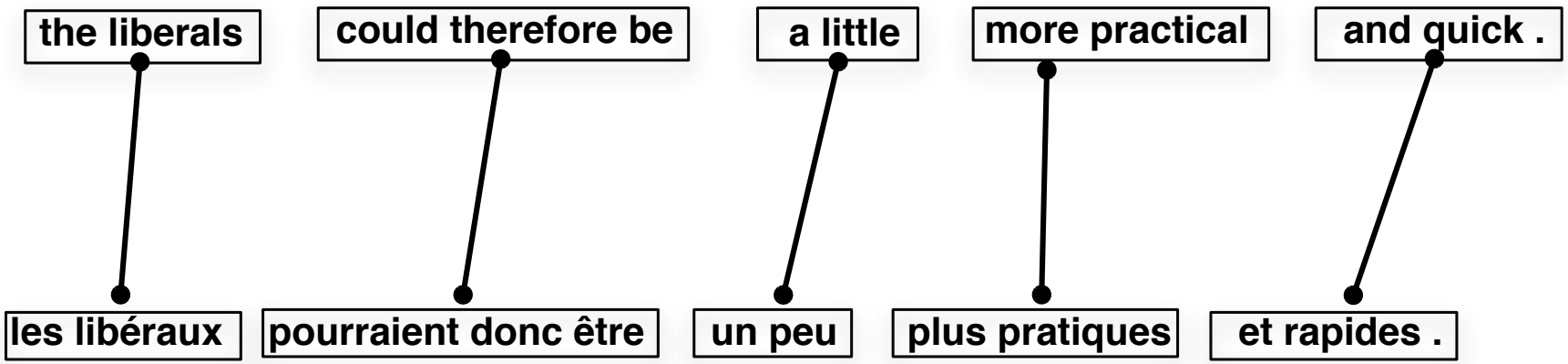
Seeding with DP-segmentation (3/3)



Seeding with Pharaoh

By using option `-t` :

SRC	les libéraux pourraient donc être un peu plus pratiques et rapides .
TRA	the liberals |0.104264|0|1| could therefore be |0.0398264|2|4| a little |0.19357|5|6| more practical |0.143042|7|8| and quick . |0.0447256|9|11|



The Scoring function

The very same function embedded in PHARAOH :

$$
\begin{aligned}
Score(e, f) = \quad & \lambda_{lm} \quad \log p_{lm}(f) && + \\
& \sum_i \quad \lambda_{tm}^{(i)} \log p_{tm}^{(i)}(f|e) && - \\
& \lambda_w \quad |f| && - \\
& \lambda_d \quad p_d(e, f)
\end{aligned}
$$

The neighborhood function

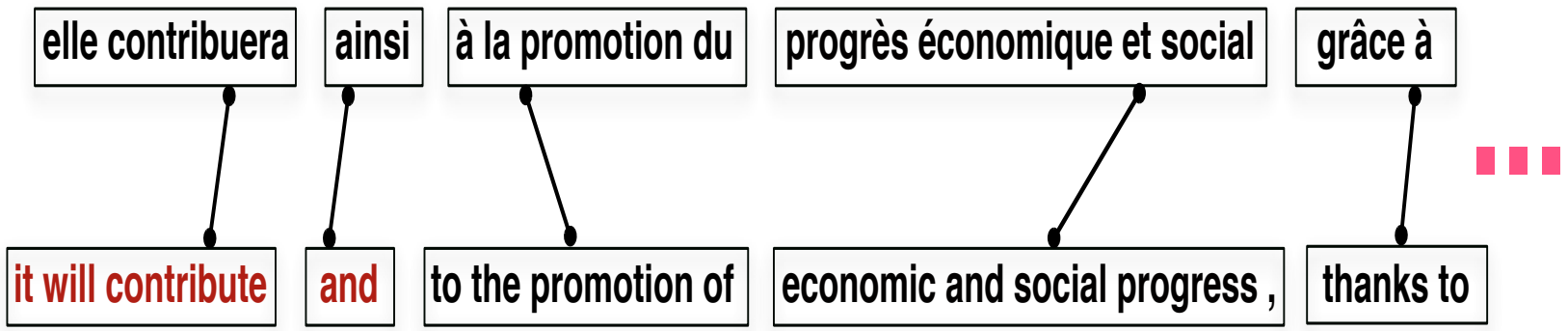
- only 5 operations encoded (+ variants)
  (first try. . .)

- many more possible
  (including inserting/deleting words)

Illustrated on 3 excerpts of translations sessions.

SWAP SPLIT MERGE MOVE REPLACE

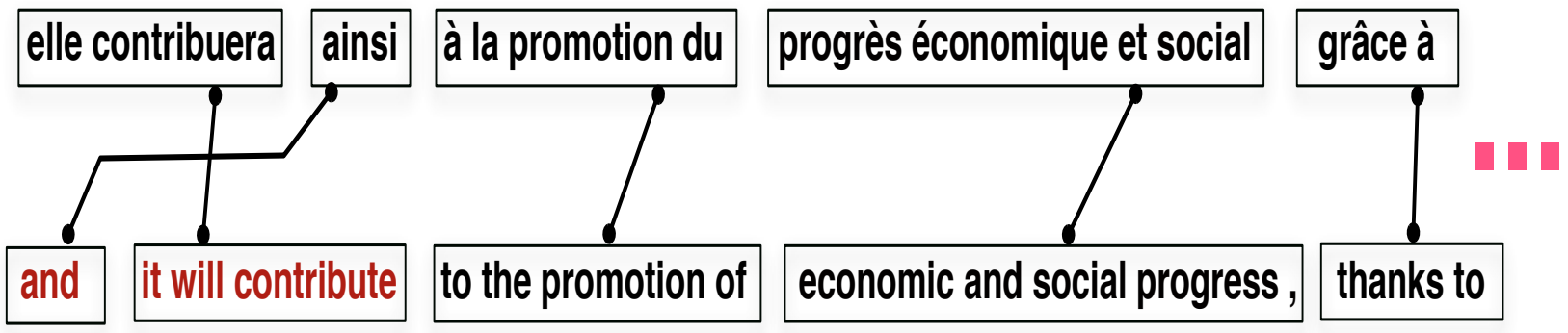
SRC : **elle contribuera ainsi à** la promotion du progrès économique et social grâce à un niveau d ' emploi élevé .



SWAP [elle contribuera ↔ it will contribute]
with [ainsi ↔ and]
STEP-3 -15.1609 → -14.6041

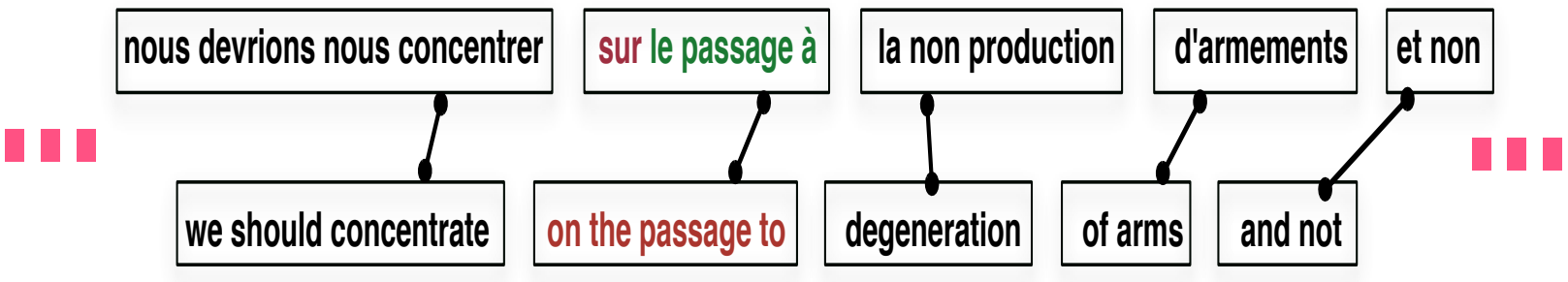
SWAP SPLIT MERGE MOVE REPLACE

SRC : **elle contribuera ainsi à** la promotion du progrès économique et social grâce à un niveau d ' emploi élevé .



SWAP **SPLIT** MERGE MOVE REPLACE

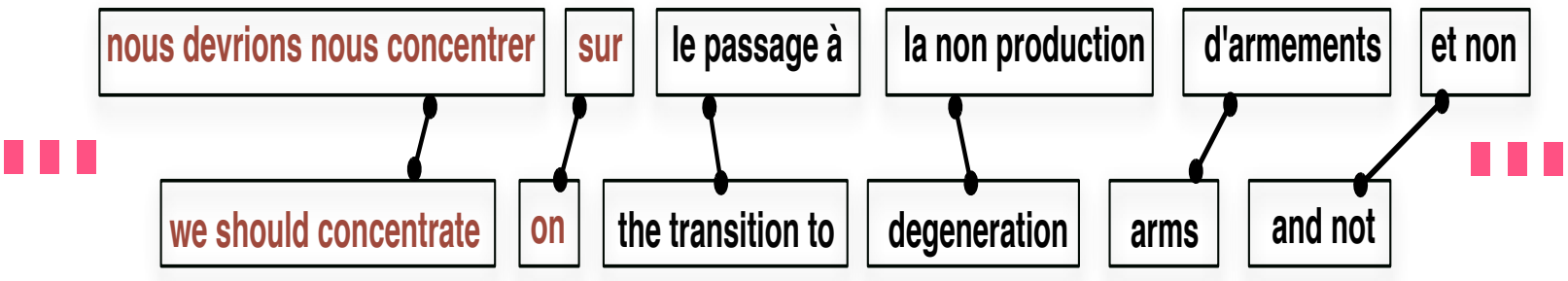
SRC : nous devrions nous concentrer **sur le passage à** la non-production d ' armements et non sur la manière dont nous allons assurer notre compétitivité par rapport aux autres pays du monde qui produisent des armements .



SPLIT	[sur le passage à ↔ on the passage to]
into	[sur ↔ on] [le passage à ↔ the transition to]
STEP-4	-35.7871 → -35.5256

SWAP SPLIT MERGE MOVE REPLACE

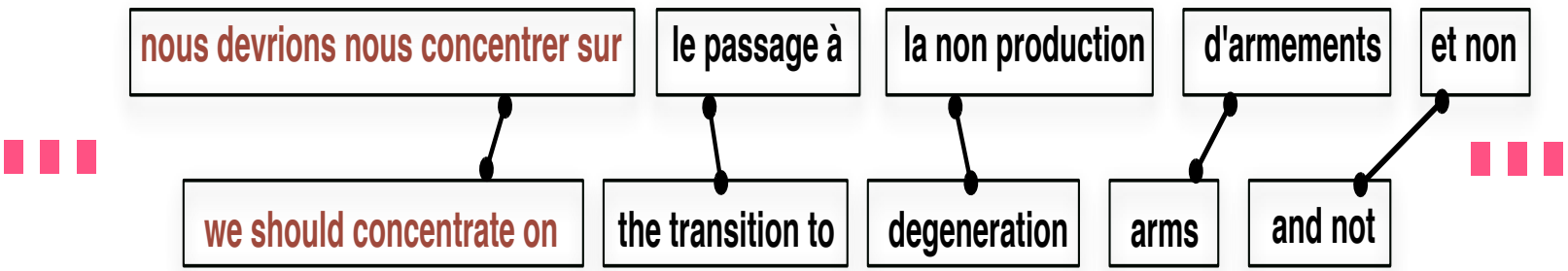
SRC : **nous devrions nous concentrer sur** le passage à la non-production d ' armements et non sur la manière dont nous allons assurer notre compétitivité par rapport aux autres pays du monde qui produisent des armements .



MERGE	[nous devrions nous concentrer] [sur]
into	[we should concentrate on]
STEP-5	-35.5256 → -35.3209

SWAP SPLIT MERGE MOVE REPLACE

SRC : nous devrions nous concentrer sur le passage à la non-production d ' armements et non sur la manière dont nous allons assurer notre compétitivité par rapport aux autres pays du monde qui produisent des armements .



SWAP SPLIT MERGE MOVE REPLACE

SRC : le groupe csu au parlement européen **se réjouit que** le présent projet de charte des droits fondamentaux rassemble et rende visibles les droits fondamentaux dont disposent les citoyens vis-à-vis des organes et institutions de l ' ue .

SEED the csu group in the european parliament **welcomes** the draft charter of fundamental rights lumps together and make visible the fundamental rights enjoyed by the citizens towards the eu institutions and bodies **that** . *(-43.8823)*

MOVE [se réjouit ↔ welcomes] [que ↔ that]
STEP-1 -43.8823 → -39.7283

SWAP SPLIT MERGE MOVE REPLACE

SRC : le groupe csu au parlement européen se réjouit que **le présent projet de** charte des droits fondamentaux rassemble et rende visibles les droits fondamentaux dont disposent les citoyens vis-à-vis des organes et institutions de l ' ue .

STEP-1 the csu group in the european parliament welcomes that **the draft** charter of fundamental rights lumps together and make visible the fundamental rights enjoyed by the citizens towards the eu institutions and bodies . *(-39.7283)*

REPLACE	[le présent projet de ↔ the draft]
by	[le présent projet de ↔ the present draft]
STEP-2	-39.7283 → -39.3657

SWAP SPLIT MERGE MOVE REPLACE

SRC : le groupe csu au parlement européen se réjouit que le présent projet de charte des droits fondamentaux **rassemble** et rende visibles les droits fondamentaux dont disposent les citoyens vis-à-vis des organes et institutions de l ' ue .

STEP-2 the csu group in the european parliament welcomes that the present draft charter of fundamental rights **lumps together** and make visible the fundamental rights enjoyed by the citizens towards the eu institutions and bodies . *(-39.3657)*

REPLACE	[rassemble ↔ lumps together]
by	[rassemble ↔ brings together]
STEP-3	-39.3657 → -39.06

SWAP SPLIT MERGE MOVE REPLACE

REF : the csu 's europe group welcomes the tabling of the final draft of the charter of fundamental rights because it summarises and makes visible the fundamental rights which the public are entitled to in respect of the institutions and bodies of the eu .

SEED : the csu group in the european parliament **welcomes the** draft charter of fundamental rights **lumps** together and make visible the fundamental rights enjoyed by the citizens towards the eu institutions and bodies **that** . *(-43.8823)*

STEP-3 the csu group in the european parliament **welcomes that the present** draft charter of fundamental rights **brings** together and make visible the fundamental rights enjoyed by the citizens towards the eu institutions and bodies . *(-39.06)*

Cascading translation engines

not a new idea

- [Berger & al. (1994)] **word-based** greedy search seeded with a word-based engine (Candide)
  no evaluation
- [Marcu (2001)] **word-based** greedy search seeded with a phrase-based translation memory
  500,000 Hansard sentences for training, 505 for testing
- [Watanabe & Sumita (2003)] **word-based** greedy search seeded with a sentence-based translation memory
  $\sim$ *150,000 BTEC sentences for training, 5,000 for testing*

Main difference here : phrase-based greedy search, evaluation on the WMT'06 shared-task

Protocol

- WMT'06 : German, French, Spanish $\leftrightarrow$ English
  - $\sim$ 700,000 pairs of sentences for training
  - 500 pairs for tuning
  - 2,000 for monitoring *(dry-run)*
  - 3,034 for testing *(in- and out-domain data)*
- Phrase-based engine made out of the scripts provided by the organizers
  - phrases up to 7 words long
  - trigram language model with SRILM
  - tuning with MERT
  - decoding with PHARAOH *(built-in default search options)*
- BLEU and WER + bootstrap resampling
  - 1,000 samples of 700 sentences each, 99% conf. level

Results

dry-run

Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	55.12	30.16	51.47	29.23
G-GLOSS		54.10	29.30	51.01	28.41
G-PHARAOH		**53.62**	**30.64**	**50.37**	**29.62**
PHARAOH	es	55.04	28.17	50.97	29.94
G-GLOSS		**53.87**	27.38	50.69	28.99
G-PHARAOH		**53.14**	**28.72**	**50.04**	**30.30**
PHARAOH	de	62.38	17.32	60.12	24.54
G-GLOSS		62.85	16.37	**57.55**	23.44
G-PHARAOH		**61.85**	**17.51**	**58.33**	24.97

Results

dry-run

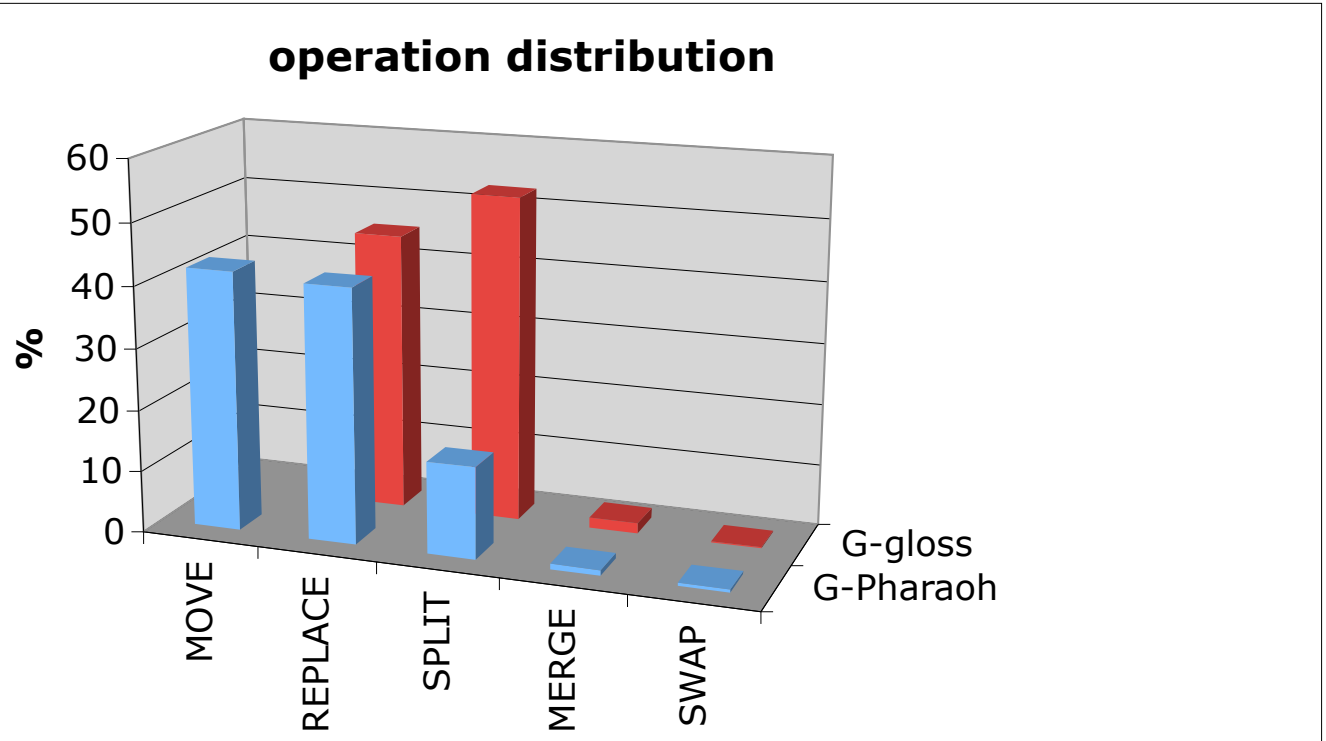
Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	55.12	30.16	51.47	29.23
G-GLOSS		54.10	29.30	51.01	28.41
G-PHARAOH		53.62	30.64	50.37	29.62
PHARAOH	es	55.04	28.17	50.97	29.94
G-GLOSS		**53.87**	27.38	50.69	28.99
G-PHARAOH		53.14	28.72	50.04	30.30
PHARAOH	de	62.38	17.32	60.12	24.54
G-GLOSS		62.85	16.37	**57.55**	23.44
G-PHARAOH		61.85	17.51	58.33	24.97

Results

dry-run

Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	55.12	30.16	51.47	29.23
G-GLOSS		54.10	29.30	51.01	28.41
G-PHARAOH		**53.62**	**30.64**	**50.37**	**29.62**
PHARAOH	es	55.04	28.17	50.97	29.94
G-GLOSS		**53.87**	27.38	50.69	28.99
G-PHARAOH		**53.14**	**28.72**	**50.04**	**30.30**
PHARAOH	de	62.38	17.32	60.12	24.54
G-GLOSS		62.85	16.37	**57.55**	23.44
G-PHARAOH		**61.85**	**17.51**	**58.33**	24.97

Dry-run, fr→ en



	MOVE	REPLACE	SPLIT	MERGE	SWAP
G-Pharaoh	42,2	41,3	14,9	0,9	0,5
G-gloss		45,1	52,8	1,7	0,2

On time issue

Time[1] for translating 1 000 sentences

PHARAOH	78	min.
G-GLOSS*	9	min.
G-PHARAOH*	$\sim$ 4	min.

* **VERY** crude implementation ! ! !

[1] Pentium computer clocked at 3 GHz

Adding a Reversed Language Model

- $p(t_1^T) \approx \prod_{i=1}^{T} p(t_i | t_{i+1} \dots t_{i+n-1})$
- difficult to plug in a standard beam-search decoder

Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	55.12	30.16	51.47	29.23
G-PHARAOH	fr	**53.62**	**30.64**	**50.37**	**29.62**
G-LMREV	fr	53.65	30.85	50.30	29.70
PHARAOH	es	55.04	28.17	50.97	29.94
G-PHARAOH	es	**53.14**	**28.72**	**50.04**	**30.30**
G-LMREV	es	52.37	29.31	50.05	30.33
PHARAOH	de	62.38	17.32	60.12	24.54
G-PHARAOH	de	**61.85**	**17.51**	**58.33**	24.97
G-LMREV	de	61.85	17.57	57.99	25.20

Adding a Reversed Language Model

- $p(t_1^T) \approx \prod_{i=1}^{T} p(t_i | t_{i+1} \dots t_{i+n-1})$
- difficult to plug in a standard beam-search decoder

Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	55.12	30.16	51.47	29.23
G-PHARAOH	fr	**53.62**	**30.64**	**50.37**	**29.62**
G-LMREV	fr	**53.65**	**30.85**	**50.30**	**29.70**
PHARAOH	es	55.04	28.17	50.97	29.94
G-PHARAOH	es	**53.14**	**28.72**	**50.04**	**30.30**
G-LMREV	es	**52.37**	**29.31**	**50.05**	**30.33**
PHARAOH	de	62.38	17.32	60.12	24.54
G-PHARAOH	de	**61.85**	**17.51**	**58.33**	24.97
G-LMREV	de	**61.85**	**17.57**	**57.99**	**25.20**

In-domain test data (2,000 sentences)

Systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
PHARAOH	fr	54.85	30.90	51.69	29.96
G-GLOSS	fr	54.27	29.83	50.93	29.13
G-PHARAOH	fr	**53.38**	**31.42**	**50.46**	**30.27**
G-BEAM5	fr	**53.46**	31.26	**50.40**	30.13
G+BEAM5	fr	**53.43**	**31.28**	**50.36**	30.17
G-LMREV	fr	**53.49**	**31.52**	**50.48**	**30.25**
PHARAOH	es	54.23	29.64	51.04	30.54
G-GLOSS	es	**53.22**	28.99	50.77	29.67
G-PHARAOH	es	**52.77**	**30.14**	**50.02**	**30.87**
G-BEAM5	es	**52.61**	30.24	**50.12**	**30.89**
G+BEAM5	es	**52.61**	**30.25**	**50.11**	**30.93**
G-LMREV	es	**52.67**	**29.79**	**50.07**	30.84

Recap

- greedy alone ;-(
- cascading greedy after PHARAOH :-)
- even if BLEU is not improved, better scores are found by greedy search...

search errors help sometimes...

Future Work

- analyzing why DP beam-search misses some targets
- coding more operations
  - at the very least, word insertion
  - global operations (modality, negation, etc.)
- comparing different ways to trade speed/quality :
  - lattice-based monotone decoding
  - local search
  - smartness in beam-search decoding [Moore and Quirk, 2007]

Conclusion

"Can the dynamic programming be adjusted – what happens when the Pharaoh default beam parameters are widened ?"

Our answer : Why not use local-search anyway ! ?

- it is cheap *(one day of coding)*
- it does not hurt *(might even improve)*
- it is fast and memory efficient
- it is a standard practice in search problems [Russell & Norvig, 1995]

ASSERT(

REPLACE(
SWAP(
SPLIT(
GLOSS(avez vous des questions ?),
avez vous, have, you),
avez,vous),
you, do you)

==

Do you have questions ?

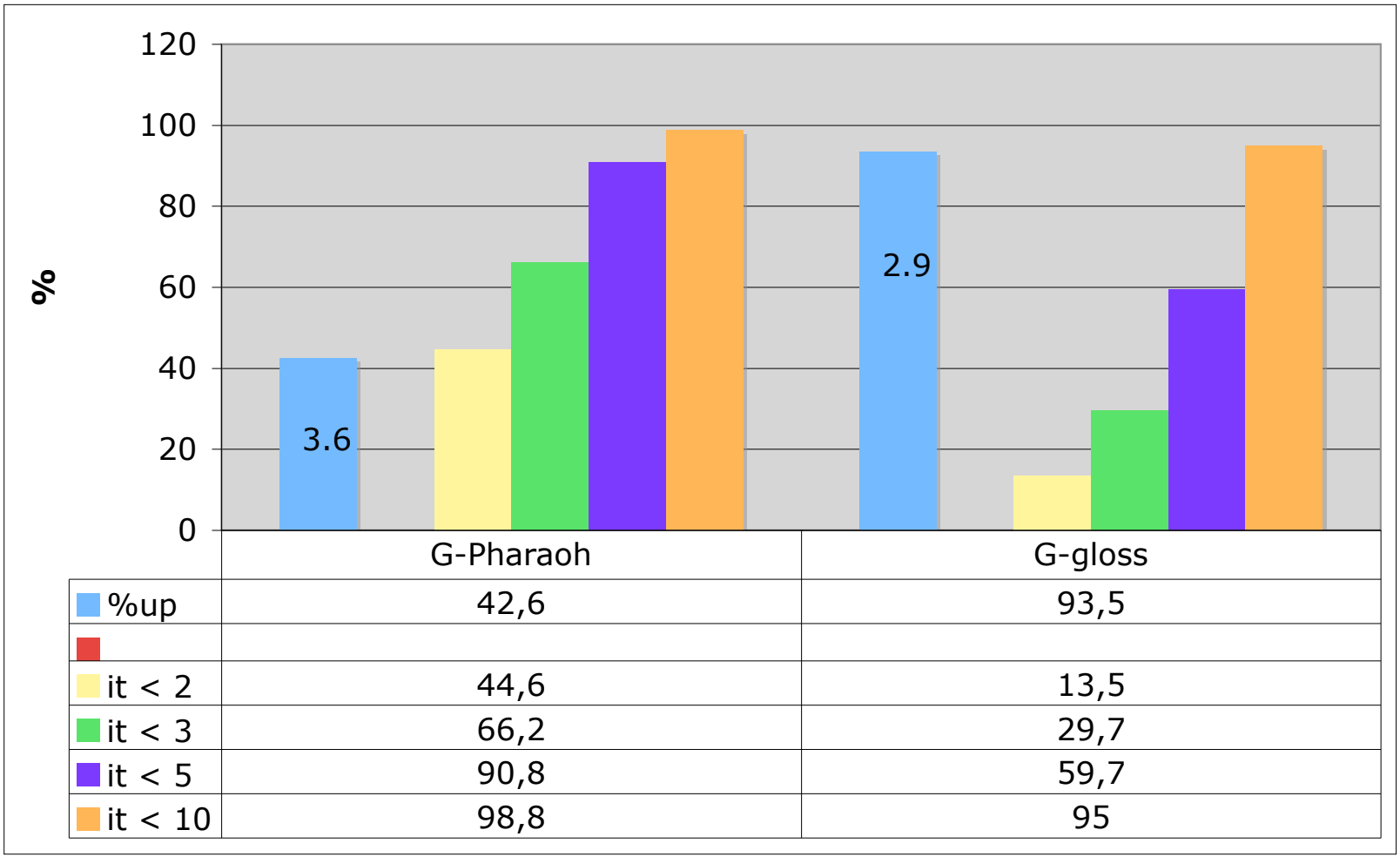
)

Increasing the search space

Dry-run, 1,000 sentences, fr→en

	Pharaoh			g-pharaoh		
stack	WER	BLEU	time	WER	BLEU	time
50	51.82	29.24	40min.	50.26	29.65	<5 min.
100	51.46	29.23	1h. 20min.	50.32	29.62	<5 min.
200	51.15	29.44	2h. 40min.	50.18	29.69	<5 min.
300	51.10	29.50	3h. 45min.	50.15	29.73	<5 min.
500	50.86	29.51	6h. 15min.	50.11	29.74	<5 min.
1000	50.64	29.54	12h. 15min.	50.04	29.74	<5 min.

Dry-run, fr→ en



	G-Pharaoh	G-gloss
%up	42,6	93,5
it < 2	44,6	13,5
it < 3	66,2	29,7
it < 5	90,8	59,7
it < 10	98,8	95

Reducing distortion

Dry-run, 1,000 sentences, fr→en

systems	L	en→L WER	en→L BLEU	L→en WER	L→en BLEU
mono	fr	-0.34	+0.15	-0.39	+0.40
dl1	fr	-1.05	+0.75	-1.55	+0.86
dl2	fr	-0.35	+0.18	-0.57	+0.44
dl3	fr	-0.06	+0.17	-0.59	+0.33
dl5	fr	-0.13	+0.07	-0.61	+0.45
PHARAOH	fr	-1.33	+2.15	-1.59	+2.71
mono	es	-0.46	+0.17	-0.10	+0.12
dl1	es	-1.18	+0.70	-1.37	+0.78
dl2	es	-0.27	+0.17	-0.35	+0.41
dl3	es	-0.09	+0.10	-0.27	+0.13
dl5	es	-0.36	+0.10	-0.41	+0.29
PHARAOH	es	-1.20	+1.81	-1.95	+3.45

A beam-search version of feGreedy

- Keeping k-best hypotheses instead of one

$\hookrightarrow$ no improvement in BLEU or WER, but :

- 20% of the translations produced by G-BEAM are different from those produced by G-PHARAOH
- 87% of those $\neq$ translations have a higher score
- If we increase the beam width, we decrease the number of downgraded translations