A multi-genre SMT system for Arabic to French

Saša Hasan and Hermann Ney

LREC 2008
Marrakech, Morocco – May 29, 2008

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Overview

- **Project TRAMES:**
  Traduction Automatique par des Méthodes Statistiques
- **Goal:**
  online system for translation of Arabic to French
- **Development over 3-year period (2005–2007):**
  - **corpus gathering**
  - **preprocessing pipeline**
  - **phrase-based SMT module (decoder)**
  - **fine-tuning for different genres**
  - **software engineering for “real-time” performance**

Rough processing pipeline (1)

- Data acquisition
  - no parallel corpora initially available for Arabic-French
  - gather data from the web (intl. organizations, news agencies, journals)
  - main data resource: Official Document System of the United Nations (ODS)
- Corpus creation
  - document and sentence alignment
  - preprocessing: tokenization, Arabic word segmentation
- Training the models
  - word alignments
  - phrase extraction
  - language modeling

Rough processing pipeline (2)

- Generation of translations (search/decoding)
  - phrase-based decoder using log-linear combination of models
  - dynamic programming beam search
  - tune parameters on development set using MERT
- Experiments
  - evaluation of the system using automatic evaluation measures
  - compare translation output to a set of reference translations
    - BLEU: *n*-gram precision w/ brevity penalty
    - TER: string edit distance allowing for block movements

Corpus creation

- **Document alignment as is (from web structure)**
- **Sentence alignment using sentence-length model and refinements from IBM model 1 probabilities**
- **Preprocessing:**
  - **tokenization and categorization for numbers, months and URLs**
  - **text normalization: remove diacritics**
  - **word segmentation:**
    prefix and suffix splitting based on finite-state automaton
  - **example:**

المَدرَسَةُ	⇒	المدرسة		
the school				
وَالمدرسة	⇒	مدرسة	ال	و
		school	**the**	**and**
مدرستهم	⇒	هم	مدرسة	
		their	**school**	

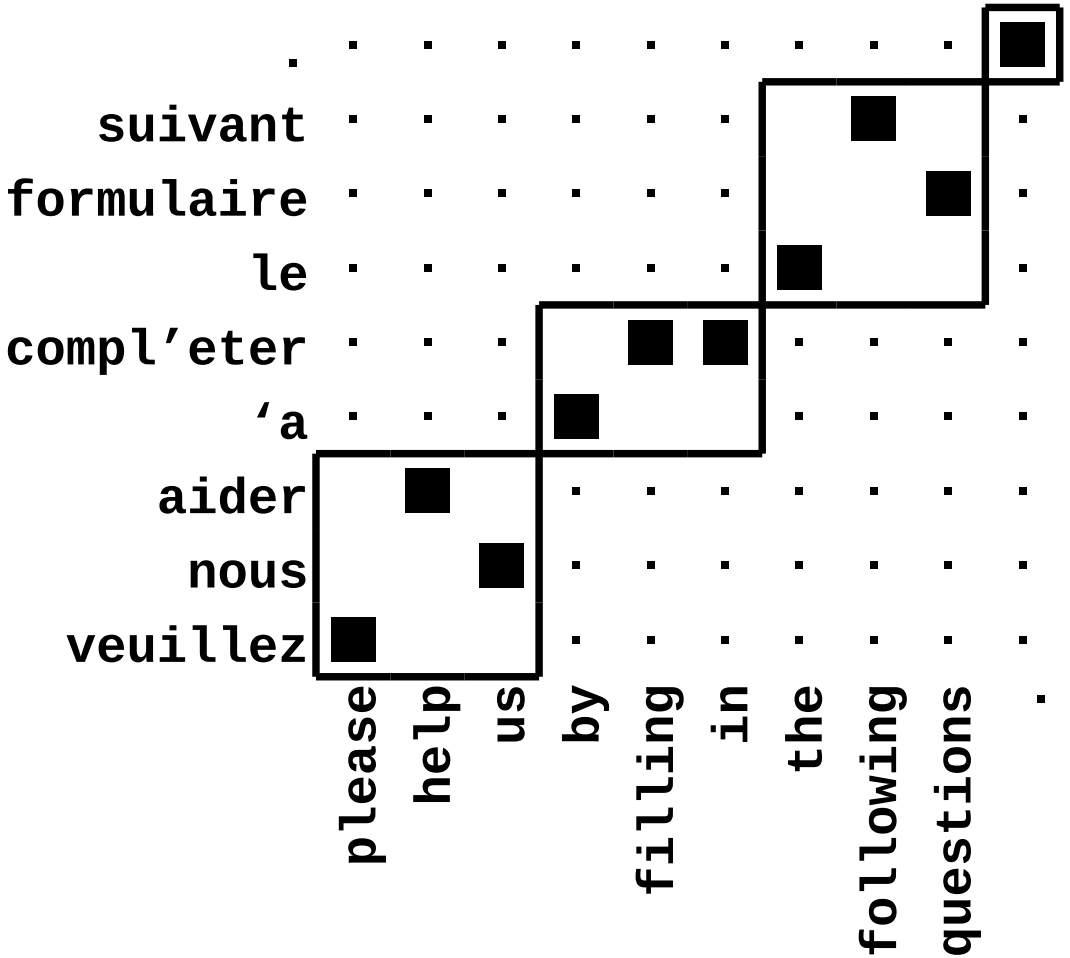
Corpus statistics

Corpus extracted from UN documents / Amnesty Int. / Le Monde Diplomatique:

	2005 system		2007 system	
	Arabic	French	Arabic	French
Doc. pairs	62K		74K	
Sent. pairs	4.7M		6.6M	
Run. words	108.1M	104.8M	151.3M	180.2M
Vocabulary	245K	288K	427K	301K

- Important data update from BN radio and TV transcripts:
  - Orient, Qatar, BBC, Alarabiya, Aljazeera, Alalam
  - 250 audio documents consisting of 90 hours radio and TV broadcasts
  - 21K sentences with 585K running words of domain-specific material for the audio domain

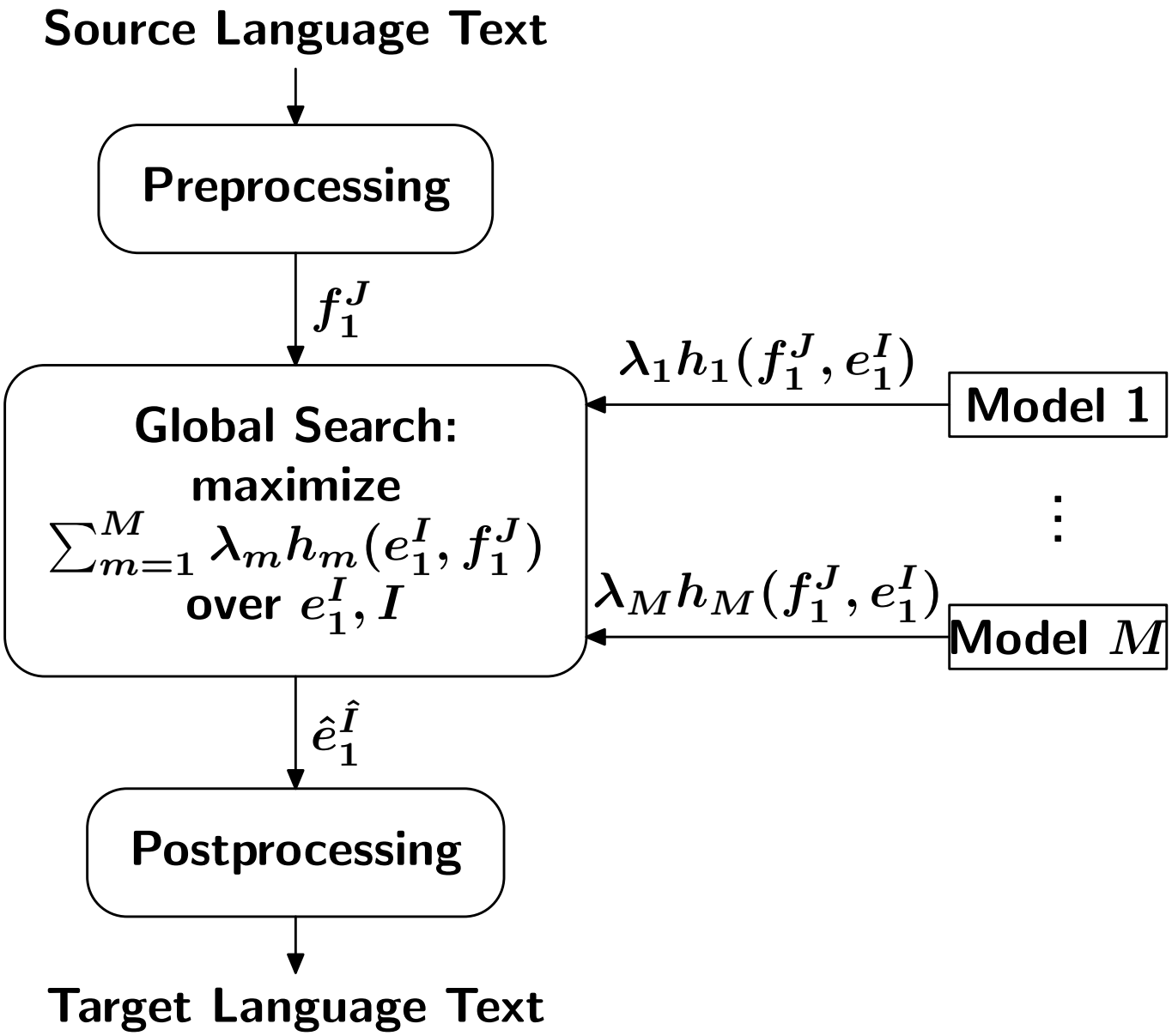
Training and Generation (1)



Idea:

1. **Segment source sentence into phrases**
2. **Translate each phrase**
3. **Concatenate these phrase translations**

Training and Generation (2)



Evaluation: progress over time

	1st sys 2005	2nd sys 2006	+BN-LM	3rd sys 2007
CESTA run2	40.8	42.9	43.8	44.8
Arabic BN				
text setting	20.9	29.7	-	34.4
audio setting	-	34.4	37.6	41.1

- System was tuned on held-out development sets
- Results shown are all on blind test sets:
  - text domain: CESTA run2 evaluation data
  - audio domain: Arabic BN transcripts from TV/radio
- Observations:
  - adding BN transcripts to the system significantly boosts performance on audio
  - genre-specific tuning makes a difference

Evaluation: comparison to Moses

	BLEU [%]	TER [%]	Translation speed [words/sec]
CESTA run2			
Moses	42.2	52.25	14.2
TRAMES	43.4	51.30	222.0
Arabic BN			
Moses	39.5	53.37	18.6
TRAMES	40.0	52.93	249.3

- Freely available: open-source phrase-based decoder Moses
- Models / search concept similar to RWTH's decoder
- Fair comparison: table shows experiments for the same training data and similar pruning parameters (histogram size 200)
- Result: TRAMES system is up to 16 times faster with up to 250 words/sec

Examples: text setting

- **Arabic source sentence:**
  ويتم التركيز على الوقاية من انتقال هذا المرض من الأم إلى الطفل واتخاذ نهج للنهوض بالوعي العام بين الشباب.
- **French translation, system update in 2005:**
  et met l'accent sur la prévention ___ de cette maladie de la mère à l'enfant et ___ une démarche pour la promotion de la sensibilisation du public chez les jeunes.
- **French translation, system update in 2006:**
  L'accent est mis sur la prévention de la transmission ___ de la mère à l'enfant et une approche pour la promotion de la sensibilisation du public chez les jeunes.
- **French translation, system update in 2007:**
  L'accent est mis sur la prévention de la transmission de la maladie de la mère à l'enfant et une approche pour promouvoir une prise de conscience parmi les jeunes.
- **French reference translation (1/4):**
  L'accent est mis sur la prévention de la transmission de cette maladie de la mère à l'enfant et l'adoption de la démarche de la généralisation de la prise de conscience parmi les jeunes.

Examples: audio setting

- **Arabic source**

  رياض محمد رصد ردود الشارع الإيراني حيال محاكمة صدام ووافانا بالتقرير التالي.

- **French sys1 2005**

  Riyad Mohammed suivi réponses la rue des UNK_إيراني pour juger Saddam et UNK_وافانا du rapport UNK_التالي .

- **French sys2 2006**

  Riad Mohamad de suivre les mesures prises par la rue iranienne par juger Saddam et nous a fait parvenir le rapport suivant.

- **French sys3 2007**

  Riad Mohamad suivi de la réponse de la rue iranienne envers le procès de Saddam et nous a fait parvenir le rapport suivant.

- **French reference translation**

  Riad Mohamed a scruté les réactions dans la rue iranienne au sujet du procès de Saddam et nous a préparé le reportage suivant.

Conclusions

- Presented a state-of-the-art SMT system for Arabic-to-French
- Multi-genre capability:
  - newswire (text domain)
  - broadcast news transcripts (audio domain)
- Real-time translation speeds of up to 250 words/sec
- Favorable performance:
  - BLEU 44.8% on text input
  - BLEU 41.1% on audio transcripts

Outlook:

- Further system updates with additional data
- Additional genres, e.g. web texts (e.g. weblogs, news groups)
- On-the-fly genre determination using text classification

Thank you for your attention

Saša Hasan

hasan@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

Test sets

Blind test sets:

- **CESTA run2 for text**
- **Arabic BN for audio setting**

	Text setting		Audio setting	
	Arabic	French	Arabic	French
Doc. pairs	30		7	
Sentences	824	3 296 (4x)	466	1 864 (4x)
Run. words	22 045	102 087	16 847	91 557
Vocabulary	4 441	6 335	5 952	6 943
OOV rate	0.40%	-	1.1%	-