Machine Translation

Hermann Ney

RWTH Aachen University

Aachen, Germany

Teams in Machine Translation

task definition: source language → target language

- **input: text (written language)**
- **input: speech (spoken language)**

MT partners:

- **LIMSI: CNRS, Paris**
- **KIT: Karlsruhe Institute of Technology, Germany**
- **RWTH: RWTH Aachen University, Germany**
- **Systran Paris**

principle approach:

- **statistical approach: requires performance measure**
  - **TER: edit operations + block moves**
  - **BLEU: n-gram accuracy with brevity penalty**
- **periodic evaluations: measure progress**

Statistical Approach: Why?

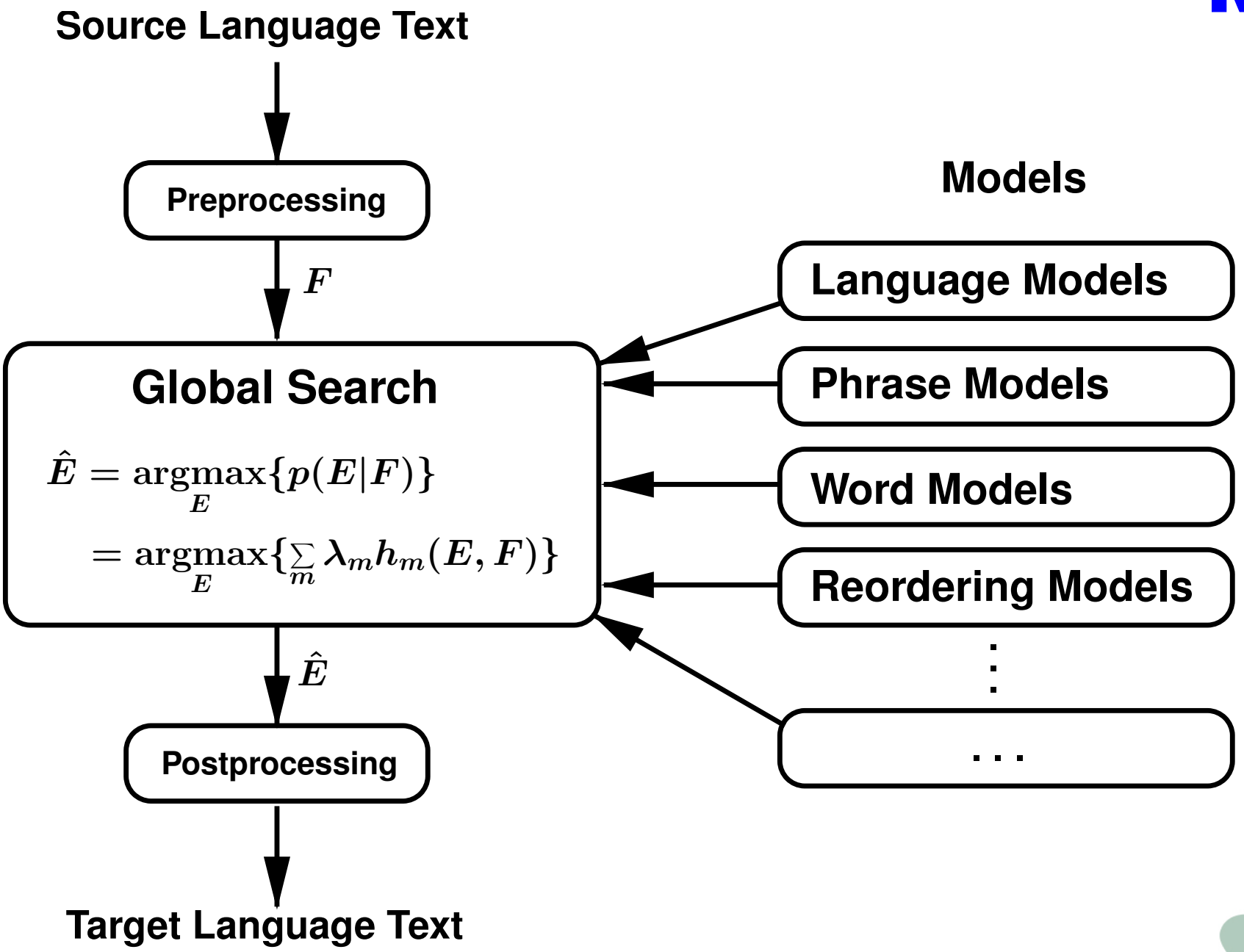
similar problem: speech recognition

machine translation: we need decisions along various dimensions:

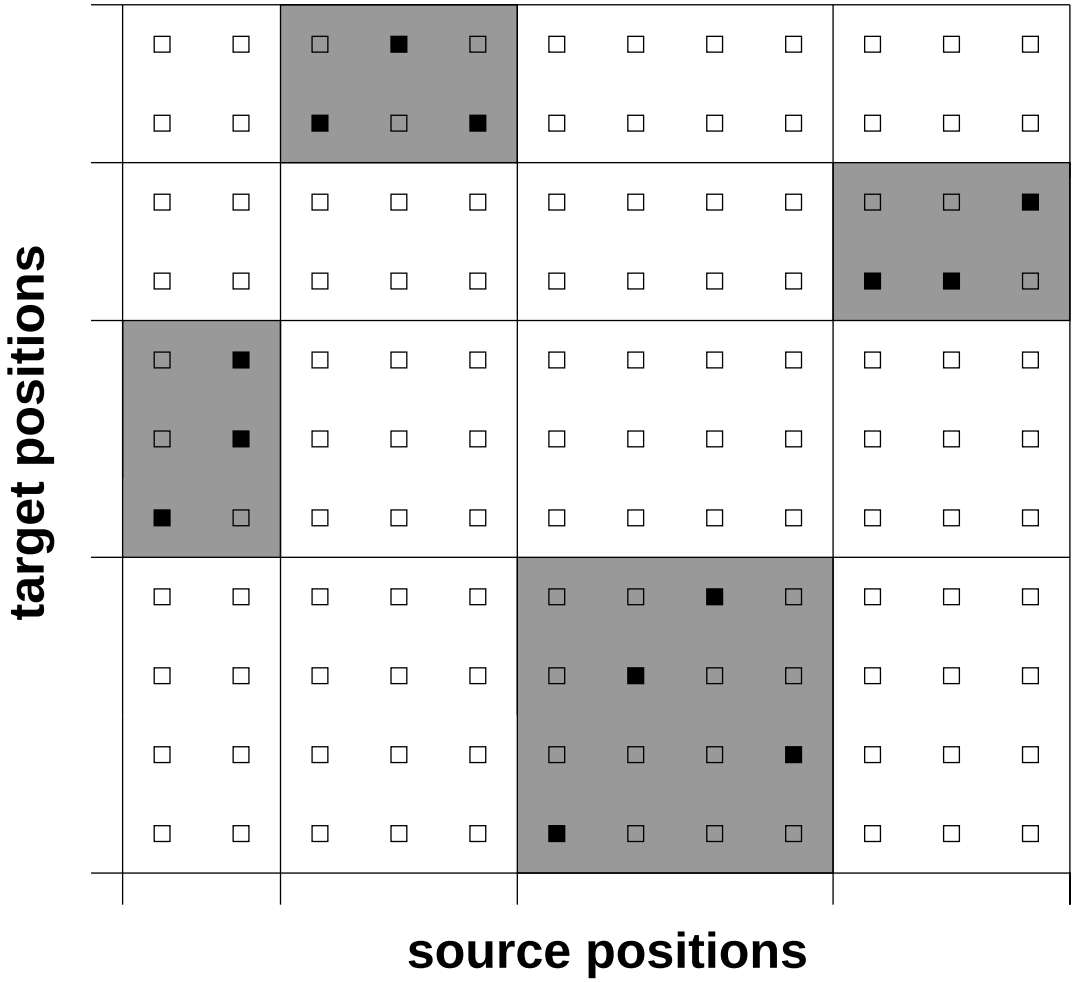
- **select the right target word**
- **select the position for the target word**
- **make sure the resulting target sentence is well formed**

advantages of statistical approach:

- **translation models can be learnt from bilingual examples**
- **interdependence of decisions is handled automatically**


Source Language Text
Preprocessing
F
Global Search
$\hat{E} = \underset{E}{\operatorname{argmax}}\{p(E|F)\}$
$= \underset{E}{\operatorname{argmax}}\{\sum_m \lambda_m h_m(E,F)\}$
$\hat{E}$
Postprocessing
Target Language Text
Models
Language Models
Phrase Models
Word Models
Reordering Models
. . .

From Words to Phrases



bilingual phrases:

- **training**
- **search/decoding**

Research Topics

- translation models:
  - hierarchical phrases
  - syntax-based models
  - re-ordering models
  - discriminative training
  - extended lexicon models

- neural networks for translation models and language models
- domain adaptation
- efficiency issues in decoding/search

Evaluation Campaign: WMT 2012

WMT (workshop on MT):
– most important evaluation for European languages
– domain: (text) news

	Fr →En		En → Fr	
	position	BLEU[%]	position	BLEU[%]
KIT	3.	31.2	3.	29.6
LIMSI	1.	31.5	1.	29.9
RWTH	6.	29.6	2.	29.6

	Ge→En		En→Ge	
	position	BLEU[%]	position	BLEU[%]
QUAERO joint subm.	1.	25.7	-	-
KIT	3.	24.5	1.	17.3
LIMSI	5.	23.9	2.	17.0
RWTH	2.	24.7	4.	16.8

Evaluation Campaign: WMT 2013

	Fr→En		En→Fr	
	position	BLEU[%]	position	BLEU[%]
KIT	3.	32.3	3.	30.6
LIMSI	2.	32.5	2.	30.8
RWTH	5.	30.8	4.	30.4

	Ge→En		En→Ge	
	Position	BLEU[%]	Position	BLEU[%]
QUAERO: joint subm.	1.	29.9	-	-
KIT	4.	28.0	2.	20.7
LIMSI	7.	26.4	5.	19.6
RWTH	2.	28.4	9.	18.2

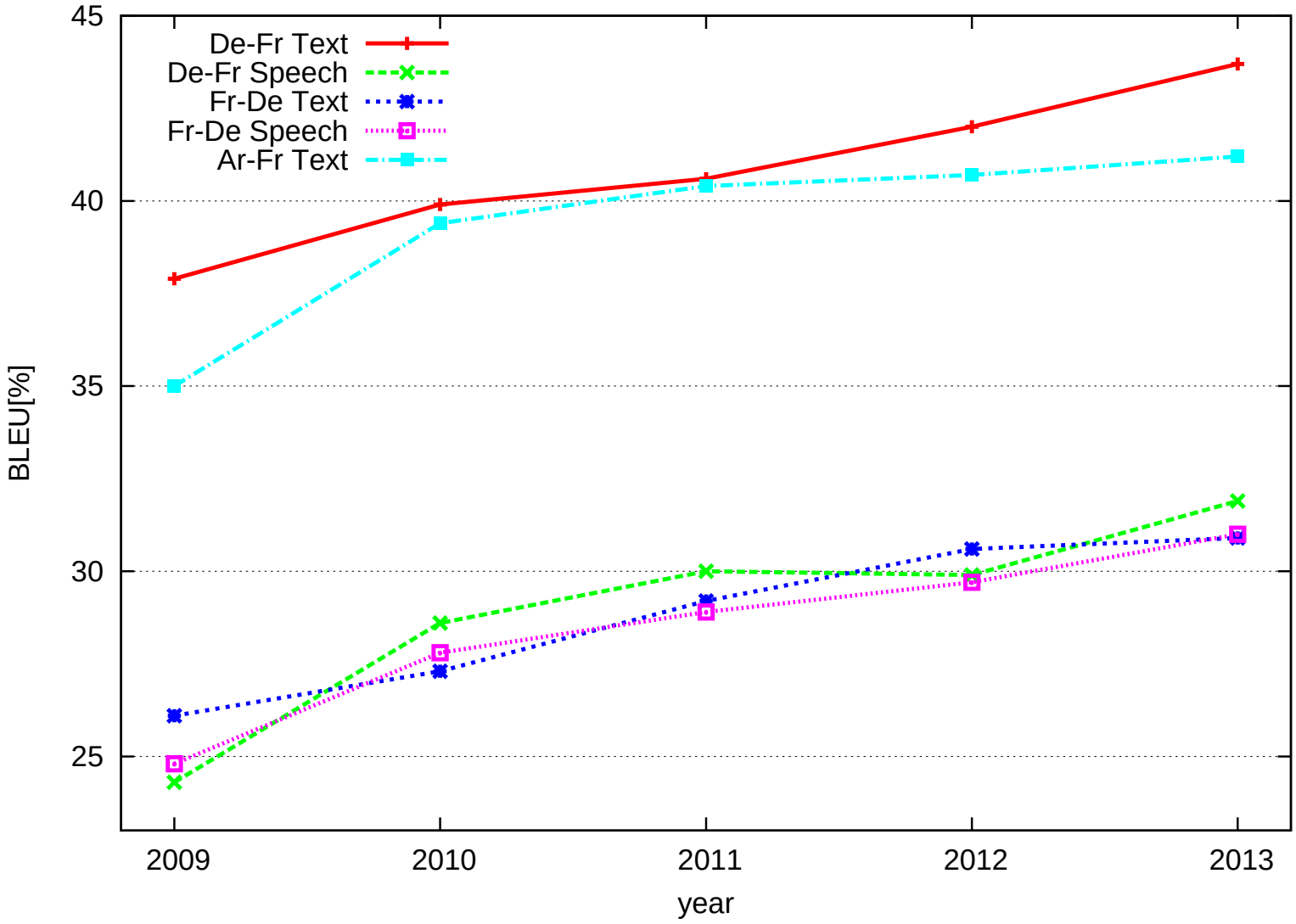
Euromatrix: BLEU[%] for 22 EU Languages

	bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	bg	46.0	41.9	41.2	39.9	56.8	46.9	34.0	33.9	48.2	34.9	45.6	33.3	36.9	32.2	43.9	42.8	46.2	43.2	37.9	43.2	42.4
cs	39.9	cs	42.9	43.3	40.1	57.5	48.3	35.5	35.0	49.7	35.3	47.8	34.4	37.5	31.9	45.0	42.8	47.5	43.5	39.3	43.9	44.2
da	38.2	46.3	da	45.0	39.7	56.3	48.3	35.9	36.2	49.9	35.3	48.0	34.1	37.7	30.9	47.4	41.9	47.8	41.5	36.8	42.1	46.6
de	31.2	44.7	43.4	de	38.7	54.5	46.6	34.6	35.2	48.5	34.9	45.9	33.5	36.3	30.0	46.1	40.6	45.6	40.0	35.4	41.2	43.1
el	39.1	45.2	42.3	41.8	el	54.0	49.4	32.8	33.3	50.4	33.4	48.8	31.9	35.4	30.8	44.5	41.0	49.1	42.1	35.8	40.7	42.5
en	46.7	53.3	50.0	47.5	45.2	en	55.5	40.5	39.4	51.4	40.6	54.8	38.9	43.1	43.5	51.9	50.5	54.9	49.4	45.1	51.2	52.1
es	40.1	47.7	45.1	44.5	43.1	59.5	es	35.2	35.5	54.9	35.2	52.2	33.8	37.1	32.5	47.2	42.5	53.6	44.1	37.9	43.3	45.2
et	35.0	40.3	37.7	39.6	33.2	51.8	41.3	et	33.7	43.5	32.6	40.3	33.8	36.7	27.2	39.5	38.6	40.4	36.5	32.2	39.0	37.8
fi	31.9	38.4	37.1	37.1	31.9	46.3	39.9	35.8	fi	40.3	34.7	38.6	32.6	34.4	25.4	38.8	34.7	39.0	34.8	30.4	35.0	37.1
fr	31.4	42.5	41.2	41.1	39.8	55.7	49.1	31.8	31.7	fr	30.7	48.7	31.6	34.4	28.2	43.1	39.3	48.6	40.6	32.9	39.3	39.9
hu	34.7	40.2	37.2	37.2	33.3	50.1	40.7	33.8	34.1	40.5	hu	39.2	32.0	35.0	27.4	39.7	37.1	39.5	36.6	32.6	37.1	36.9
it	40.5	48.3	45.3	45.2	43.7	59.9	53.0	35.9	36.1	55.5	35.2	it	34.4	37.8	32.8	47.6	43.2	52.6	44.8	37.9	43.6	45.2
lt	33.9	39.7	35.4	36.9	32.0	50.5	40.2	34.7	31.2	42.0	31.9	39.2	lt	38.5	26.8	37.6	36.8	39.4	35.8	30.7	37.6	36.2
lv	35.3	40.9	36.1	37.7	32.9	52.0	41.3	34.9	30.9	43.2	32.0	40.3	37.7	lv	27.0	38.5	38.0	40.3	37.1	31.6	39.0	37.3
mt	42.5	48.2	43.4	42.6	37.5	69.8	50.1	35.7	35.4	51.2	36.5	48.9	35.0	39.2	mt	45.6	45.4	49.2	45.4	40.8	46.2	45.2
nl	39.4	47.1	45.6	45.7	37.4	57.4	49.8	35.5	36.1	51.1	36.2	49.2	34.1	37.4	31.9	nl	42.5	48.9	42.1	37.5	42.8	45.1
pl	40.2	46.1	41.4	43.2	38.1	60.2	46.2	36.7	33.4	49.5	34.7	45.4	35.4	38.7	32.2	43.5	pl	45.5	41.9	37.9	44.9	42.7
pt	40.1	47.5	45.0	44.4	43.4	59.8	54.2	35.5	35.4	55.7	34.6	52.5	33.9	37.2	32.4	47.1	42.5	pt	44.2	37.7	43.1	45.3
ro	41.0	47.5	42.8	42.3	41.2	59.9	49.8	34.4	34.3	52.6	34.9	49.1	33.3	36.9	33.0	45.0	42.6	49.2	ro	37.5	43.3	44.0
sk	40.8	49.9	42.8	41.8	39.2	59.4	47.2	35.0	34.6	47.9	35.9	45.9	34.4	38.1	33.2	44.6	44.1	46.4	42.6	sk	45.4	43.6
sl	41.2	47.4	42.1	43.8	38.5	60.6	46.9	37.0	33.7	49.2	35.0	45.8	35.9	39.6	32.9	44.4	45.1	46.0	42.1	39.0	sl	43.1
sv	37.6	45.9	44.8	43.4	39.4	58.0	47.4	35.0	35.6	48.5	34.8	46.5	33.4	36.6	31.5	45.3	41.5	46.8	41.2	36.6	42.5	sv

(blue: direct systems; test on JRC-EU Acquis Communautaire)

Progress over Time

Performance (BLEU) for winning systems in evaluations



Improvements (2009→2013): Examples

INPUT: Aber wird das reichen, um all das auszugleichen, was derzeit in den größeren Volkswirtschaften passiert?
2009: Mais, pour compenser tout ce que les plus grandes économies?
2013: Mais cela suffira à compenser tout ce qui se passe actuellement dans les grandes économies?

INPUT: In wenigen Jahren wird die große Mehrheit der Weltbevölkerung Zugang zu drahtlosem Breitbandinternet haben.
2009: En quelques années, la grande majorité de la population mondiale, l'accès drahtlosem internet à large bande.
2013: Dans quelques années , la grande majorité de la population mondiale aura accès à internet à haut débit sans fil.

Improvements (2009→2013): Examples

INPUT: Die Globalisierung könnte auch die legale Steuerumgehung erleichtern.
2009: La mondialisation pourrait également la légalité.
2013: La mondialisation pourrait également faciliter l'évasion fiscale légale .

INPUT: Als katholische Sozialdoktrin bis ins späte zwanzigste Jahrhundert prägt der Korporatismus immer noch Verfassungen, Gesetze und Einstellungen auf der ganzen Welt.
2009: Comme la doctrine sociale catholique jusqu'au tardive vingtième siècle marque le corporatisme toujours constitutions, des lois et des attitudes dans le monde entier.
2013: En tant que doctrine sociale catholique jusqu'à la fin du xxe siècle, le corporatisme imprègne encore les constitutions, des lois et des attitudes dans le monde entier.

Summary: MT

- significant technical progress
- many text MT systems for various domains:
  news, Europarl, patents
- many speech MT systems for various domains:
  broadcast news, parliamentary speeches, lectures
- many languages:
  – fully fledged systems: Fr, Ge, En, Sp, ..., Ar, Ch, Ja
  – baseline systems for all 22 EU languages

THE END