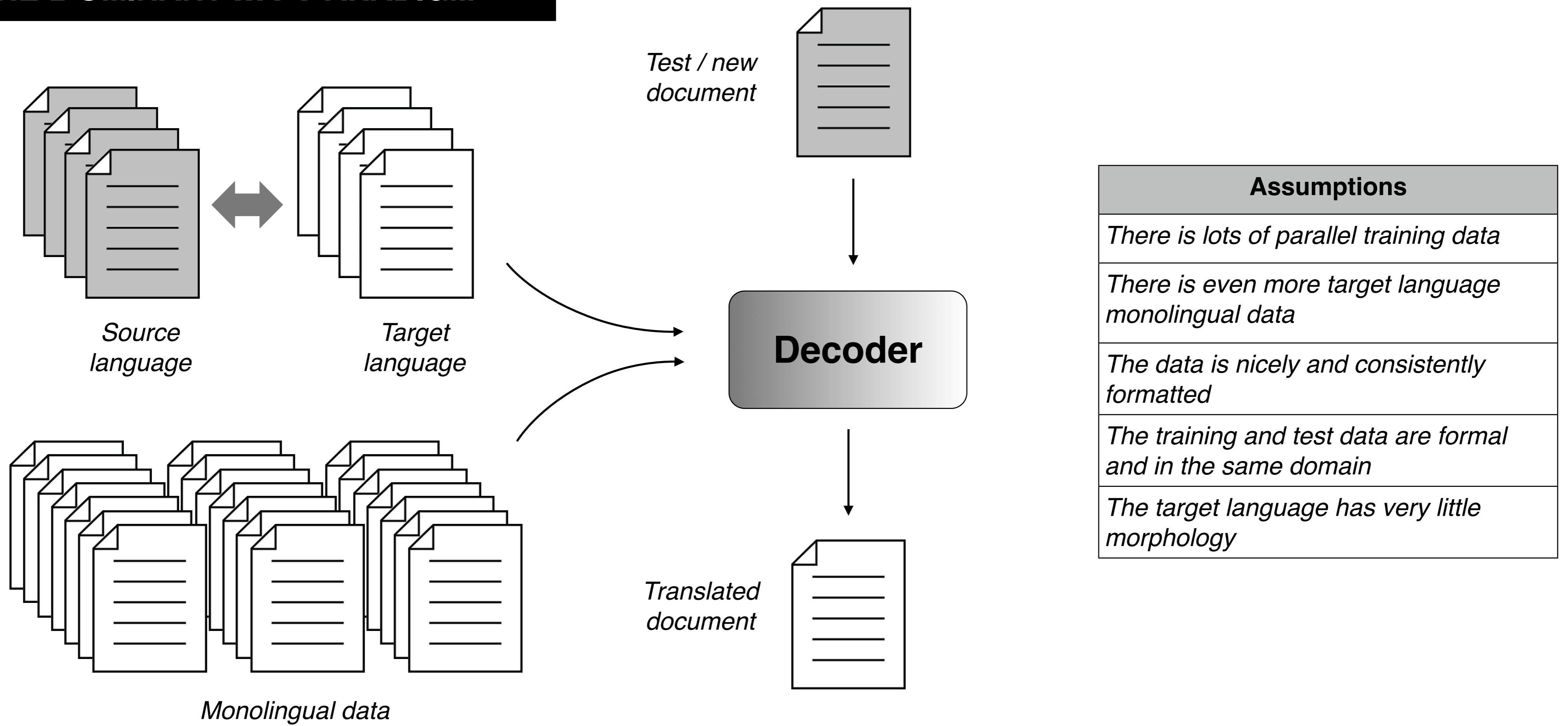# Beyond bitext: Five open problems in machine translation

Adam Lopez and Matt Post

THE DOMINANT MT PARADIGM
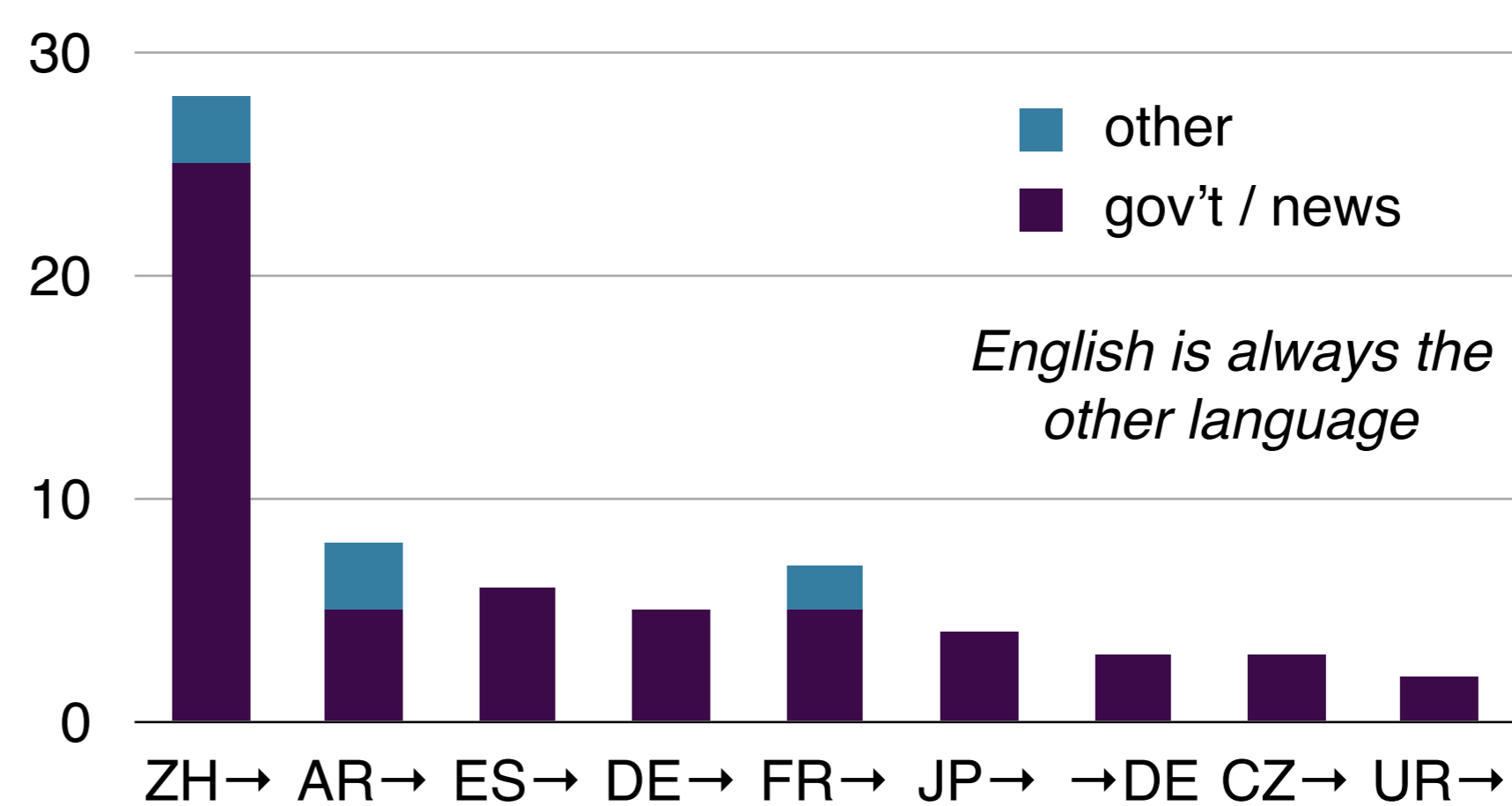
Source language

Target language

Monolingual data

Test / new document

Decoder

Translated document

Assumptions
There is lots of parallel training data
There is even more target language monolingual data
The data is nicely and consistently formatted
The training and test data are formal and in the same domain
The target language has very little morphology

FAVORS THE FOLLOWING

Industrial-scale engineering:

1st place	18
2nd place	12
3rd place	3
4th or worse	1

"ONLINE-B" performance across all language tasks in WMT 2010–13, according to human judgment

Languages of first-world countries:

other
gov't / news

English is always the other language

ZH→ AR→ ES→ DE→ FR→ JP→ →DE CZ→ UR→

THE NEXT 20 YEARS

JOHNS HOPKINS
U N I V E R S I T Y

human language technology
center of excellence