Translation quality, neural machine translation and language resources

Kenneth Heafield, University of Edinburgh

Transition to Neural Models

What types of models won pairs in the Conference on Machine Translation?

Year	**Neural**	**Phrase**	**Rule**
2015	1	9	1
2016	6	6	1
2017	14	0	0

Neural is better at agreement

Source Byl to bratr, který bral věci takové, jaké jsou.
Reference He was the brother that went with the flow.
Phrase It was a brother who took things as they are.
Neural He was a brother who took things the way they are.

Decisions based on whole sentence, not just local context.

Word selection

Source Seit Jahrzehnten fördert Langer den Nachwuchs.
Reference Langer has been encouraging up-and-coming talent for years.
Phrase For decades, Langer promotes the offspring.
Neural For decades, Langer has been promoting the young.

Larger context $\Longrightarrow$ generally better at fluency.

Rare words are hard

Source	Jennifer Aniston:	Ich werde immer in Schubladen gesteckt
Reference	Jennifer Aniston:	I'm always pigeonholed
Phrase	Jennifer Aniston:	I am always plugged in drawers
Neural	Jennifer Aniston:	I'll always be put in drawers

$\Longrightarrow$ Translations covering your terminology are important.

Style

Source	Erkek kardeşim, her duruma uyum sağlardı.
Reference	He was the brother that went with the flow.
Phrase	My brother, sağlardı fit in every situation.
Neural	My brother was harmonised in every situation.

Neural fits style more closely:

Great if you add in-domain data

Awkward if you don't. Subtitles swear a lot.

Data Sizes

Total Data

Out-of-the-box neural beats phrase-based on >10,000,000 words
(Koehn and Knowles, 2016)

Recent work: neural beats phrase-based on >200,000 total words
(Nguyen and Chiang, 2017)

In-Domain Data

10,000 words: enough to test
50,000 words: light customization
Even monolingual data can buy 1–4 BLEU points

Case Study: Patent Translation

European Patent Office

Refuses to share translations publicly

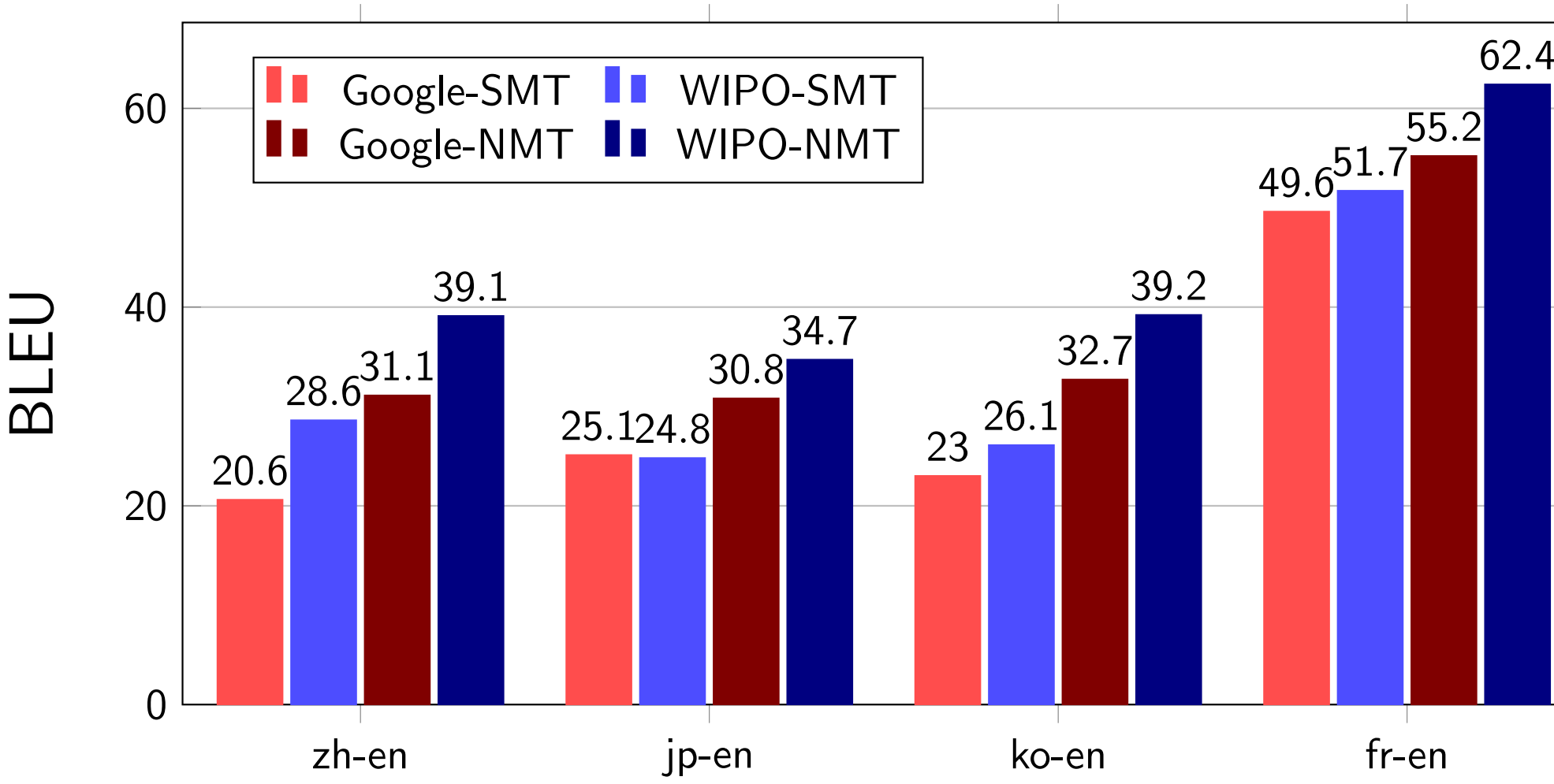
Gave translations to Google, got generic system

World Intellectual Property Organization

Has European Patent Office data

Hired academic consultant, got custom system

Google/EPO loses to WIPO



(Plot courtesy Marcin Junczys-Dowmunt and Bruno Pouliquen, WIPO)

EPO will not share data they gave Google.

"Several hundred thousand high-quality translations of patents"

Google's Solution is Inferior

Custom WIPO system wins by 6.4 BLEU on average

EPO will not share data they gave Google.

"Several hundred thousand high-quality translations of patents"

Google's Solution is Inferior

Custom WIPO system wins by 6.4 BLEU on average

EPO Mission

"support innovation, competitiveness and economic growth across Europe. . ."

EPO will not share data they gave Google.

"Several hundred thousand high-quality translations of patents"

Google's Solution is Inferior

Custom WIPO system wins by 6.4 BLEU on average

EPO Mission

"support innovation, competitiveness and economic growth across Europe. . ."

Patents are Public

"Specifications of European patents shall be published in the language of the proceedings and shall include a translation of the claims in the other two official languages of the European Patent Office." (Article 14, European Patent Convention)

ParaCrawl Project

Co-financed by the European Union

Connecting Europe Facility

Mine web for translations in all 24 EU languages:

September 2017	Kickoff
January 2018	12 languages
June 2018	18 languages
March 2019	24 languages

Expect 1 billion translated words for 8 languages, millions for others.

ParaCrawl Project

Co-financed by the European Union

Connecting Europe Facility

Mine web for translations in all 24 EU languages:

September 2017	Kickoff
January 2018	12 languages
June 2018	18 languages
March 2019	24 languages

Expect 1 billion translated words for 8 languages, millions for others.

ELRC guidance: manually review 3% of the corpora

→ Would cost more than entire budget.

→ EU agreed 2000 sentences/language.

ParaCrawl Does Not Replace Your Data

ParaCrawl

Broad domain $\rightarrow$ improve MT@EC coverage

Big, but noisy

Your Data

In-domain $\rightarrow$ improve your use case

Clean, but smaller

WIPO got 6.4 BLEU from in-domain data!

Funding

HimL
ModernMT
ParaCrawl
QT21
SUMMA
TraMOOC