Beyond bitext: Five open problems in machine translation

Adam Lopez and Matt Post, Johns Hopkins University

human language technology
center of excellence

Q Is machine translation a solved problem?
A Yes, if by *machine translation* you mean *machine learning from big bitext.*

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

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

"ONLINE-B" placed second this year in three tasks...to a system with a terabyte language model.

Q So, in what scenarios don't I have big bitext?
A Most of them! Here are **five open problems** in machine translation.

Translation of low-resource languages

Tamil அவர் மக்களாட்சியை ஒழித்து இன வாரியான புது உலக அடைவை வற்புறுத்தினார்.
English He destroyed the democracy and encouraged governement based on caste.
MT Sex-wise to get rid of democracy, he insisted that the new global directory.

Translation of informal text

Spanish yo la kiero ver pq me encanta Yonghwa jijiji
English I want to see it because I love Yonghwa hehehe
MT I love kiero see Yonghwa jijiji pq

Translation across domains

French mode et voie(s) d'administration
English method and route(s) of administration
MT fashion and voie(s) of directors

source: Measuring Machine Translation Errors in New Domains *(Irvine, Morgan, Carpuat, Daumé III, and Munteanu, TACL 2013)*

Translation of speech

Spanish E- ella su ma- el marido de ella es de aquí ¿verdad?
English H- her husb- her husband is from here, right?
MT E-ma-she her husband she is here right?

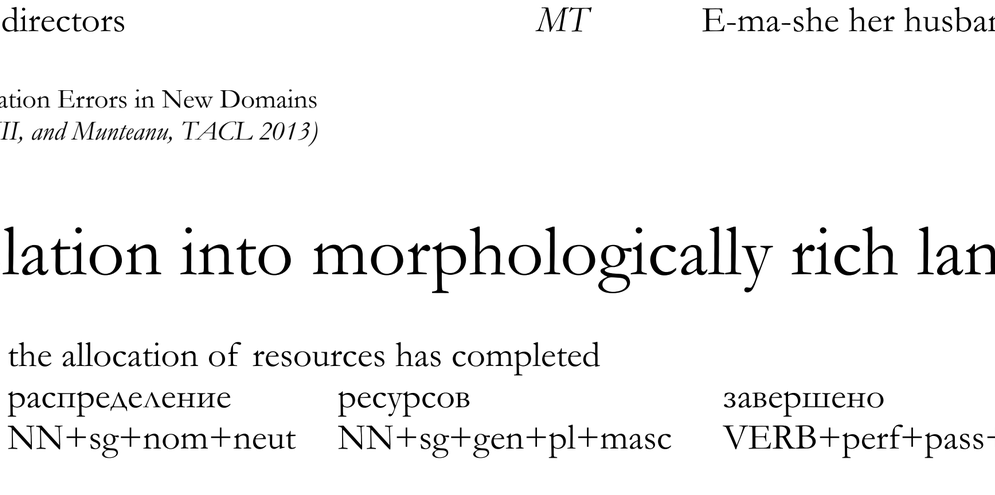
Translation into morphologically rich languages

English	the allocation of	resources	has completed
Russian	распределение	ресурсов	завершено
Gloss	NN+sg+nom+neut	NN+sg+gen+pl+masc	VERB+perf+pass+part+neut+sg

source: Generating Complex Morphology for Machine Translation *(Minkov, Toutanova, and Suzuki, ACL 2007)*

Q But isn't the research community already working on these problems?
A Yes, but the vast majority of research is on big bitext.

Exhibit B: Test sets used in 51 papers at ACL 2013 that contained machine translation experiments.



Except for the column labeled "other", English was *always* one of the languages.

Q What can I do to change the situation?
A Research in MT is driven by shared tasks and available test sets. We need *sustained* focus on shared tasks and test sets for these scenarios.