
SYSTRAN Software
Intelligent Language Technologies
Building Renewable
Language Assets in
Government Domains

Agenda

- Why SYSTRAN
- Background of Government Work
- Current Direction

Why SYSTRAN

Expertise

- Leadership role for 40+ years in machine translation
- 200+ language technology professionals in the US, France and Korea

Best Fit Technology

- Multiple deployment options (on-premise, desktop, mobile, cloud)
- Specialized engines for domains

Continuous Improvement

- Re-investment of 25% revenue into R&D
- Partnerships with academia, currently focused in the area of Neural MT

Trusted Industry Partner

- Organic growth through government partner referrals
- Flexible approach to meet mission requirements

Background of Government Work

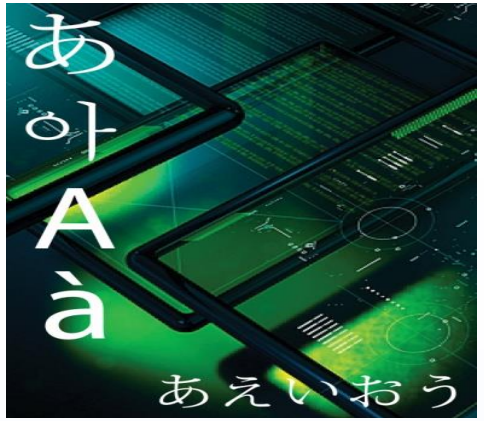
IT Systems

- IC Networks
- Command, Control, Communication, Computers & Intelligence (C4I) Systems

Language expansion resulting from world events, starting with Russian

Language Requirements

Technology Evolution



1. Migration off the mainframe
2. Arrival of the internet
3. Statistical MT
4. **Neural MT**

What have we done in the past

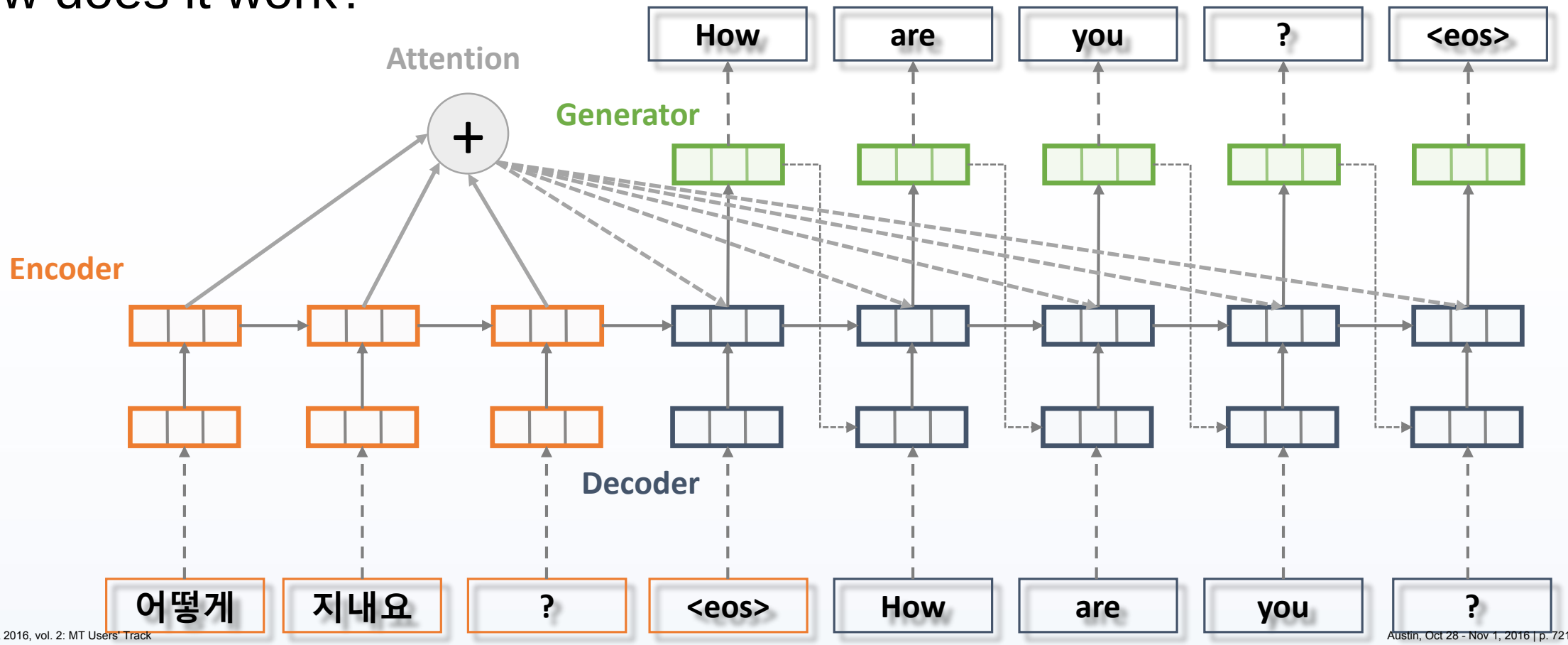
- Full Rule-Based Backbone for English target
  - Morphology, part of speech, normalization
  - Mix rules and statistical decision models
- Statistical post-editing (SPE)
  - Statistical layer to fine tune output based on bilingual corpus
  - Allows for greater customization
  - Allows for quality translations in domains with little bilingual corpus
- Dictionaries
  - Extensive dictionaries – user defined
  - Cover terminology not found in bilingual corpus
  - Cover low resource domains and languages – Science and Technology
- Entity recognition

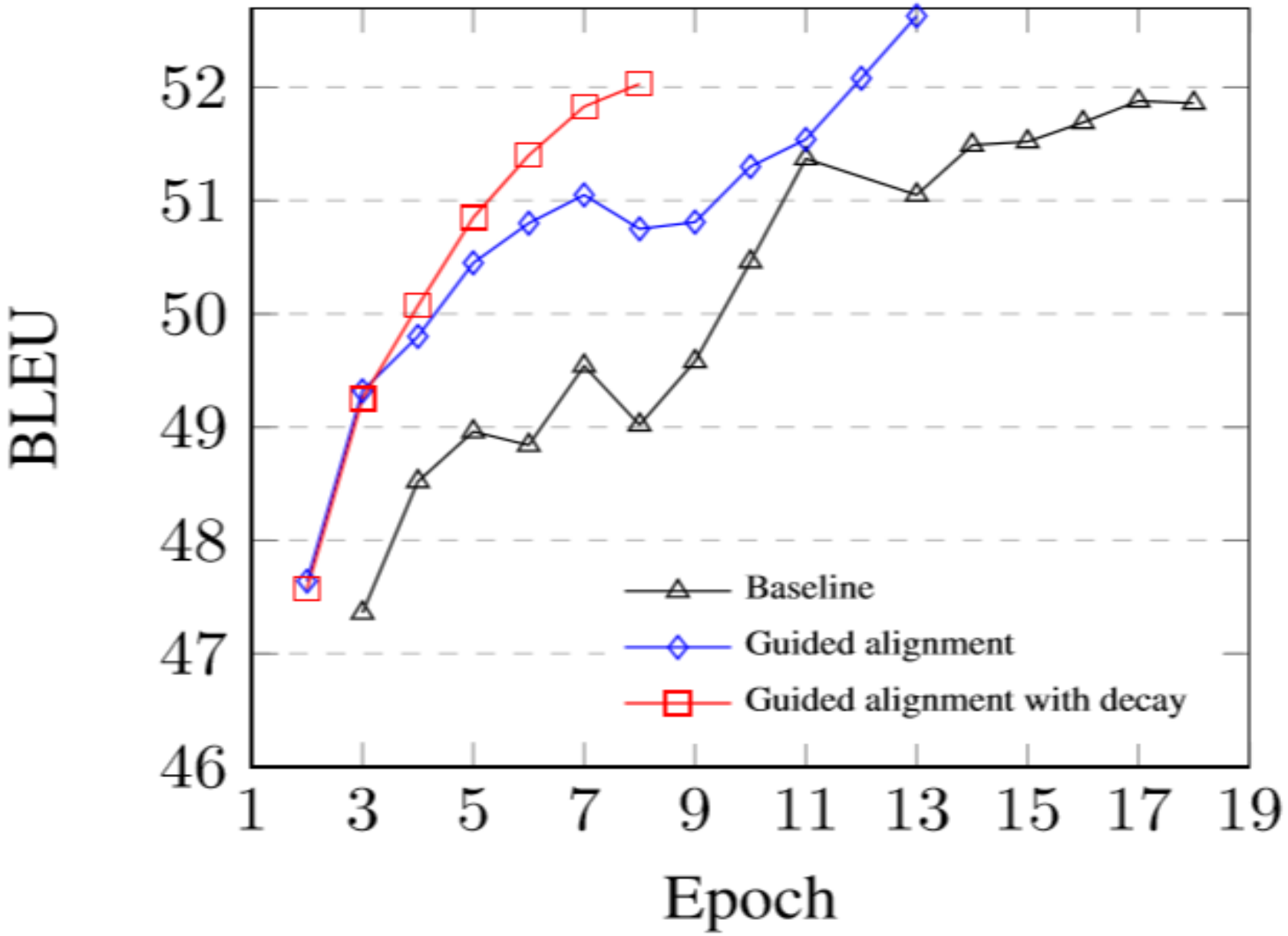
Pure Neural Machine Translation

- http://demo-pnmt.systran.net
  - 14 language pairs with more on the way
- https://arxiv.org/pdf/1610.05540v1.pdf
- Systran specialization
  - Architecture
  - Pre-processing
  - Features
  - Post-processing
  - Domain specialization
- Incorporate standard customization features
  - User Dictionaries, Translation Memory etc.

Neural Machine Translation

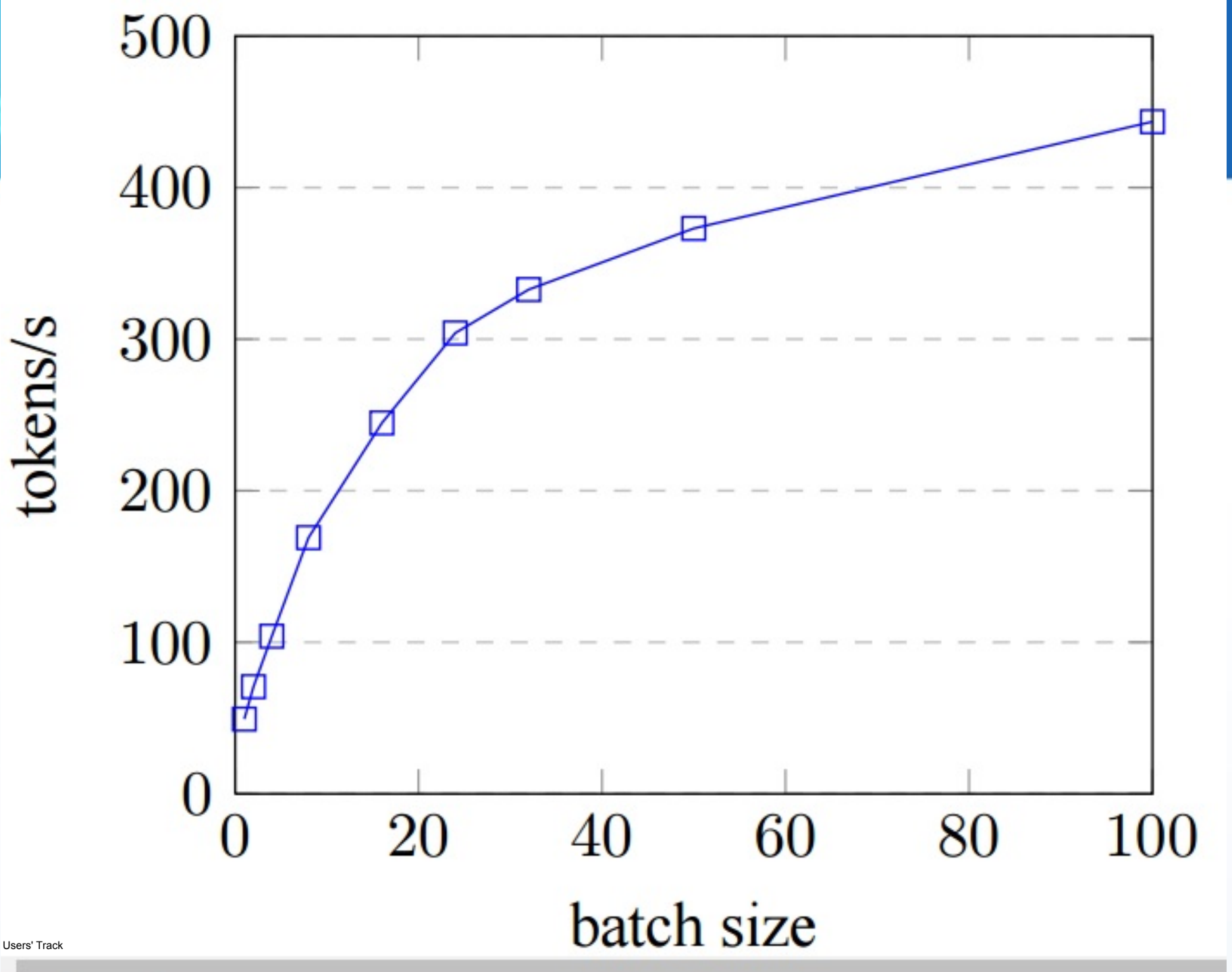
- How does it work?




BLEU
52
51
50
49
48
47
46
1
3
5
7
9
11
13
15
17
19
Epoch
Baseline
Guided alignment
Guided alignment with decay

Architecture

- seq2seq-attn (https://github.com/harvardnlp/seq2seq-attn)
  - Harvard NLP group
  - Sequence to sequence RNN
  - Guided attention with decay
  - Features in source and target
  - Open source with ability to tune several parameters
- Run time can now be done on CPU
  - 4 threads on a desktop Intel i7 CPU
  - Pruning
  - Distillation
  - Batch modes with beam size

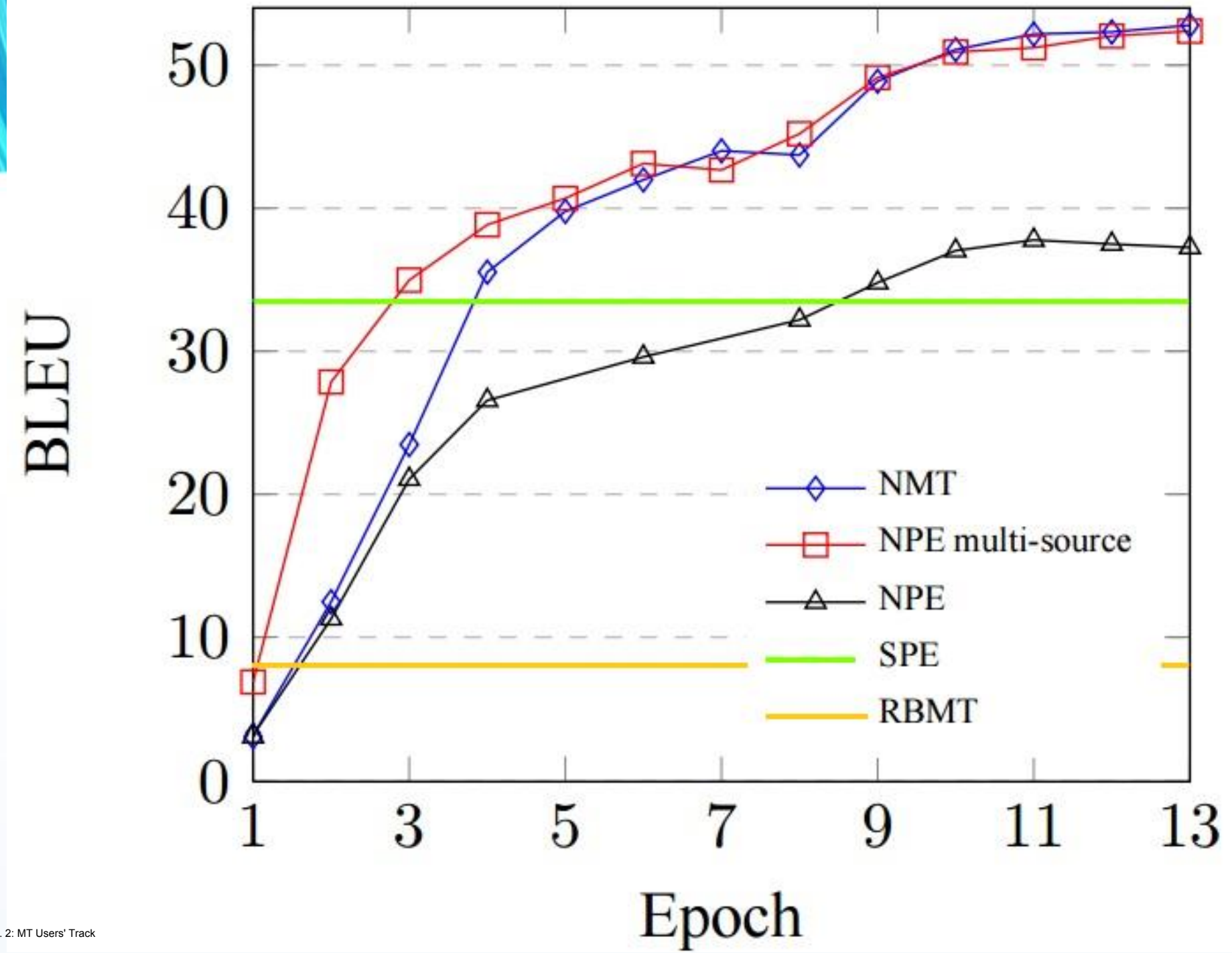

500
400
300
200
100
0
tokens/s
0
20
40
60
80
100
batch size

Pre-processing

- Normalization
- Tokenization
  - Word based
  - Special tokenization for CJK, German and Arabic
  - More experimentation (BPE, character, combo)
- Entity recognition – replace with token (__ent_numeric)
  - Replacement needs to be in both to learn

Linguistic Features

- Dictionaries, entities, capitalization
  - Maybe include part of speech, parse
- Formality mode – Korean
  - Domain mode
- Include traditional translation
  - Smart "Neural Post-Editing"
- Can't neural networks learn everything, given enough data?
  - Depends on complexity of language and amount of data
  - We know NMT has issues with OOV

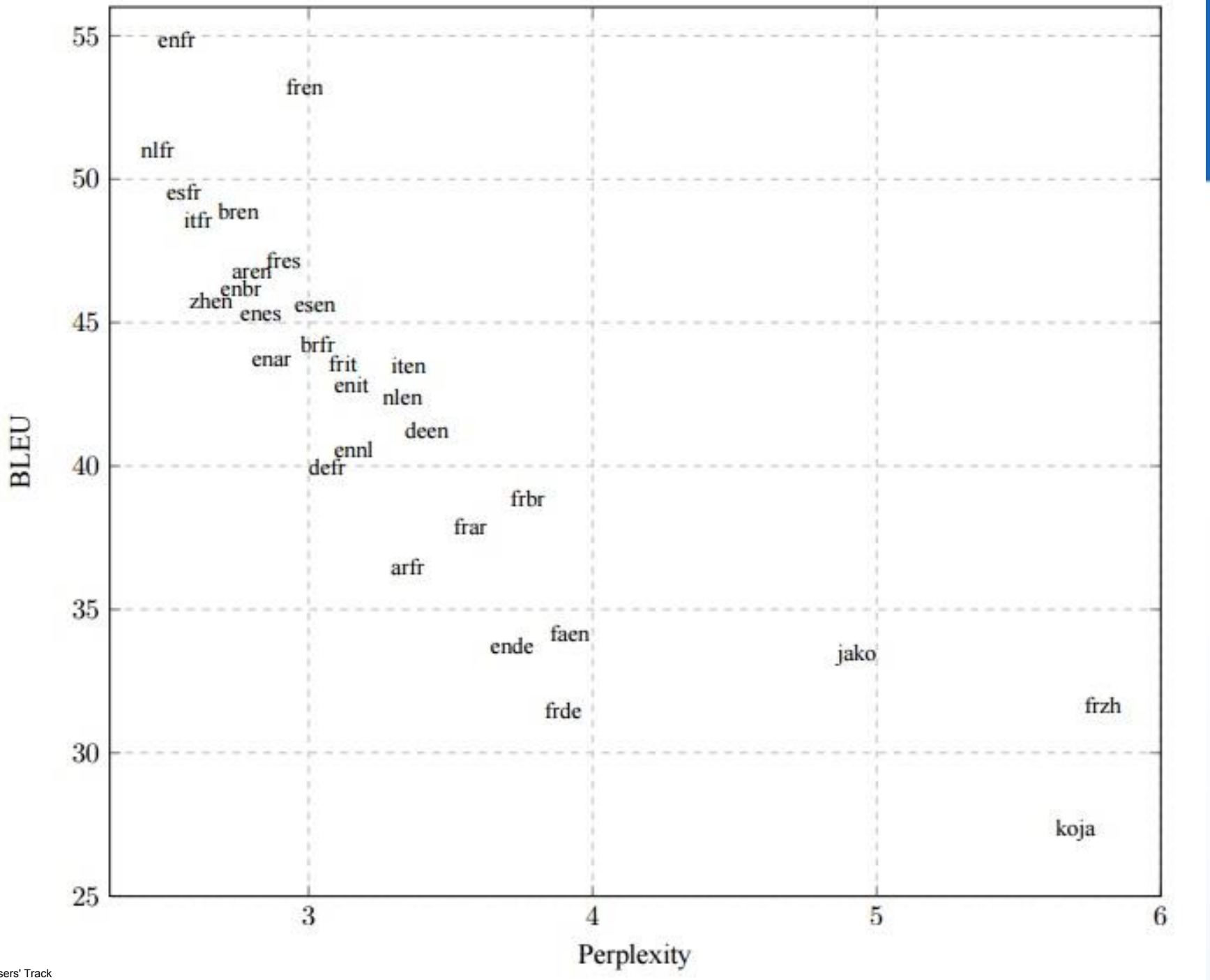

BLEU
50
40
30
20
10
0
1
3
5
7
9
11
13
Epoch
NMT
NPE multi-source
NPE
SPE
RBMT

Post-processing

- Restore entities and dictionary terminology
- Apply features (capitalization)
- Restore punctuation
- Out of Vocabulary
  - Look them up with dictionary
  - Use SMT-style phrase table
  - Allow NMT to transliterate

Domain specialization

- Stack neural networks
  - Adds another neural network trained on in-domain corpus
  - Requires some tweaks on vocabulary
  - Enable full specialization in a few hours
- Marker for domain
- Synthetic corpus
  - Makes use of well-written, in-domain sentences from target language
- Inject terminology
  - Full support for User Dictionaries
  - Explicit recognition of entities
  - Other linguistic knowledge?


55
50
45
40
35
30
25
BLEU
3
4
5
6
Perplexity
enfr
fren
nlfr
esfr
itfr
bren
fres
aren
enbr
zhen
enes
esen
brfr
enar
frit
iten
enit
nlen
deen
ennl
defr
frbr
frar
arfr
faen
ende
jako
frde
frzh
koja

Examples

- FARSI: فرهنگ هر کشور دارای هویت و ویژگی های خاص خود می باشد.
- SPE: Culture of each country has identity and its special features.
- PNMT: The culture of each country has its own identities and particularities.

- FARSI: به دلیل اختلافات ایدئولوژیکی و عدم توافق بر سر اهداف مورد حمله, بسیاری از هسته ها از این سازمان تازه بیرون کشیدند یا هیچوقت به آن نپیوستند.
- SPE: For an ideological reason for disagreements and discordance, goals of case of attack pulled out many of nuclei of this new organization over or they never joined that.
- PNMT: Many cells pulled out of the new organization or never joined it because of ideological differences and disagreements over the targeted targets.

QUESTIONS

Beth Flaherty
Beth.flaherty@systrangroup.com

Joshua Johanson
Joshua.johanson@systrangroup.com