Interactive Machine Translation: From Research to Practice

Spence Green

spence@lilt.com

30 October 2016

Joint work with:

John DeNero	Jason Chuang
Saša Hasan	Thang Luong Minh
Joern Wuekber	Sida Wang
Franz Och	Jeff Heer
Daniel Cer	Chris Manning

2012: First Stanford System

Text from a document about: **a mathematical theory**.

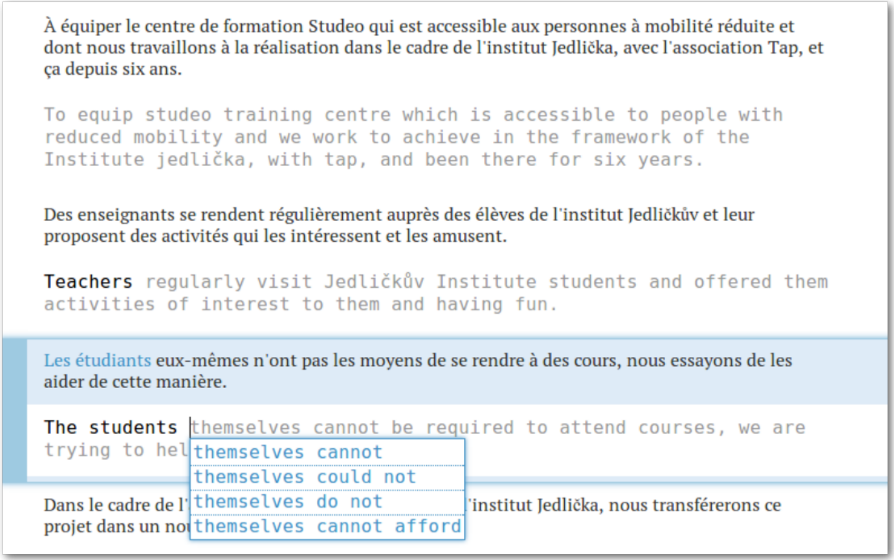
The physicist Arthur Eddington drew on Borel's image further in The Nature of the Physical World (1928), writing: If I let my fingers wander idly over the keys of a typewriter it might happen that my screed made an intelligible sentence.

```
Le physicien Arthur Eddington a attiré sur l'image de
Borel dans le caractère du monde physique (1928), écrit:
Si je laisse mes doigts se promener les bras croisés sur
les touches de la machine à écrire, il peut arriver que
mon chape fait une phrase intelligible.
```

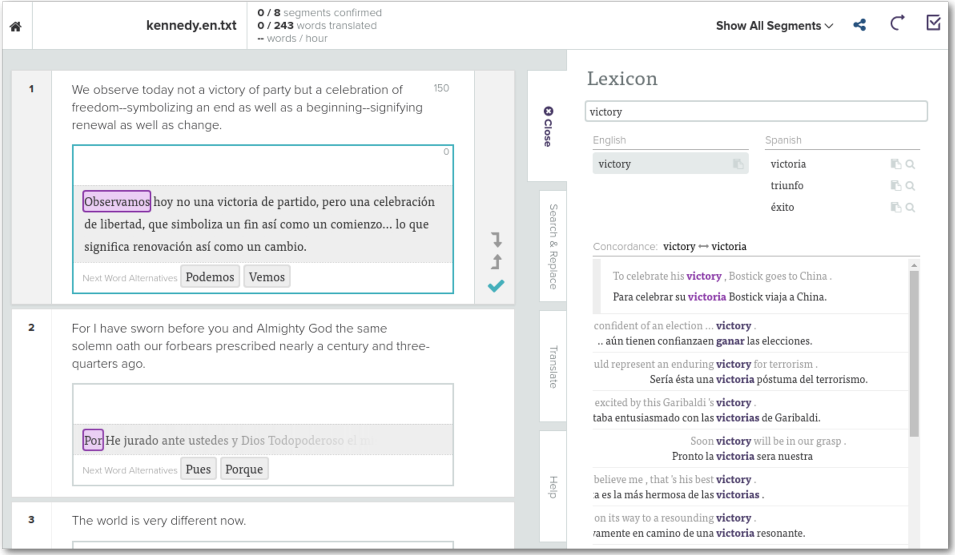
Idle: 00:09 (maximum: 03:00)

Submit

2014: Predictive Translation Memory



2016: Lilt



kennedy.en.txt
0 / 8 segments confirmed
0 / 243 words translated
-- words / hour
Show All Segments
1
We observe today not a victory of party but a celebration of freedom--symbolizing an end as well as a beginning--signifying renewal as well as change.
150
0
Observamos hoy no una victoria de partido, pero una celebración de libertad, que simboliza un fin así como un comienzo... lo que significa renovación así como un cambio.
Next Word Alternatives
Podemos
Vemos
2
For I have sworn before you and Almighty God the same solemn oath our forbears prescribed nearly a century and three-quarters ago.
Por He jurado ante ustedes y Dios Todopoderoso el mi
Next Word Alternatives
Pues
Porque
3
The world is very different now.
Close
Search & Replace
Translate
Help
Lexicon
victory
English
Spanish
victory
victoria
triunfo
éxito
Concordance: victory ⟷ victoria
To celebrate his victory , Bostick goes to China .
Para celebrar su victoria Bostick viaja a China.
confident of an election ... victory .
.. aún tienen confianzaen ganar las elecciones.
uld represent an enduring victory for terrorism .
Sería ésta una victoria póstuma del terrorismo.
excited by this Garibaldi 's victory .
taba entusiasmado con las victorias de Garibaldi.
Soon victory will be in our grasp .
Pronto la victoria sera nuestra
believe me , that 's his best victory .
a es la más hermosa de las victorias .
on its way to a resounding victory .
ramente en camino de una victoria resonante.

Classic Use Cases for MT

Assimilation – user pulls translation

- "Gisting" – Google Translate / browser integration
- Full-sentence MT
- Main focus of the research community

Classic Use Cases for MT

Assimilation – user pulls translation

"Gisting" – Google Translate / browser integration

Full-sentence MT

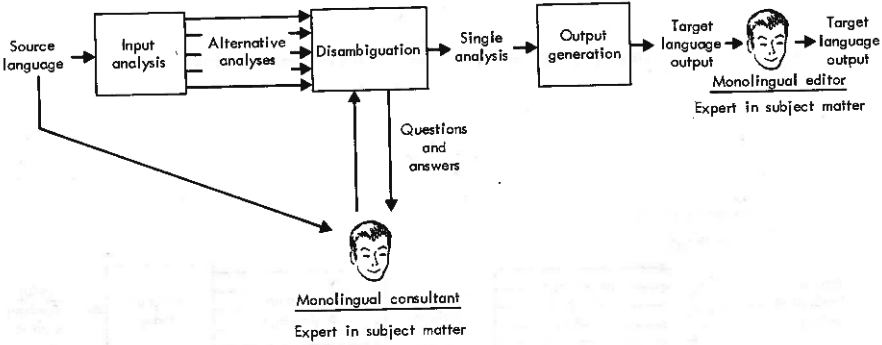
Main focus of the research community

Dissemination – content publishing

Intent to communicate

"Post-editing" – MT as productivity enhancer

Main focus of the **translation industry**



*the **man and the machine are collaborating** to produce not only a translation…but also a device…that is being constantly enhanced.*

[Kay 1980]

Mixed-Initiative Systems

Translation is a classic **mixed-initiative** task

> **Mixed-initiative** A human-computer discourse in which each party takes turns driving the task

Mixed-Initiative Systems

Translation is a classic **mixed-initiative** task

> **Mixed-initiative** A human-computer discourse in which each party takes turns driving the task

Principles of Mixed-initiative User Interfaces [Horvitz 1999]

Mixed-Initiative Systems

Translation is a classic **mixed-initiative** task

> **Mixed-initiative** A human-computer discourse in which each party takes turns driving the task

Principles of Mixed-initiative User Interfaces [Horvitz 1999]

Basic Post-editing	**Violates Horvitz' Principle**
Provenance / trust	#11 – remember recent interactions

Mixed-Initiative Systems

Translation is a classic **mixed-initiative** task

> **Mixed-initiative** A human-computer discourse in which each party takes turns driving the task

Principles of Mixed-initiative User Interfaces [Horvitz 1999]

Basic Post-editing	**Violates Horvitz' Principle**
Provenance / trust	#11–remember recent interactions
Static	#9–collaborative refinement

Mixed-Initiative Systems

Translation is a classic **mixed-initiative** task

> **Mixed-initiative** A human-computer discourse in which each party takes turns driving the task

Principles of Mixed-initiative User Interfaces [Horvitz 1999]

Basic Post-editing	**Violates Horvitz' Principle**
Provenance / trust	#11 – remember recent interactions
Static	#9 – collaborative refinement
Domain mismatch	#12 – learn by observing

Prior Work: One-slide Summary

Basic post-editing

Makes translators faster (c. 2010 research)

Produces higher quality (c. 2013)

Prior Work: One-slide Summary

Basic post-editing

Makes translators faster (c. 2010 research)

Produces higher quality (c. 2013)

Qualitative assessment: still poor, esp. for experts [O'Brien and Moorkens 2014]

Prior Work: One-slide Summary

Basic post-editing

Makes translators faster (c. 2010 research)

Produces higher quality (c. 2013)

Qualitative assessment: still poor, esp. for experts [O'Brien and Moorkens 2014]

State of Interactive MT (c. 2014)

Speed / quality assessment: no better than PE

Translators like the idea, in practice suggestions are distracting

2014: Predictive Translation Memory

À équiper le centre de formation Studeo qui est accessible aux personnes à mobilité réduite et dont nous travaillons à la réalisation dans le cadre de l'institut Jedlička, avec l'association Tap, et ça depuis six ans.

To equip studeo training centre which is accessible to people with reduced mobility and we work to achieve in the framework of the Institute jedlička, with tap, and been there for six years.

Des enseignants se rendent régulièrement auprès des élèves de l'institut Jedličkův et leur proposent des activités qui les intéressent et les amusent.

Teachers regularly visit Jedličkův Institute students and offered them activities of interest to them and having fun.

Les étudiants eux-mêmes n'ont pas les moyens de se rendre à des cours, nous essayons de les aider de cette manière.

The students themselves cannot be required to attend courses, we are trying to hel

themselves cannot
themselves could not
themselves do not
themselves cannot afford

Dans le cadre de l' ... 'institut Jedlička, nous transférerons ce projet dans un no

Interactions

Source comprehension – simple lexicon, source coverage

Target gisting – partial and complete suggestions

Target generation – autocomplete, re-ordering

Interactions

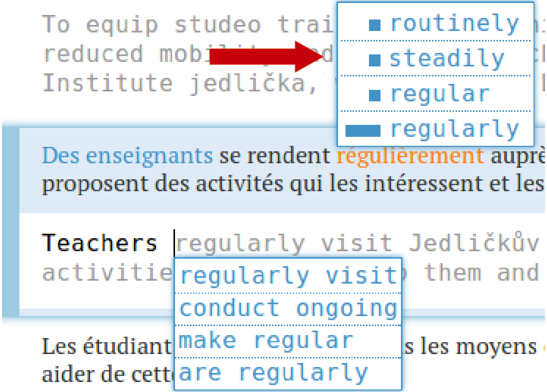
Source comprehension – simple lexicon, source coverage

Target gisting – partial and complete suggestions

Target generation – autocomplete, re-ordering

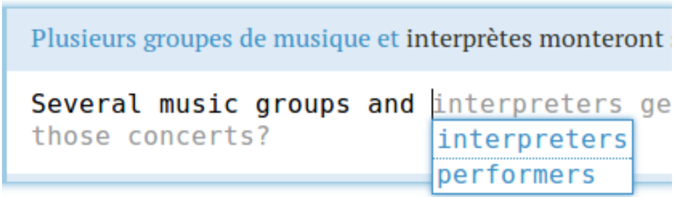
Informed by Horvitz' (1999) mixed-initiative principles

Source Comprehension



Horvitz #6 – allowing efficient direct invocation and termination

Target Generation | Autocomplete



Horvitz #5 – employing dialog to resolve key uncertainties

Human Subjects Experimental Design

Task	Translate French-English or English-German
Conditions	Post-edit (pe) and PTM
Expert Subjects	16 per language pair (from Proz)
Source Data	≈3,000 tokens of News / Medical / Software

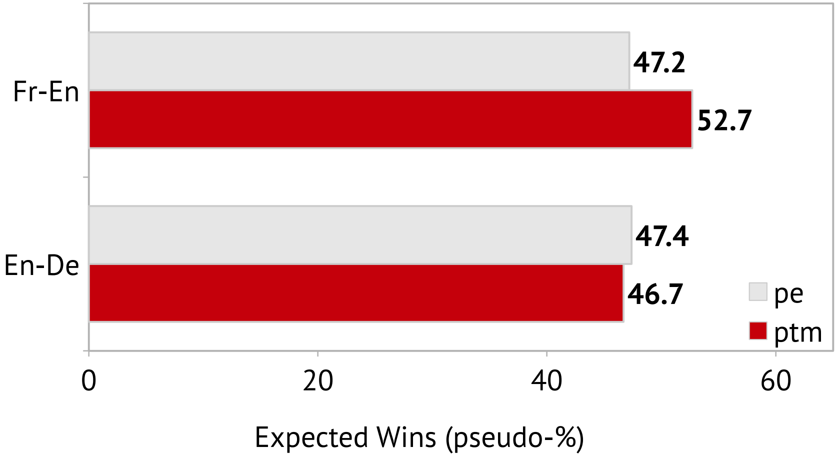
Human Subjects Experimental Design

Task	Translate French-English or English-German
Conditions	Post-edit (pe) and PTM
Expert Subjects	16 per language pair (from Proz)
Source Data	≈3,000 tokens of News / Medical / Software

Three research questions:

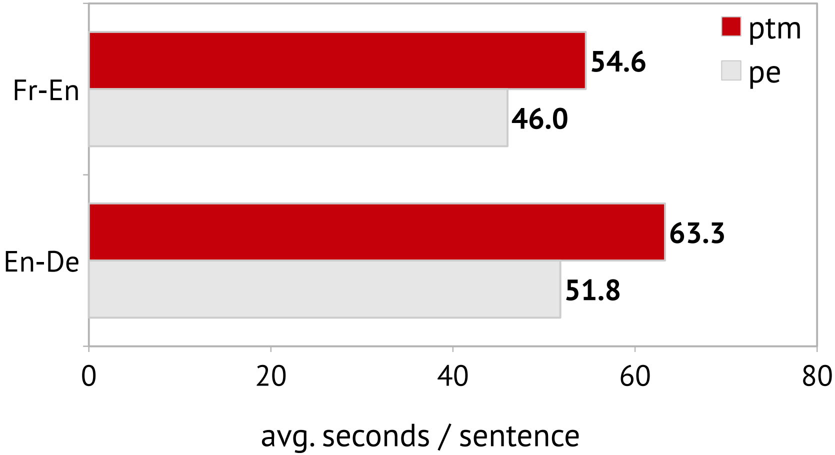
1. Does PTM reduce **time**?
2. Does PTM increase **quality**?
3. Do subjects **use interactive aids**?

Question #1 | Quality



PTM is better for Fr-En ($p < 0.05$)

Question #2 | Time



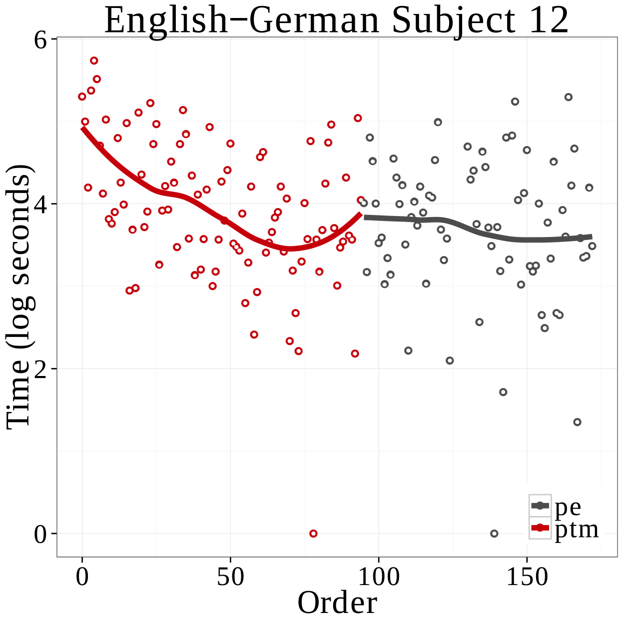
PTM is slower for En-De ($p < 0.01$)

Time: Learning curve?

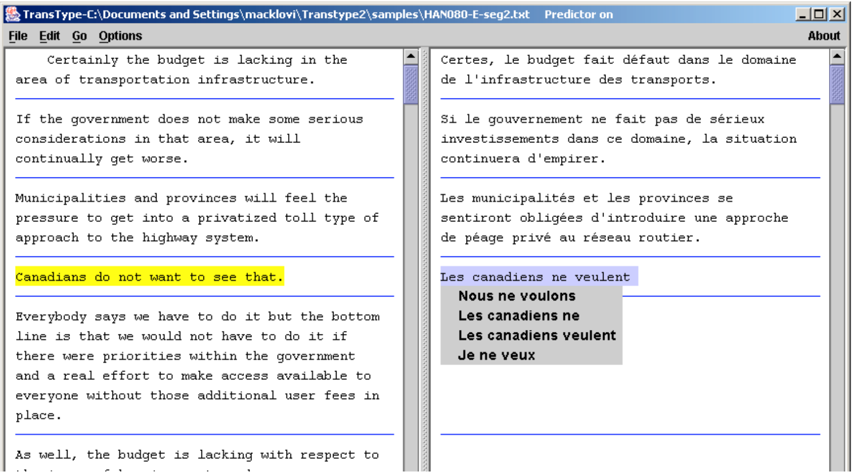
Post-edit mode was easier at first, but **the interactive mode was better once I got used to it.**

If I had time to use the interactive tool *and grow accustomed to its way of functioning, it would be quite useful…*

I am used to this [post-edit]*, this is how Trados works.*

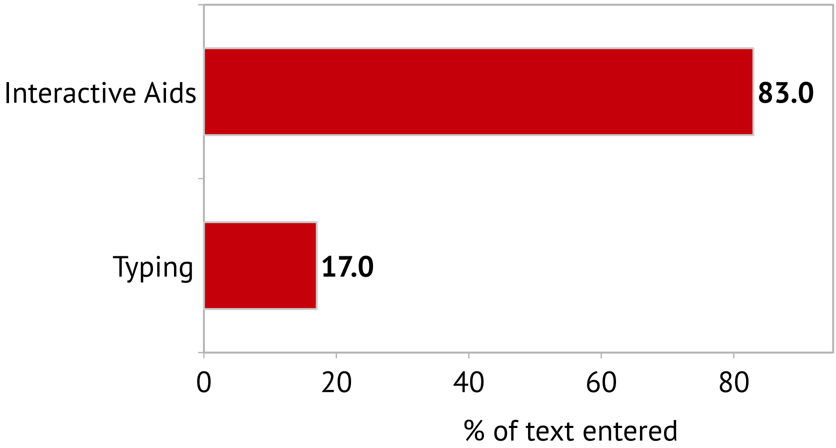

English−German Subject 12
Time (log seconds)
Order
0
2
4
6
0
50
100
150
pe
ptm

Question #3 | Interactive Usage



TransType – users typed 69% of text [Langlais and Lapalme 2002]

Question #3 | Interactive Usage



TransType – users typed 69% of text [Langlais and Lapalme 2002]

End of 2014: Open Problems

Online adaptation

Not online, no domain adaptation

Not tuned for prefixes

End of 2014: Open Problems

Online adaptation

Not online, no domain adaptation

Not tuned for prefixes

Prefix decoding was poor

Low next word accuracy

No diversity

End of 2014: Open Problems

Online adaptation

Not online, no domain adaptation

Not tuned for prefixes

Prefix decoding was poor

Low next word accuracy

No diversity

UI issues

Dropdown is distracting

High learning curve

End of 2014: Open Problems

Online adaptation

Not online, no domain adaptation

Not tuned for prefixes

Prefix decoding was poor

Low next word accuracy

No diversity

UI issues

Dropdown is distracting

High learning curve

Hierarchical Incremental Adaptation

Goal: Reduce repeated MT errors

Hierarchical Incremental Adaptation

Goal: Reduce repeated MT errors

- After each segment is translated, update the model

Hierarchical Incremental Adaptation

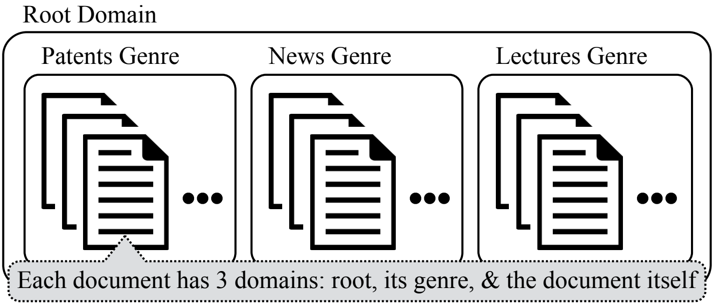
Goal: Reduce repeated MT errors

- After each segment is translated, update the model
- Translation model: updates to features and weights

Hierarchical Incremental Adaptation

Goal: Reduce repeated MT errors

- After each segment is translated, update the model
- Translation model: updates to features and weights
- Hierarchical domains: genres and documents



[Wuebker et al. 2015]

Incremental Adaptation: Approach

Weight adaptation

- Feature augmentation (FEDA) [Daumé III 2007]

  → all weights are replicated for all domains

Incremental Adaptation: Approach

Weight adaptation

- Feature augmentation (FEDA) [Daumé III 2007]

  → all weights are replicated for all domains

- For each example (f, e), active domain features are added

Incremental Adaptation: Approach

Weight adaptation

- Feature augmentation (FEDA) [Daumé III 2007]

  → all weights are replicated for all domains
- For each example (f, e), active domain features are added
- SGD with AdaGrad [Duchi et al. 2011]

Incremental Adaptation: Approach

Weight adaptation

- Feature augmentation (FEDA) [Daumé III 2007]

  → all weights are replicated for all domains
- For each example (f, e), active domain features are added
- SGD with AdaGrad [Duchi et al. 2011]

Translation model adaptation

- Learn separate genre-specific translation model features

Incremental Adaptation: Approach

Weight adaptation

- Feature augmentation (FEDA) [Daumé III 2007]

  → all weights are replicated for all domains
- For each example (f, e), active domain features are added
- SGD with AdaGrad [Duchi et al. 2011]

Translation model adaptation

- Learn separate genre-specific translation model features
- Stream-based updates with suffix arrays [Levenberg et al. 2010]

Incremental Adaptation: Workflow

Initialize: baseline weights w_t

For each new sentence pair (f, e):

1. Stochastic weight update w_{t+1}

Incremental Adaptation: Workflow

Initialize: baseline weights w_t

For each new sentence pair (f, e):

1. Stochastic weight update w_{t+1}
2. align (f, e) via forced decoding or IBM-2

   $\Rightarrow$ obtain alignment a

Incremental Adaptation: Workflow

Initialize: baseline weights w_t

For each new sentence pair (f, e):

1. Stochastic weight update w_{t+1}
2. align (f, e) via forced decoding or IBM-2

   $\Rightarrow$ obtain alignment a
3. add (f, e, a) to genre-specific translation model corpora

Incremental Adaptation: Experiments

German→English training data

- 6.4M parallel sentences: WMT 2015 + Patents (PatTr)
- 4B monolingual English tokens

Test data

- News (WMT)
- Lectures (IWSLT)
- Patents (PatTR)

MT system

Phrasal w/ 5gm LM [Green et al. 2014]

Incremental Adaptation: BLEU (%) Results

	Lectures	News	Patents
baseline	25.8	24.9	49.0
+ genre weights	26.6 (+0.8)	25.1 (+0.2)	49.4 (+0.4)
+ genre TM	**27.7 (+1.9)**	**25.7 (+0.8)**	**53.2 (+4.2)**

Incremental Adaptation: BLEU (%) Results

	Lectures	News	Patents
baseline	25.8	24.9	49.0
+ genre weights	26.6 (+0.8)	25.1 (+0.2)	49.4 (+0.4)
+ genre TM	**27.7 (+1.9)**	**25.7 (+0.8)**	**53.2 (+4.2)**
+ document weights	**28.0 (+2.2)**	**25.7 (+0.8)**	**53.4 (+4.4)**

Incremental Adaptation: BLEU (%) Results

	Lectures	News	Patents
baseline	25.8	24.9	49.0
+ genre weights	26.6 (+0.8)	25.1 (+0.2)	49.4 (+0.4)
+ genre TM	**27.7 (+1.9)**	**25.7 (+0.8)**	**53.2 (+4.2)**
+ document weights	**28.0 (+2.2)**	**25.7 (+0.8)**	**53.4 (+4.4)**
+ sparse features	**28.1 (+2.3)**	**25.9 (+1.0)**	**54.3 (+5.3)**

Boldface: $p < 0.05$ vs. baseline

End of 2014: Open Problems

Online adaptation

- Not online, no domain adaptation
- Not tuned for prefixes

Prefix decoding was poor

- Low next word accuracy
- No diversity

UI issues

- Dropdown is distracting
- High learning curve

Prefix-Constrained Translation Inference

A user enters a prefix, MT system predicts the rest

Prefix-Constrained Translation Inference

A user enters a prefix, MT system predicts the rest

Example:
Yemeni media report that there is traffic chaos in the capital.

Once the user has typed:
Jemenitische Medien berichten von einem Verkehrschaos

The system suggests:
in der Hauptstadt.

Prefix-Constrained Translation Inference

A user enters a prefix, MT system predicts the rest

Example:
Yemeni media report that there is traffic chaos *in the capital.*

Once the user has typed:
Jemenitische Medien berichten von einem Verkehrschaos

The system suggests:
in der Hauptstadt.

- Align the prefix to the source to determine what remains

Prefix-Constrained Translation Inference

A user enters a prefix, MT system predicts the rest

Example:
Yemeni media report that there is traffic chaos in the capital.

Once the user has typed:
Jemenitische Medien berichten von einem Verkehrschaos

The system suggests:
in der Hauptstadt.

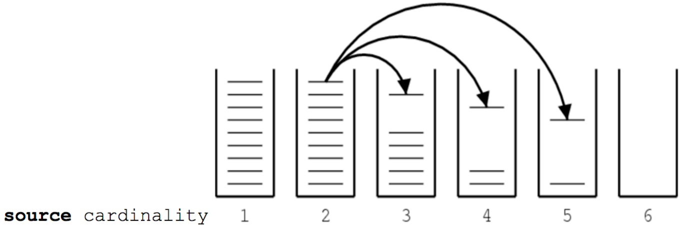
- Align the prefix to the source to determine what remains
- Predict a suffix for the prefix

Inference: Baseline

Prior work [Barrachina et al. 2008; Ortiz-Martínez et al. 2009]

- Standard beam search (force decoding)

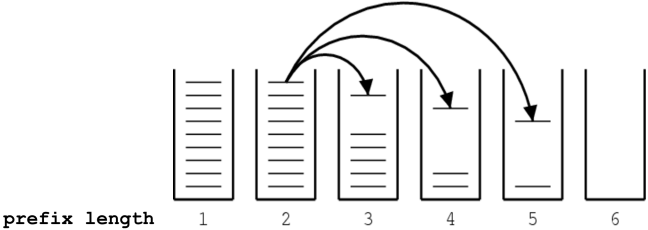
One beam per source cardinality



- Discard all hypotheses that violate prefix e_p

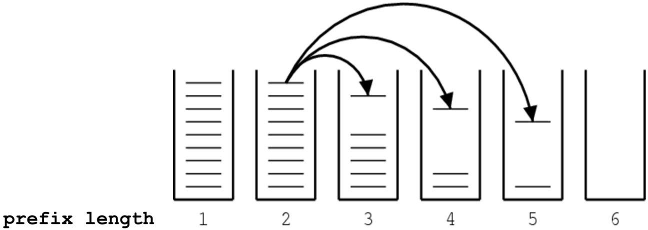
Inference: Target Beam Search

1. **Phrase alignment of source f and prefix e_p**
   - Associate each beam with *target* cardinality



Inference: Target Beam Search

1. **Phrase alignment of source f and prefix e_p**
   - Associate each beam with *target* cardinality



2. **Generate suffix e_s with standard beam search**
   - Copy partial hypotheses to source beams
   - Standard cube-pruning beam search

Prefix Tuning

Example:
Yemeni media report that there is traffic chaos in the capital.

Prefix: Jemenitische Medien berichten von einem Verkehrschaos
Suffix: in der Hauptstadt.

Prefix Tuning

Example:
Yemeni media report that there is traffic chaos in the capital.

Prefix: Jemenitische Medien berichten von einem Verkehrschaos
Suffix: in der Hauptstadt.

⇒ **Hierarchical, incremental tuning** [Wuebker et al. 2015]

Four feature domains:

- ROOT

Prefix Tuning

Example:
Yemeni media report that there is traffic chaos in the capital.

Prefix: Jemenitische Medien berichten von einem Verkehrschaos
Suffix: in der Hauptstadt.

⇒ **Hierarchical, incremental tuning** [Wuebker et al. 2015]

Four feature domains:

- ROOT
- PREFIX

Prefix Tuning

Example:
Yemeni media report that there is traffic chaos in the capital.

Prefix: Jemenitische Medien berichten von einem Verkehrschaos
Suffix: in der Hauptstadt.

⇒ **Hierarchical, incremental tuning** [Wuebker et al. 2015]

Four feature domains:

- ROOT
- PREFIX
- SUFFIX

Prefix Tuning

Example:
Yemeni media report that there is traffic chaos in the capital.

Prefix: Jemenitische Medien berichten von einem Verkehrschaos
Suffix: in der Hauptstadt.

⇒ **Hierarchical, incremental tuning** [Wuebker et al. 2015]

Four feature domains:

- ROOT
- PREFIX
- SUFFIX
- OVERLAP

Results: Phrase-based

English→French (newstest2014)

	pxBleu ↑		WPA ↑		KSR ↓	
baseline	40.9		38.0		61.7	
target beam	44.1	*+3.2*	49.4	*+11.4*	51.1	*-10.6*
prefix tuning	**44.7**	*+0.6*	**50.9**	*+1.5*	**50.5**	*-0.6*

pxBleu Prefix-Bleu (Bleu for the suffix only)

WPA Word Prediction Accuracy [Koehn et al. 2014]

KSR Key-Stroke-Ratio [Och et al. 2003]

Prefix NMT?

$$\log p(e|f) = \sum_{i=1}^{|e|} \log p(e_i|e_{<i}, f; \theta)$$

$$\theta = \arg\min_{\theta} \sum_{f,e} \sum_{i} -\log p(e_i|e_{<i}, f; \theta)$$

- $e_{<i}$ can be interpreted as target prefix e_p

Prefix NMT?

$$\log p(e|f) = \sum_{i=1}^{|e|} \log p(e_i | e_{<i}, f; \theta)$$

$$\theta = \arg\min_{\theta} \sum_{f,e} \sum_{i} - \log p(e_i | e_{<i}, f; \theta)$$

- $e_{<i}$ can be interpreted as target prefix e_p

Modification

- Condition on user prefix e_p instead of partial hypothesis

Results: Prefix NMT

English→German (autodesk)

	Bleu ↑		pxBleu ↑		WPA ↑		sec/seg
target beam	**44.5**		62.2		46.0		**0.051**
NMT single	40.6	*-3.9*	61.2	*-1.0*	52.3	*+6.3*	1.6
NMT ensem.	44.3	*-0.2*	**64.7**	*+2.5*	**54.9**	*+8.9*	7.7

NMT system: [Luong et al. 2015]

Results: Prefix NMT

English→German (autodesk)

	Bleu ↑		pxBleu ↑		WPA ↑		sec/seg
target beam	**44.5**		62.2		46.0		**0.051**
NMT single	40.6	*-3.9*	61.2	*-1.0*	52.3	*+6.3*	1.6
NMT ensem.	44.3	*-0.2*	**64.7**	*+2.5*	**54.9**	*+8.9*	7.7

NMT system: [Luong et al. 2015]

Recent NMT Work

Idea: add NMT as feature [Junczys-Dowmunt and Grundkiewicz 2016]

- en-de model of [Sennrich et al. 2016]

Recent NMT Work

Idea: add NMT as feature [Junczys-Dowmunt and Grundkiewicz 2016]

- en-de model of [Sennrich et al. 2016]

Reduced beam size: 25

Recent NMT Work

Idea: add NMT as feature [Junczys-Dowmunt and Grundkiewicz 2016]

- en-de model of [Sennrich et al. 2016]

Reduced beam size: 25

Virtualized CPU (Google Cloud); native ops in OpenBLAS

decoding cores	ms / scored token	ms / translated token
1	60	2,500
16	32	1,300

Interactive NMT Open Issues

Decoding time

- Distillation / pruning
- Reduced precision arithmetic

Interactive NMT Open Issues

Decoding time

- Distillation / pruning
- Reduced precision arithmetic

Periodic updates to model architecture

- PBMT: add feature, initialize to zero

Interactive NMT Open Issues

Decoding time

- Distillation / pruning
- Reduced precision arithmetic

Periodic updates to model architecture

- PBMT: add feature, initialize to zero

Dynamic updates to word-piece / segmentation model

- Hypothesis: high user sensitivity to dropped content words

End of 2014: Open Problems

Online adaptation

- Not online, no domain adaptation
- Not tuned for prefixes

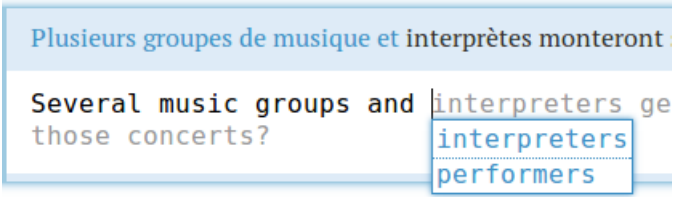
Prefix decoding was poor

- Low next word accuracy
- No diversity

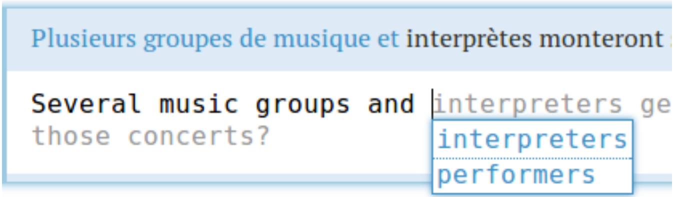
UI issues

- Dropdown is distracting
- High learning curve

Revisiting the Autocomplete Dropdown

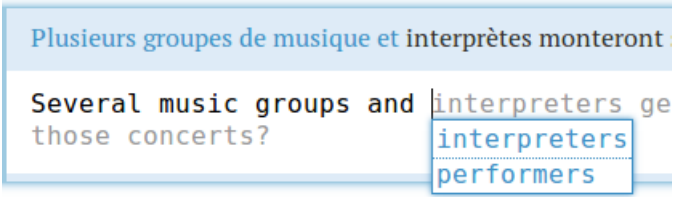


Revisiting the Autocomplete Dropdown



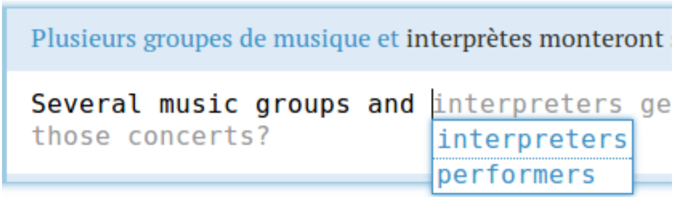
1. Autocomplete works best for limited vocabularies

Revisiting the Autocomplete Dropdown



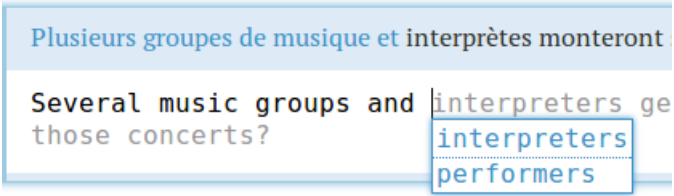
1. Autocomplete works best for limited vocabularies
2. **Distracting**: 100% visual acuity for only 4–5 characters

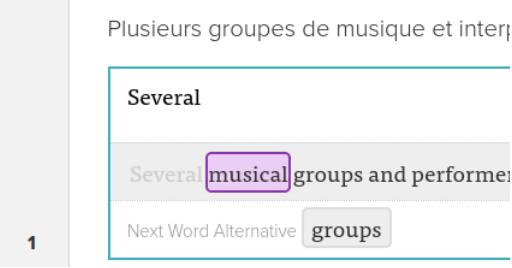
Revisiting the Autocomplete Dropdown



1. Autocomplete works best for limited vocabularies
2. **Distracting**: 100% visual acuity for only 4–5 characters
3. No post-edit mode

2014 vs. 2016


Plusieurs groupes de musique et interprètes monteront
Several music groups and interpreters ge
those concerts?
interpreters
performers


1
Plusieurs groupes de musique et inter
Several
Several musical groups and performe
Next Word Alternative
groups

(Demo)

Interactive Machine Translation: From Research to Practice

Spence Green

spence@lilt.com

30 October 2016

References I

Barrachina, Sergio et al. (2008). "Statistical approaches to computer-assisted translation". In: *Computational Linguistics* 35.1, pp. 3–28.

Daumé III, Hal (2007). "Frustratingly Easy Domain Adaptation". In: *ACL*.

Duchi, J., E. Hazan, and Y. Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *JMLR* 12, pp. 2121–2159.

Green, Spence, Daniel Cer, and Christopher D. Manning (2014). "Phrasal: A Toolkit for New Directions in Statistical Machine Translation". In: *WMT*.

Horvitz, Eric (1999). "Principles of Mixed-initiative User Interfaces". In: *CHI*. Pittsburgh, Pennsylvania, USA. isbn: 0-201-48559-1. doi: 10.1145/302979.303030.

Junczys-Dowmunt, Marcin and Roman Grundkiewicz (2016). "Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing". In: *Proceedings of the First Conference on Machine Translation*.

Kay, Martin (1980). *The Proper Place of Men and Machines in Language Translation*. Tech. rep. CSL-80-11. Xerox Palo Alto Research Center (PARC), pp. 3–23.

Koehn, Philipp, Chara Tsoukala, and Herve Saint-Amand (2014). "Refinements to Interactive Translation Prediction Based on Search Graphs". In: *ACL*.

Langlais, Philippe and Guy Lapalme (2002). "TransType: Development-Evaluation Cycles to Boost Translator's Productivity". English. In: *Machine Translation* 17.2, pp. 77–98. issn: 0922-6567. doi: 10.1023/B:COAT.0000010117.98933.a0.

References II

- Levenberg, Abby, Chris Callison-Burch, and Miles Osborne (2010). "Stream-based Translation Models for Statistical Machine Translation". In: *NAACL*.
- Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning (2015). "Effective Approaches to Attention-based Neural Machine Translation". In: *EMNLP*.
- O'Brien, S. and J. Moorkens (2014). "Towards intelligent post-editing interfaces". In: *20th FIT World Congress*.
- Och, Franz Josef, Richard Zens, and Hermann Ney (2003). "Efficient Search for Interactive Statistical Machine Translation". In: *EACL*.
- Ortiz-Martínez, D., I. García-Varea, and F. Casacuberta (2009). "Interactive Machine Translation Based on Partial Statistical Phrase-based Alignments". In: *RANLP*.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Edinburgh Neural Machine Translation Systems for WMT 16". In: *Proceedings of the First Conference on Machine Translation*.
- Wuebker, Joern, Spence Green, and John DeNero (2015). "Hierarchical Incremental Adaptation for Statistical Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1059–1065.