
Computer Aided Translation

Philipp Koehn

26 April 2016



Why Machine Translation?



Assimilation — reader initiates translation, wants to know content

- user is tolerant of inferior quality
- focus of majority of research■

Communication — participants don't speak same language, rely on translation

- users can ask questions, when something is unclear
- chat room translations, hand-held devices
- often combined with speech recognition■

Dissemination — publisher wants to make content available in other languages

- high demands for quality
- currently almost exclusively done by human translators

Why Machine Translation?



Assimilation — reader initiates translation, wants to know content

- user is tolerant of inferior quality
- focus of majority of research

Communication — participants don't speak same language, rely on translation

- users can ask questions, when something is unclear
- chat room translations, hand-held devices
- often combined with speech recognition

Dissemination — publisher wants to make content available in other languages

- high demands for quality
- currently almost exclusively done by human translators

Goal: Helping Human Translators



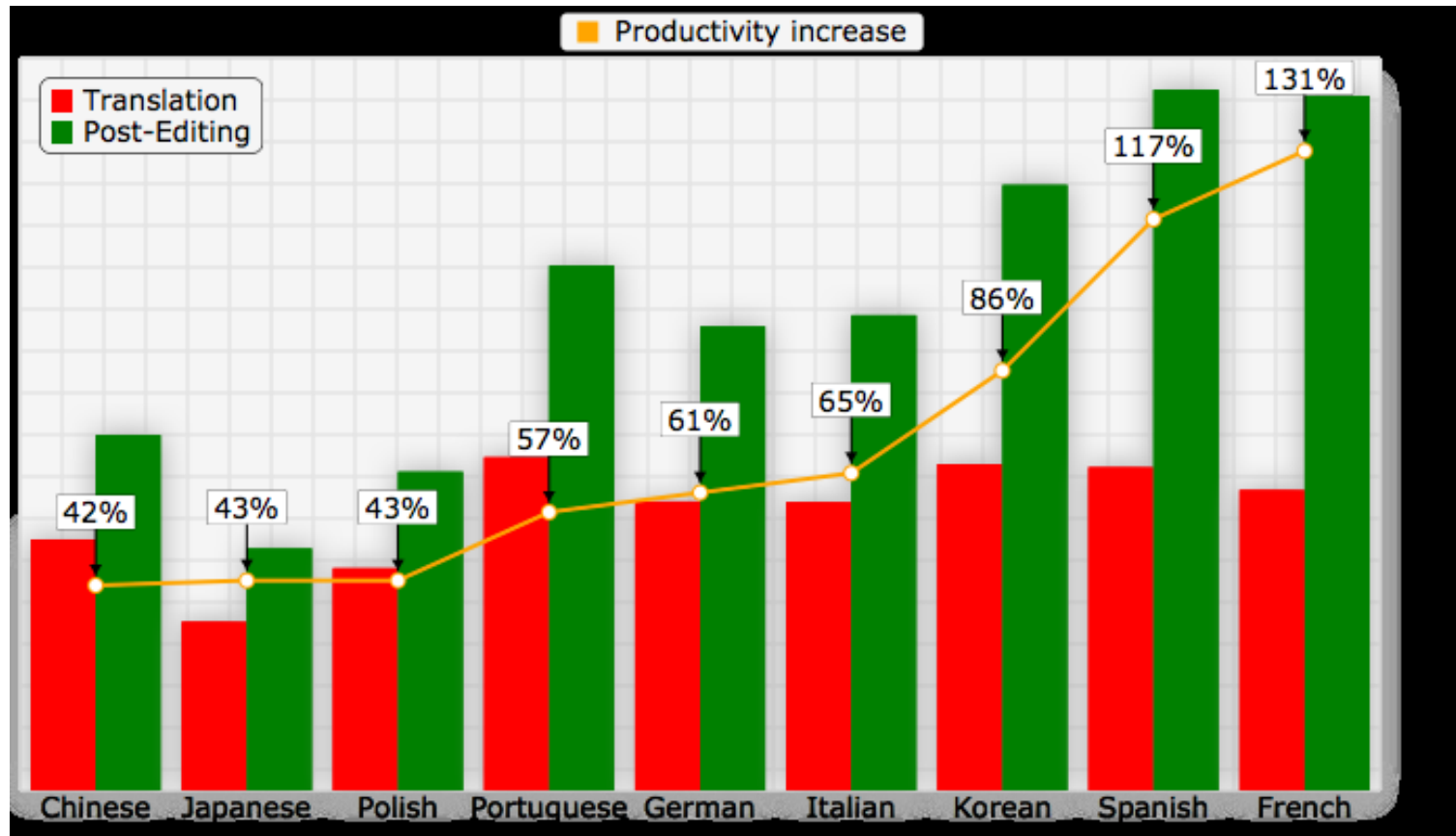
If you can't beat them, join them.■

→ How can machine translation help human translators?

Post-Editing Machine Translation



4



(source: Autodesk)

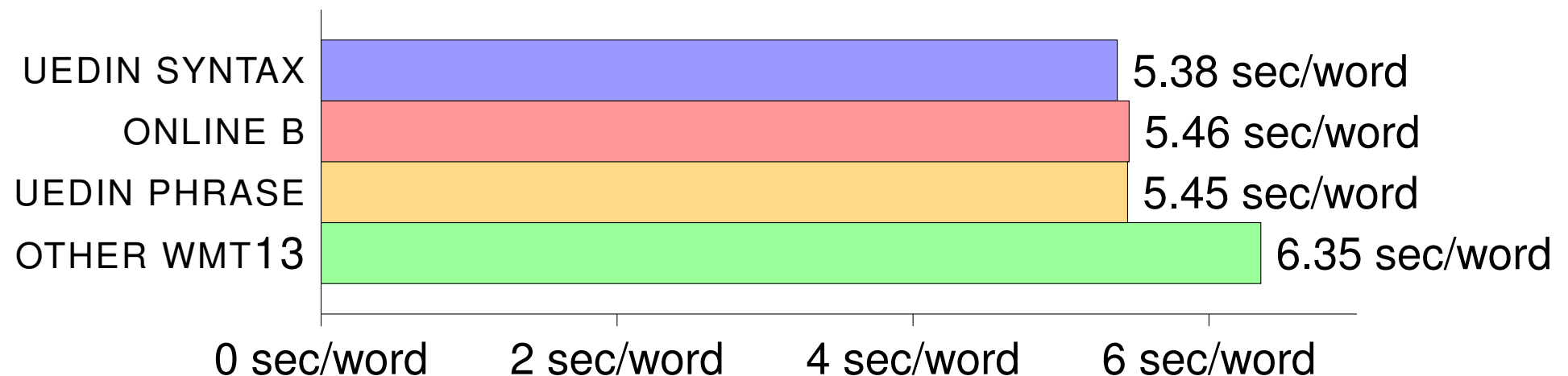
Machine Translation Quality Matters



Experiment:

Post-editing with different machine translation systems

English–German, news stories



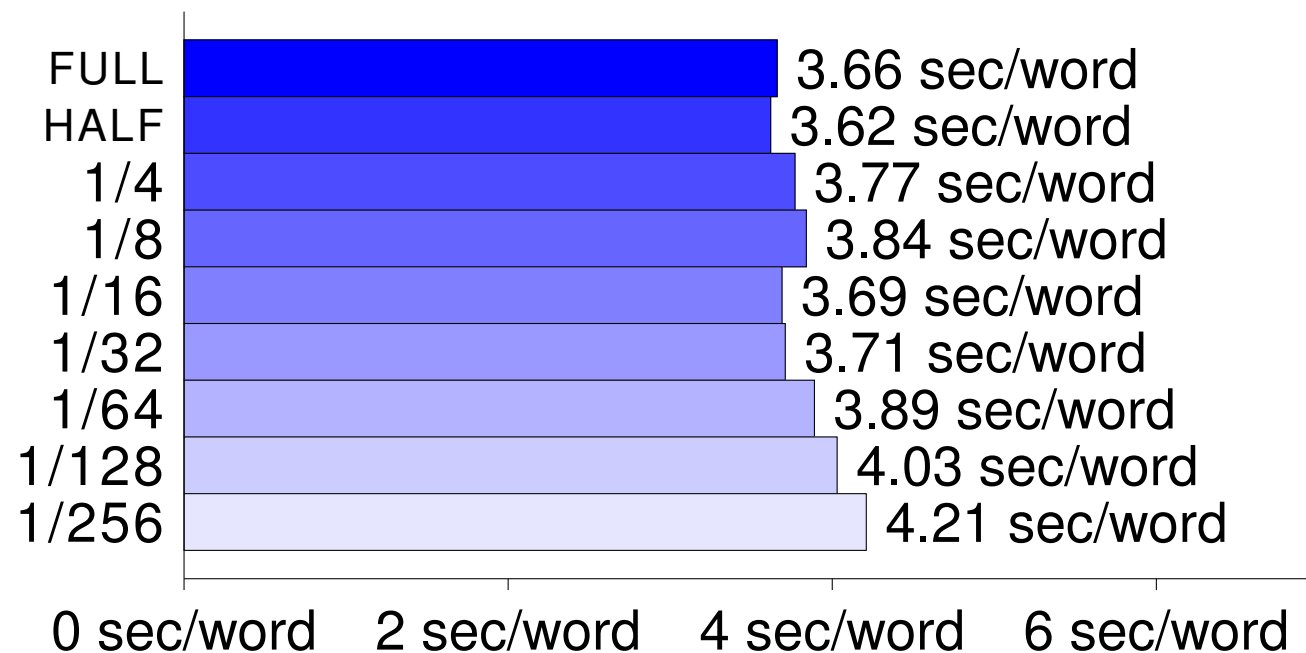
[Koehn and Germann, 2014]

Machine Translation Quality Matters



Experiment:

Post-editing with system trained on different data size
Spanish–English, news stories



[Torrón et al., 2016]

- **Interactivity**
- Choices
- Confidence
- Adaptation

- Traditional professional translation approaches
 - translation from scratch
 - post-editing translation memory match
 - post-editing machine translation output
- More interactive collaboration between machine and professional?

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

|

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

| He

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

He | has

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

He has | for months

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

He planned |

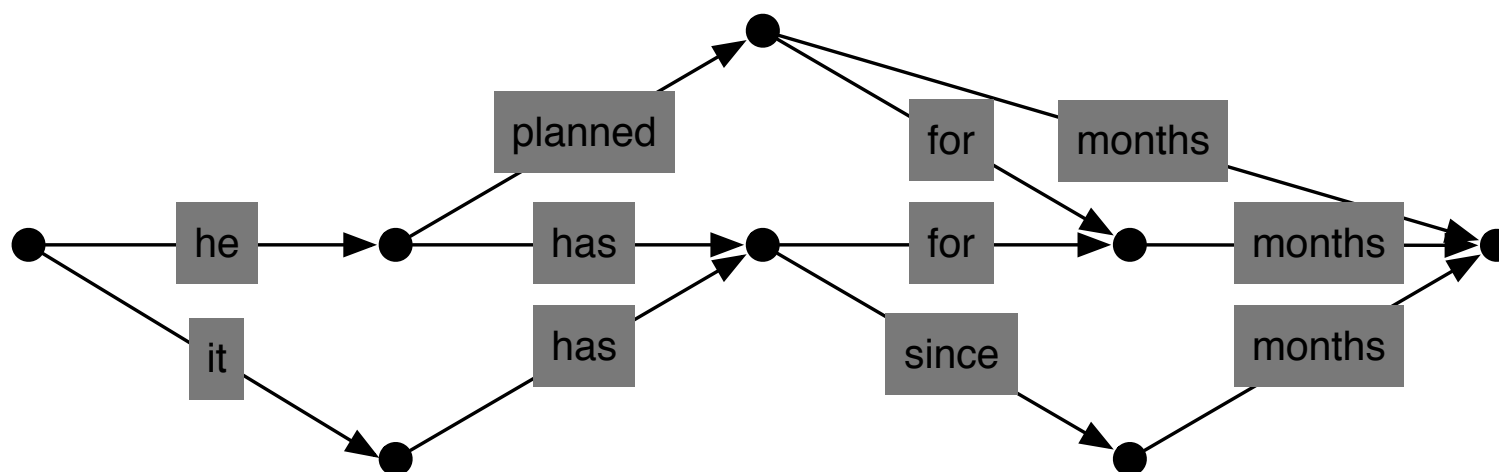
Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

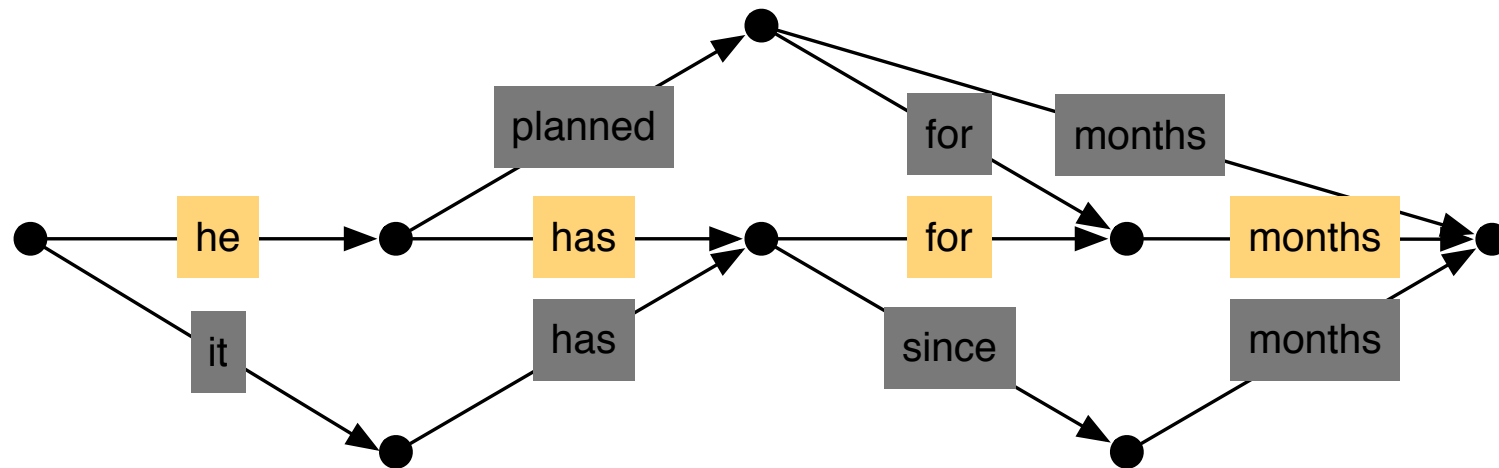
He planned | for months

Prediction from Search Graph



Search for best translation creates a graph of possible translations

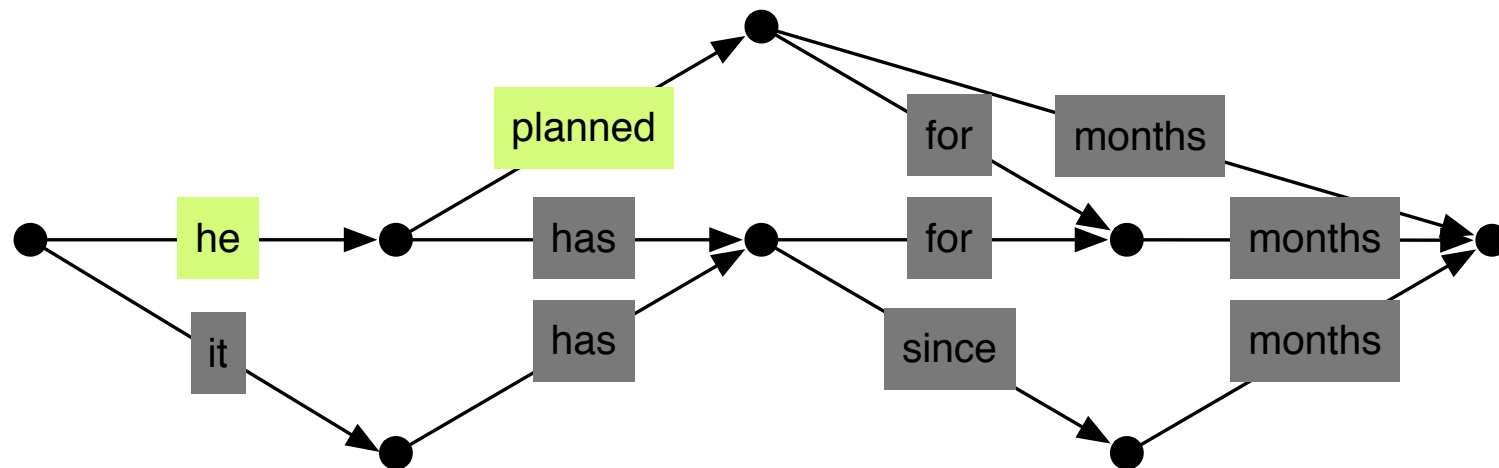
Prediction from Search Graph



One path in the graph is the best (according to the model)

This path is suggested to the user

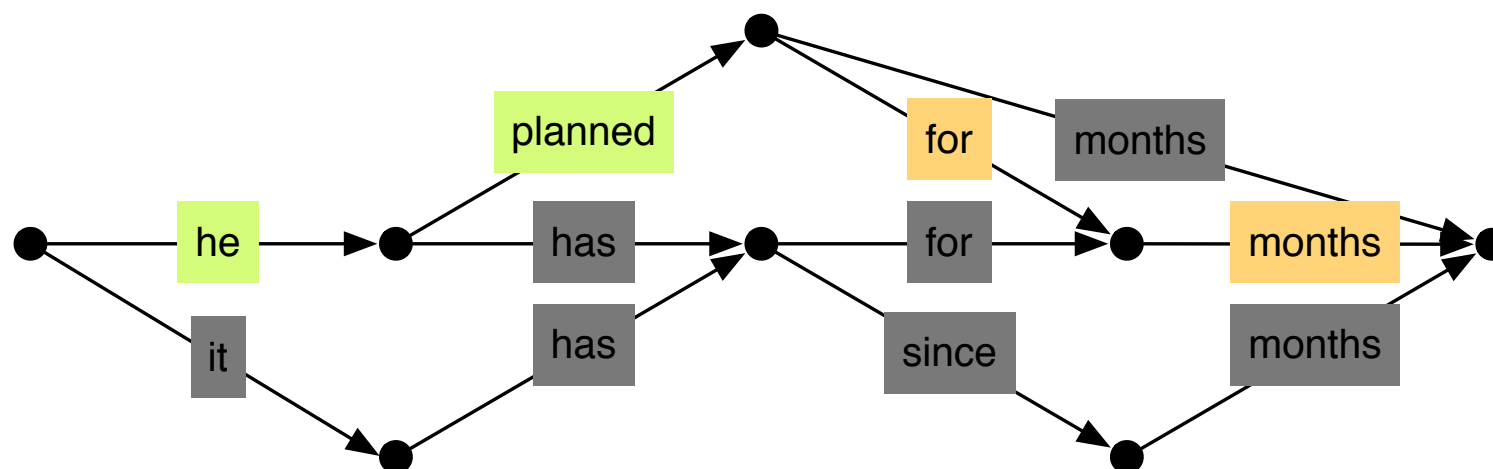
Prediction from Search Graph



The user may enter a different translation for the first words

We have to find it in the graph

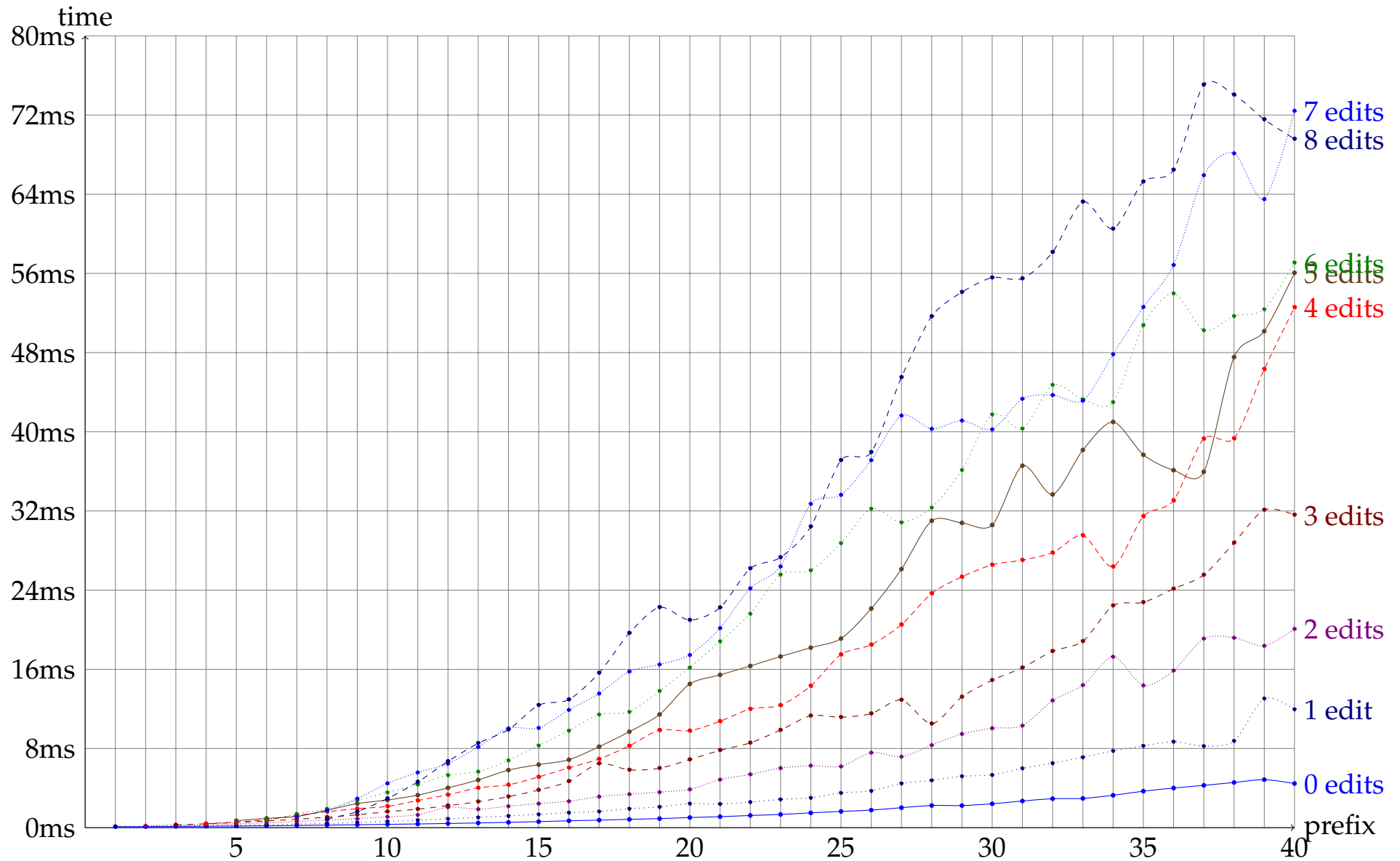
Prediction from Search Graph



We can predict the optimal completion (according to the model)

Run Time

19



Word Alignment Visualization

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

He planned for months to give a lecture in Baltimore | in

Input Sentence

Er hat seit Monaten geplant, **im** April einen Vortrag in Baltimore zu halten.

Professional Translator

He planned for months to give a lecture in Baltimore | **in**

Shading off Translated Material

Input Sentence

Er hat seit Monaten geplant, **im** April einen Vortrag in Baltimore zu halten .

Professional Translator

He planned for months to give a lecture in Baltimore | **in**

- How can we do this?
 - word alignments by-product of matching against search graph
 - automatic word alignments (as used in training)
- User feedback
 - users like interactive machine translation
 - ... but they may be slower than with post-editing machine translation
 - user like mouse-over word alignment highlighting
 - user do not like at-cursor word alignment highlighting

- Interactivity
- **Choices**
- Confidence
- Adaptation

- Trigger the passive vocabulary
- Display multiple translations for words and phrases

er	hat	seit	Monaten	geplant	,	im	März	einen	Vortrag	...
he has		for months		the plan		in March		a lecture		...
it has		for months now		planned	,	in	March	a presentation		...
he was		for several months		planned to		in the March		a speech		...
he has made		since	months	the pipeline		in March of		a statement		...
he did		for many months		scheduled		the March		a general		...

- Rank and color-highlight by probability of each translation
- Prefer diversity

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Professional Translator

He planned for months to **give a lecture** in Baltimore in April.

give a presentation

present his work

give a speech

speak

User requests alternative translations for parts of sentence.

Bilingual Concordancer

27



entre autres(560/1554)

...d and made recommendations , " **inter alia** " , with respect to the follow...
...on (EC) No 1995 / 2000 imposing , **inter alia** , a definitive anti @-@ dumping dut...
...ervices . this increase , arising , **inter alia** , as a result of economic growth , ...
...of paragraph 1 the Commission may , **inter alia** , bring forward :
... of stocks of obsolete pesticides , **inter alia** , by supporting projects aimed at s...
...wn rules of procedure which shall , **inter alia** , contain provisions for convening ...
...uch specific agreements may cover , **inter alia** , financing provisions , assignment...
...he internal market and concerning , **inter alia** , health and environmental protecti...
...e product concerned) originating , **inter alia** , in Belarus and Russia (the count...
...e product concerned) originating , **inter alia** , in India .

... des recommandations concernant , **entre autres** , les questions spécifiques suiva...
...995 / 2000 du Conseil instituant , **entre autres** , un droit antidumping définitif ...
...nsports . cette augmentation , due **entre autres** facteurs à la croissance économi...
...aragraphe 1 , la Commission peut , **entre autres** , présenter :
...r les stocks de vieux pesticides , **entre autres** en soutenant des projets à cet ef...
...lement intérieur , qui contient , **entre autres** dispositions , les modalités de c...
...ords spécifiques peuvent porter , **entre autres** , sur les mécanismes financiers s...
...hé intérieur et qui concernent , **entre autres** , la santé et la protection de l&...
...it concerné ") originaire , **entre autres** , du Belarus et de Russie (ci @-@ ...
...t concerné ") originaires , **entre autres** , de l ' Inde .

notamment(447/1554)

... the EU budget by addressing " **inter alia** " the problems of accountabili...
...ates , the Commission has adopted , **inter alia** , Decision 2003 / 526 / EC (3) wh...
...d equitable development involving , **inter alia** , access to productive resources , ...
...ertain products which could be used **inter alia** , as equipment on board ships but w...
...nexes , taking into consideration , **inter alia** , available scientific , technical ...
...w that it is absolutely necessary , **inter alia** , because of enlargement , to find ...
...paragraphs 1 and 2 as appropriate , **inter alia** , by conducting studies and compili...
...liability and efficiency , caused , **inter alia** , by insufficient technical and adm...
...in the Programme shall be pursued , **inter alia** , by the following means :

...get de l' Union , ce qui passe **notamment** par la résolution du problème de r...
...es États membres , la Commission a **notamment** arrêté la décision 2003 / 526 / C...
... durable et équitable , impliquant **notamment** l' accès aux ressources produc...
...usceptibles d' être utilisés **notamment** comme équipements mis à bord , mai...
...ion et à ses annexes , compte tenu **notamment** des informations scientifiques , tec...
...os ; il est absolument nécessaire , **notamment** en raison de l' élargissement ...
...rgraphes 1 et 2 le cas échéant , **notamment** en menant des études et en compilan...
... et d' efficacité en raison , **notamment** , d' une interopérabilité tec...
...nis dans le programme , il convient **notamment** de mettre en oeuvre les moyens ci @-...

- Interactivity
- Choices
- **Confidence**
- Adaptation

- Machine translation engine indicates where it is likely wrong (also known as quality estimation — Lucia Specia)■
- Different Levels of granularity
 - document-level (SDL's "TrustScore")
 - sentence-level
 - word-level■
- What are we predicting?
 - how useful is the translation — on a scale of (say) 1–5
 - indication if post-editing is worthwhile
 - estimation of post-editing effort
 - pin-pointing errors

- Translators are used to “Fuzzy Match Score”
 - used in translation memory systems
 - roughly: ratio of words that are the same between input and TM source
 - if less than 70%, then not useful for post-editing
- We would like to have a similar score for machine translation■
- Even better
 - estimation of post-editing time
 - estimation of from-scratch translation time
 - can also be used for pricing
- Active research question, see also shared task at WMT 2013

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Machine Translation

He has for months planned in April give a lecture in Baltimore.

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

Machine Translation

He has for months planned in April give a lecture in Baltimore.

Note: different color for wrong words and reordered words
(inserted words? missing words?)

- Can we identify errors in human translations?
 - missing / added information
 - inconsistent use of terminology

Input Sentence

Er hat seit Monaten geplant, im April einen Vortrag in Baltimore zu halten.

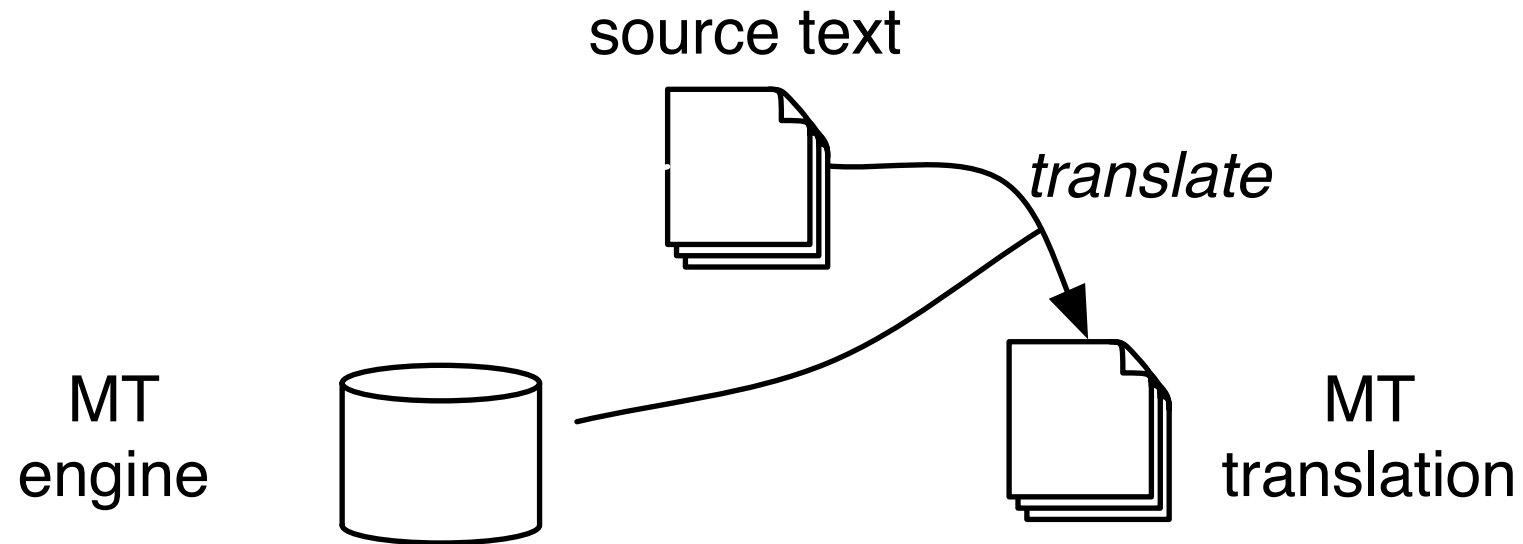
Human Translation

Moreover, he planned for months to give a lecture in Baltimore.

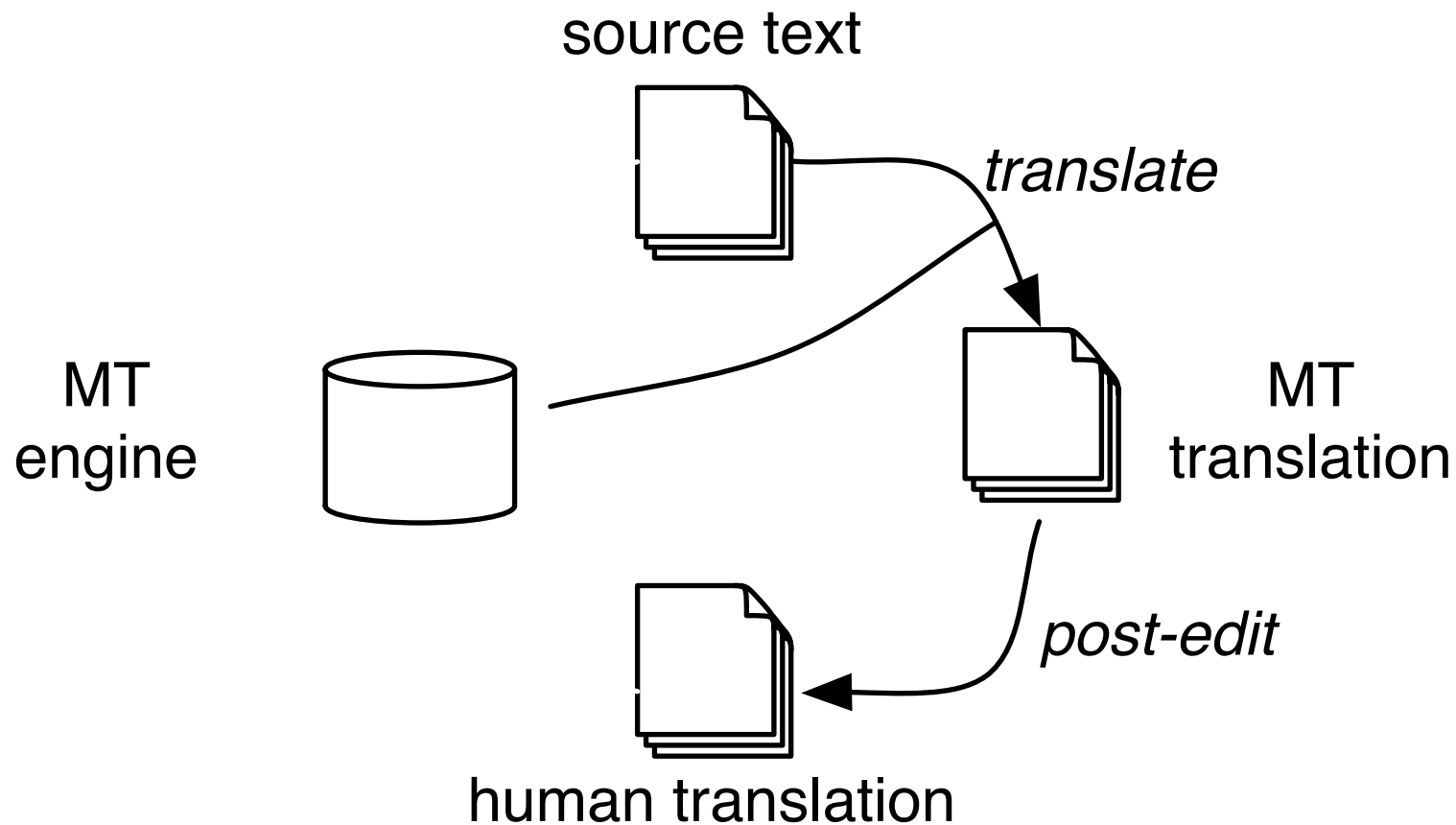
- Interactivity
- Choices
- Confidence
- **Adaptation**

- Machine translation works best if optimized for domain
- Typically, large amounts of out-of-domain data available
 - European Parliament, United Nations
 - unspecified data crawled from the web
- Little in-domain data (maybe 1% of total)
 - information technology data
 - more specific: IBM's user manuals
 - even more specific: IBM's user manual for same product line from last year
 - and even more specific: sentence pairs from current project
- Various domain adaptation techniques researched and used

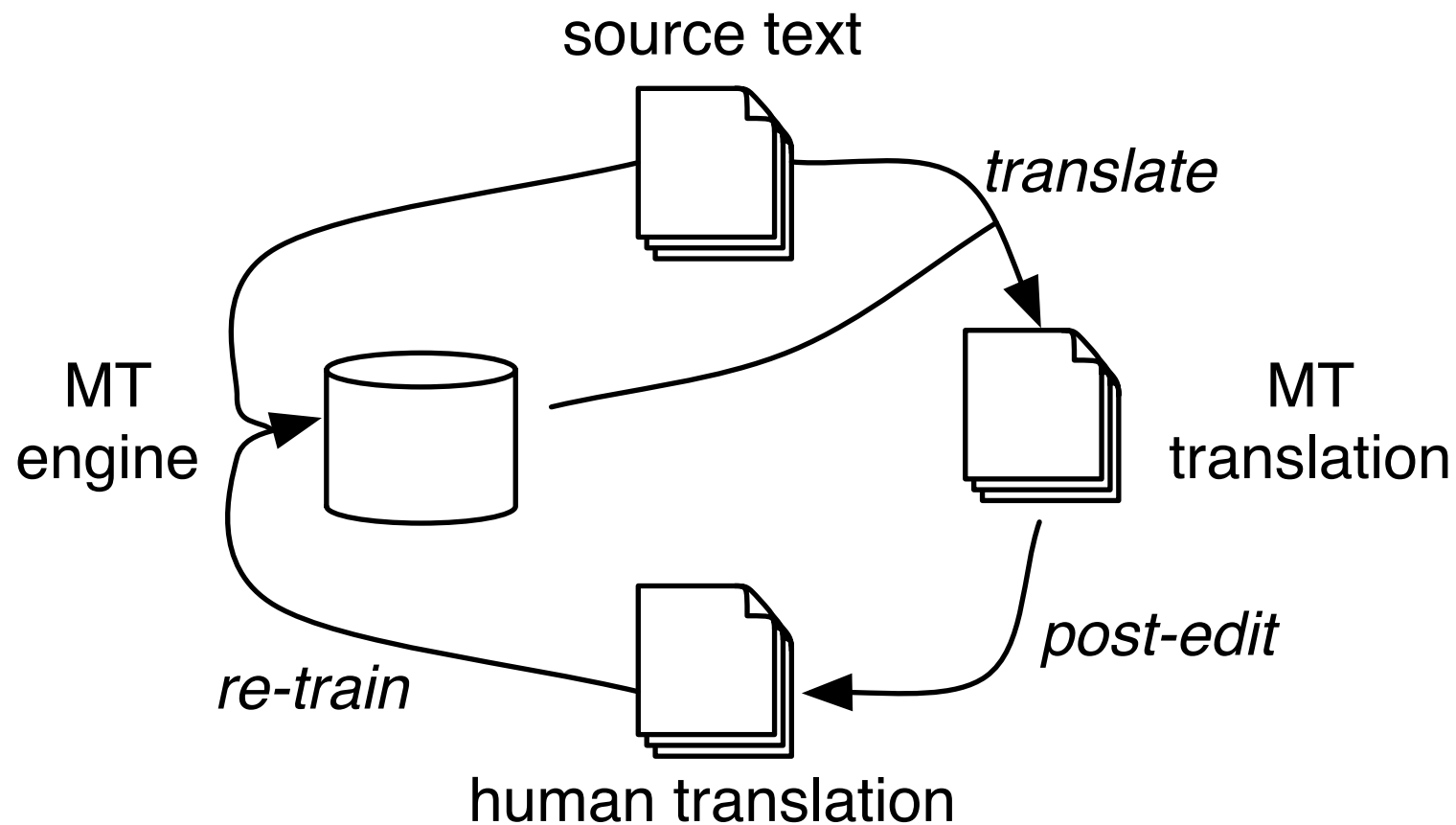
Incremental Updating



Incremental Updating



Incremental Updating



Updatable Translation Table

- Required: quickly add a sentence pair to the translation table
 - word alignment
 - phrase extraction
 - phrase table building
- Online word alignment
 - use existing model to align new sentence pair
- Online phrase table building
 - Store parallel corpus in memory
 - Index corpus with suffix array
 - Extract phrases on the fly
- This can be done in less than 1 second

Suffixes

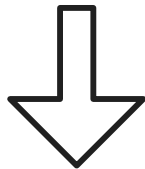
- 1 government of the people , by the people , for the people
- 2 of the people , by the people , for the people
- 3 the people , by the people , for the people
- 4 people , by the people , for the people
- 5 , by the people , for the people
- 6 by the people , for the people
- 7 the people , for the people
- 8 people , for the people
- 9 , for the people
- 10 for the people
- 11 the people
- 12 people

Sorted Suffixes

5 , by the people , for the people
9 , for the people
6 by the people , for the people
10 for the people
1 government of the people , by the people , for the people
2 of the people , by the people , for the people
12 people
4 people , by the people , for the people
8 people , for the people
11 the people
3 the people , by the people , for the people
7 the people , for the people

Suffix Array

5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people



suffix array: sorted index of corpus positions

Querying the Suffix Array

5

, by the people , for the people

9

, for the people

6

by the people , for the people

10

for the people

1

government of the people , by the people , for the people

2

of the people , by the people , for the people

12

people

4

people , by the people , for the people

8

people , for the people

11

the people

3

the people , by the people , for the people

7

the people , for the people

Query: people

Querying the Suffix Array

5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people

Query: people

Binary search: start in the middle

Querying the Suffix Array

5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	→ of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people

Query: people

Binary search: discard upper half

Querying the Suffix Array

5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people

Query: people

Binary search: middle of remaining space

Querying the Suffix Array

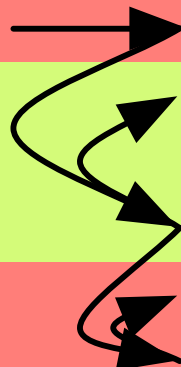
5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people

Query: people

Binary search: match

Querying the Suffix Array

5	, by the people , for the people
9	, for the people
6	by the people , for the people
10	for the people
1	government of the people , by the people , for the people
2	of the people , by the people , for the people
12	people
4	people , by the people , for the people
8	people , for the people
11	the people
3	the people , by the people , for the people
7	the people , for the people



Query: people

Finding matching range with additional binary searches for start and end

- Phrase translation probability

$$p(\bar{e}|\bar{f}) = \frac{\text{count}(\bar{e}, \bar{f})}{\text{count}(\bar{f})}$$

- Suffix array allows quick estimation of $\text{count}(\bar{f})$
- By looping through the \bar{f} , the $\text{count}(\bar{e}, \bar{f})$ can be collected as well
- If there are too many \bar{f} , we can resort to sampling

Summary

- Use of machine translation in translation industry becomes standard
- Interaction between machine and human open problem
- Not very much research in this area
- Open source toolkit: CASKAT