

Searchable Translation Memories

Chris Callison-Burch

LINEAR B

What's wrong with translation memories?

- A very valuable tool:
 - Productivity gains
 - Increased consistency
 - Simplified quality control
 - Improved terminology management
- However, limited to sentence / term retrieval

The Future of TMs

- "Certainly there are improvements that can be made to current TM applications and their underlying technologies ... the next new wave of language technology innovation will come from extending the recycling of translated material to different levels of granularity, and to combine them. Today we mainly recycle on the term level, sentence level, or document level. In the future, we want to extend this to recycling on the phrase level. This will require its own smart linguistic algorithms."

(Trados CTO, article in Clientside News)

Searchable Translation Memories

- Tools that allow Google-style searching of translation memories.
- Find past translations which are not normally retrieved in current TMs
- Better exploit knowledge in TM. Further increase consistency, improved productivity, etc.

Talk Overview

- Motivation
- Linear B's Statistical Machine Translation
- Technology for Searchable Translation Memories
- Demo
- Summary
- Sneak Peek

Linear B: Who we are

- Linear B was founded in October 2002 by Informatics PhD students at the University of Edinburgh
- Funded by venture capital and a SMART technology grant from UK government
- Linear B creates statistical machine translation software

Statistical Machine Translation

- Data-driven - Automatically learns to translate from example translations
- Eliminates the need for a staff of linguists to hand-craft rules
- Can be applied to any language pair

Parallel corpus

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Advantages of SMT

- Can be applied to any language pair
- Quick to develop
- Improves as more data becomes available
- (Recently) High-quality

Example Source

- Honorables sénateurs, tandis que la guerre en Irak entre dans sa troisième semaine, nous ne devons pas oublier qu'il faut prendre des mesures pour éviter une crise humanitaire dans la population civile. À cet égard, il y a de nombreux domaines dans lesquels les Canadiens doivent faire preuve de leadership.

Maintenant que la guerre fait rage en Irak et que des pénuries de produits alimentaires et de fournitures médicales commencent à se produire, les pays qui ont accès à des ressources ont la responsabilité d'essayer de minimiser les effets négatifs du conflit sur la population irakienne. Je crois personnellement que le Canada devrait jouer un plus grand rôle à cet égard.

Au Canada, nous disposons en abondance de blé et d'autres produits alimentaires. Nous pouvons également fournir des articles médicaux aux citoyens de l'Irak. Le Canada a une fière réputation pour ce qui est d'offrir une aide humanitaire aux gens quand ils en ont besoin.

Word-based SMT

- Honourable senators, while that the war in Iraq between in his third week, we not must not forget that it must take of measures to avoid a crisis humanitarian in the people calendar. In many areas which Canadians must show leadership in this regard

Now that the war fact raging in Iraq and that of shortages of products food and of supplies medical beginning to be produce, the country which have access to of resources have the responsibility of try of minimize the effects negative of conflict on the people irakienne. I personally believe this regard great role Canada should play more

In other food products wheat Canada have abundant We can provide the citizens medical articles Iraq The offer Canada has a reputation proud

Linear B's Translation

- Honourable senators, while the war in Iraq extending into a third week, the minister should not forget the one we must take some steps to prevent a crisis humanitarian face in the civilian populations. In this regard, there are a number of areas in which the Canadians must concentrate to show leadership.

That the war is raging in Iraq and that a shortage of food and medical supplies are beginning to take place, those countries that have access to the resources are the responsibility for doing try to minimize the negative effects on workers on the people Irakienne. I personally believe that Canada should play a more significant role in this regard.

In Canada, we have in abundance of wheat, as have other food products. We can also provision of medical articles to the people on the other Iraq. Canada has a proud record in so far, as would be to give humanitarian assistance to people when they need it.

Large Units of Human Translated Text

Je crois personnellement que le	Canada devrait jouer	un plus grand rôle	à cet égard.
---------------------------------	----------------------	--------------------	--------------

I personally believe that	Canada should play	a more significant role	in this regard.
---------------------------	--------------------	-------------------------	-----------------

Synergy Between SMT and TMs

- Quality of statistical MT improves when large TMs are available
- Better chance of retrieving larger blocks of human translated text
- Techniques used in statistical MT can be used to align phrases within TMs
- Might be used to increase usefulness of TMs

Statistical MT to Searchable TMs

- The technology allows us to build an index of phrasal translations
- Like a concordance, but also lets us identify the translation of a phrase

Index of Phrases 2

	we	owe	it	to	the	taxpayers	to	keep	the	costs	in	check
wir	■										■	■
sind											■	■
es			■								■	■
den				■	■						■	■
steuerzahlern						■					■	■
schuldig		■									■	■
die									■		■	■
kosten										■	■	■
unter	■	■	■	■	■	■	■	■	■	■	■	■
kontrolle	■	■	■	■	■	■	■	■	■	■	■	■
zu						■						
haben							■					

Highlighting Translations

- An index of phrasal translation allows us to highlight the translation that someone is looking for:

Search for translations of *paiement initial*

Down Payment

Ils n'ont pas les quelques milliers de dollars nécessaires pour

le **paiement initial** afin d'acheter leur propre maison.

They do not have the several thousand dollars needed for

a **down payment** on their own shelter.

Ranking Translations

- Knowing how many times a translation occurred in the index lets us rank results by frequency:

Search for translations of *paiement initial*

Initial Payment - 15 sentences matched

Down Payment - 8 sentences matched

Initial Payments - 1 sentence matched

Demo

Evaluation

- Created a "gold standard" set of correct translations for 120 phrases that occurred in a translation memory containing 50,000 German-English sentences
- Evaluated the correctness of the suggestions that our software presents, and how many of the correct translations it managed to find

Results

- 78% of phrases were correct
- 81% of the "gold standard" translation were retrieved
- The top translation that our system returned for a phrase was correct 87% of the time

Conclusion

- Presented Linear B's first foray into aides to the human translation process
- Technology developed for statistical machine translation can be used to create searchable translation memories
- Allows us to better exploit the knowledge contained within TMs

Future Directions

- Preview of Linear B's coming technology:
 - Improved machine translation through post-editing
 - Human-aided machine translation

Improved Translation Through Post-editing

1. 電子メール
/Fax/郵送にて校正
原稿をお送りくださ
い。
2. お見積、納期をお
客様にお知らせいた
し

Linear B Machine Translation

1. Please send a
proofreading
manuscript by the
E-mail / Fax /
mailing.
2. I announce a
visitor an estimate
and time for
delivery.

they post-edit

1. Please send a
manuscript for
proofreading by e-
mail, fax, or post.
2. You will receive
a notification of
receipt containing a
time estimate

Manuscript sent
to client

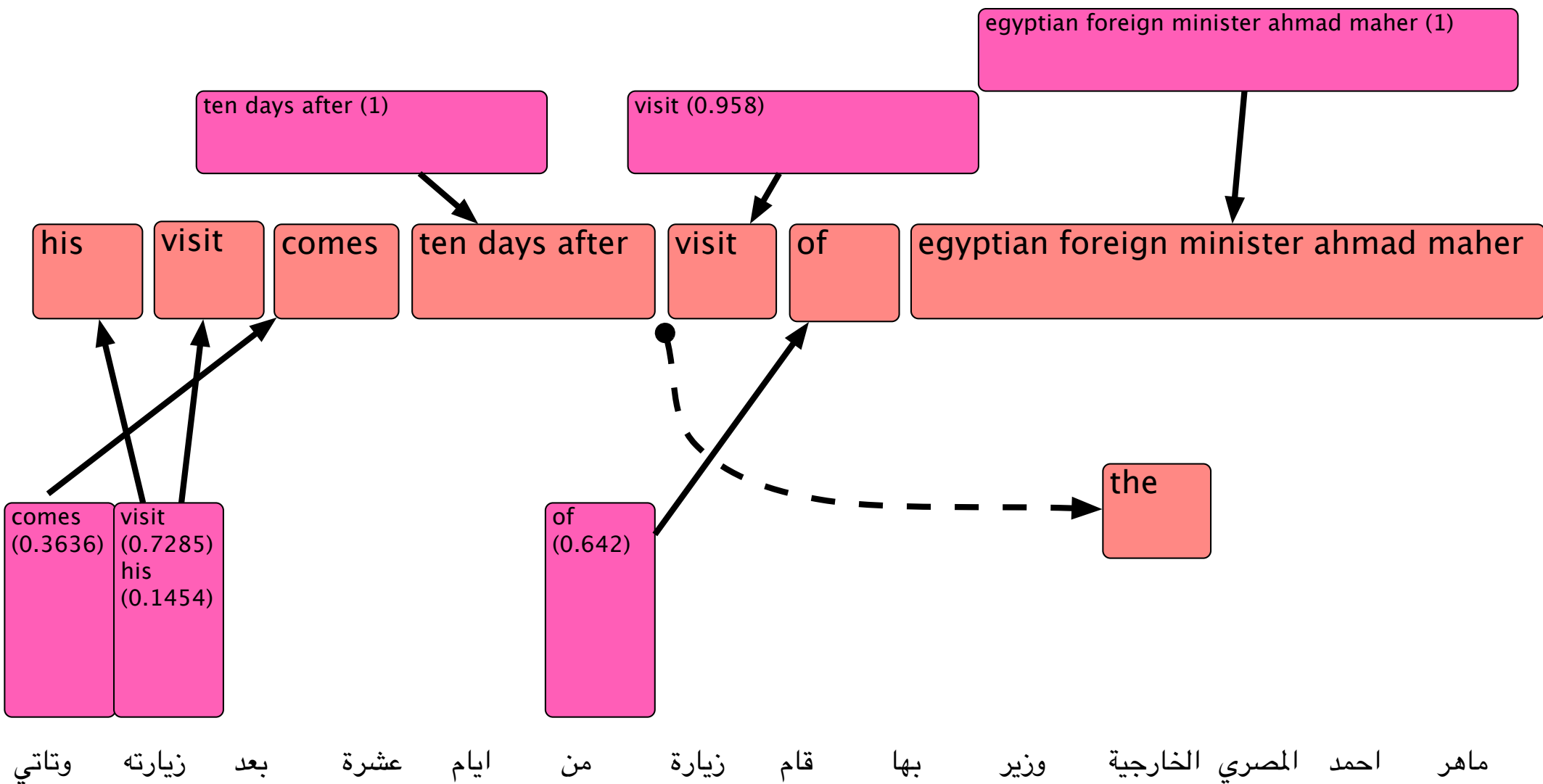
1. 電子メール
/Fax/郵送にて校正
原稿をお送りくださ
い。
2. お見積、納期をお
客様にお知らせいた
し

1. Please send a
manuscript for
proofreading by e-
mail, fax, or post.
2. You will receive
a notification of
receipt containing a
time estimate

Parallel text sent to
Linear B for retraining

Human-aided Statistical Machine Translation

Manual re-ordering



Resulting Translation

- Ramallah (West Bank) 1/1 (AFP) An Agence France-Presse correspondent said that the Egyptian presidential advisor Osama Baz arrived on Thursday in Ramallah for talks with Palestinian authority president Yasser Arafat. Minister Hun and Authority Chairman Katif participated in talks and negotiations with Interior Minister Saib Ariqat. Baz refused to make statements immediately upon arrival to the province. ***His visit comes ten days after the visit of Egyptian foreign minister Ahmed Maher.*** Baz began his visit to Ramallah last August with a new truce idea for Palestinians and Israel, however Palestinian factions have not agreed to this Egyptian sponsored proposal. It is expected that Egyptian endeavours will continue next week with a planned visit by the head of general intelligence.