Directed MT Research for Commercial Settings

Adrià de Gispert

13 September 2016

About me

- 2002-06: PhD at Univ. Politècnica de Catalunya (UPC, Barcelona)
  - Ngram-based SMT, translation into morphologically-rich languages

- 2007-12: post-doc at Univ. of Cambridge (UK)
  - Hierarchical phrase-based SMT using finite-state automata (large lattices), minimim bayes risk decoding, lattice-based confidence metrics, push-down automata for MT, hiero grammar design,...

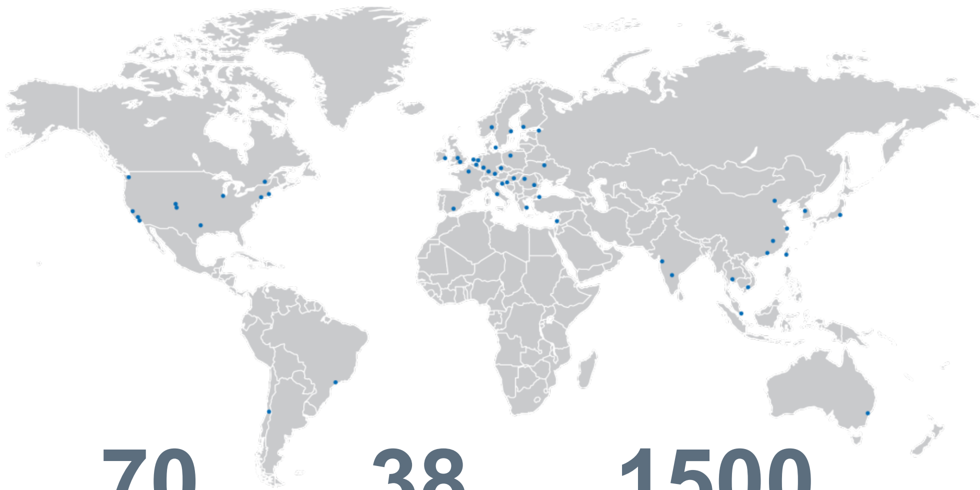
- 2012-today: research scientist at SDL Research (UK)
  - bringing research ideas to actual MT products

2700
Employees Worldwide

Founded in 1992

Publicly traded company (LSE:SDL)

$430M annual revenue



70 offices

38 countries

1500 enterprise customers

SDL

Leading Global Language and Content Capabilities

- Analytics
- Social
- Campaigns
- eCommerce
- Language
- Web
- Documentation

Leader in professional translation for over 20 yrs

Powering marketing campaigns for 400+ global brands

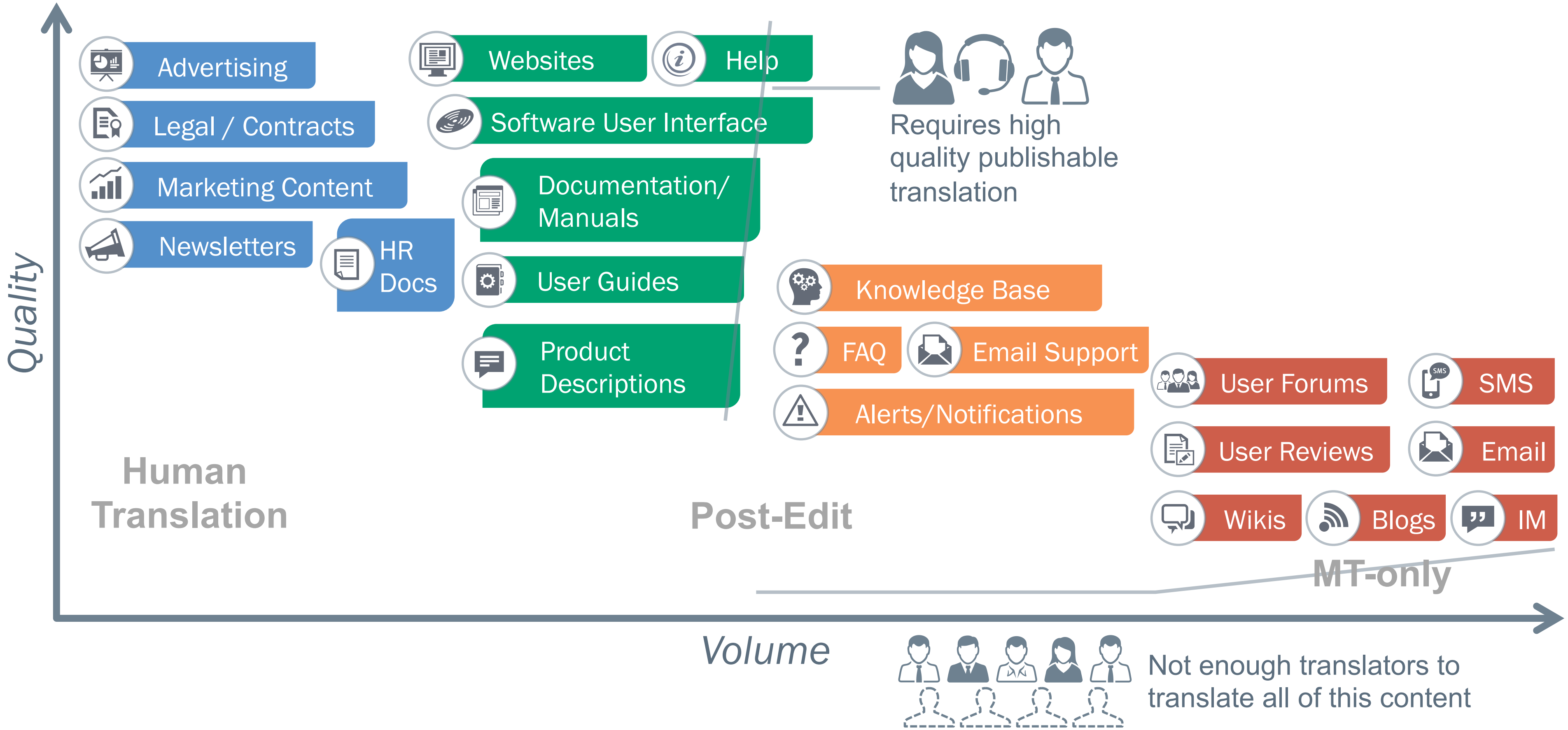
Around 70% of the largest global companies work with SDL

>10 BILLION words translated every month

Enabling companies to communicate with customers in 100+ countries

Driving $14B in online revenue annually with our ecommerce technology

So much to translate

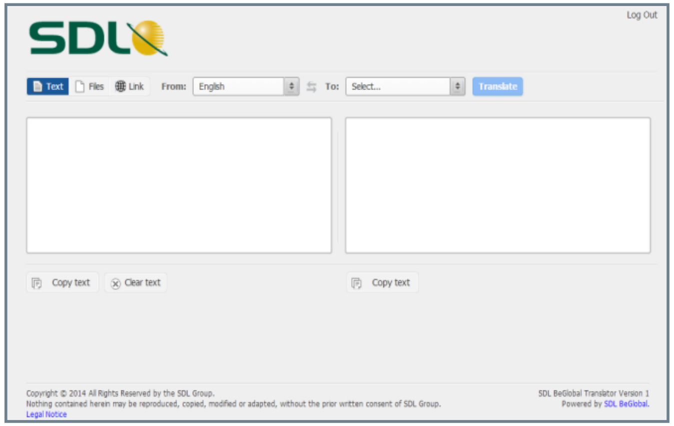


Machine Translation – customer use cases

Corporate Translation Portal

Description

- Immediate translation of text, files or web pages through a browser-based translation portal interface



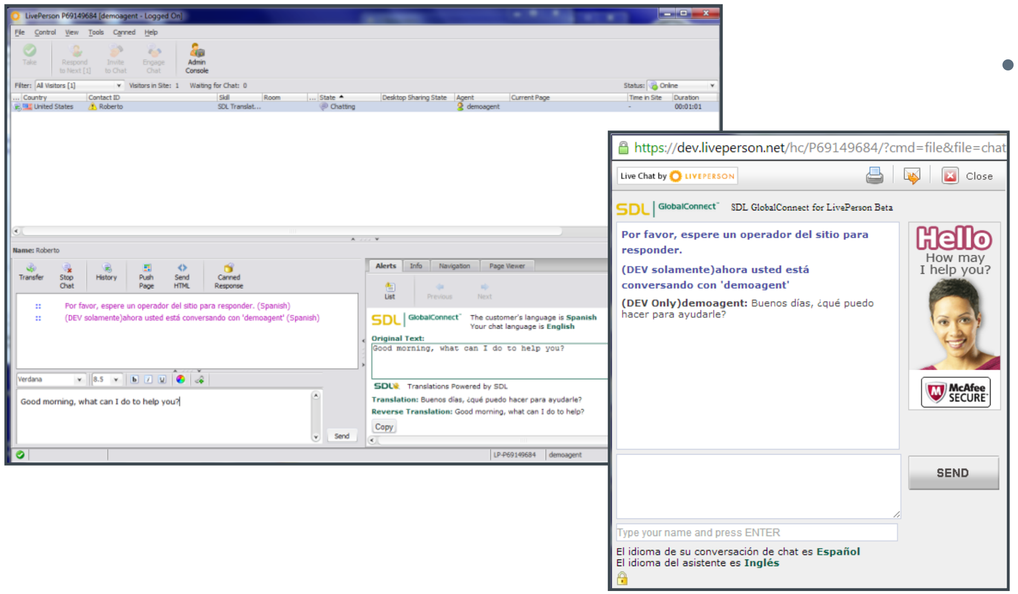
Benefits

- Virtually immediate response to translation requests
- Term and Brand support
- No concern about the mining/harvesting of sensitive information by the translation provider
- Can be branded with customer logo

Live Chat Translation

Description

- Real-time translation of web-based chat conversations



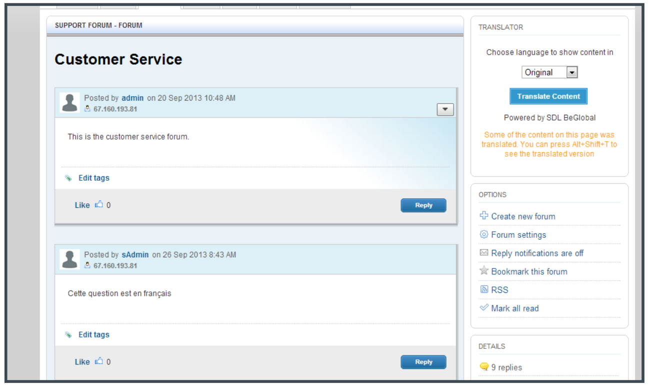
Benefits

- Reduces cost of staffing the support/ sales operations as they do not need multi-lingual agents
- Customer acquisition rates are much higher if you engage the customer in chat.

Community Forum Translation

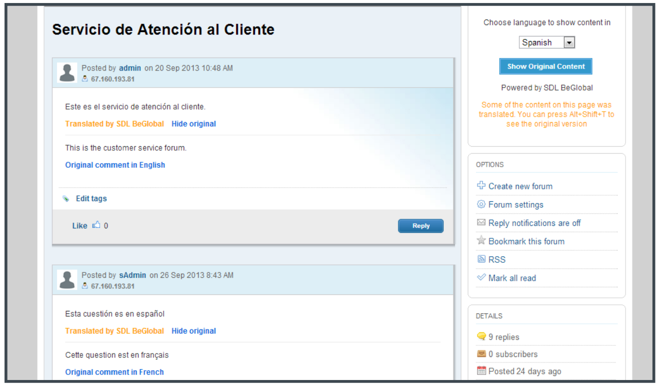
Description

- Translation of user-generated content in web-based community forums



Benefits

- Enable interactions between customers who speak different languages
- Leverage community expertise across languages instead of only within the language of community experts

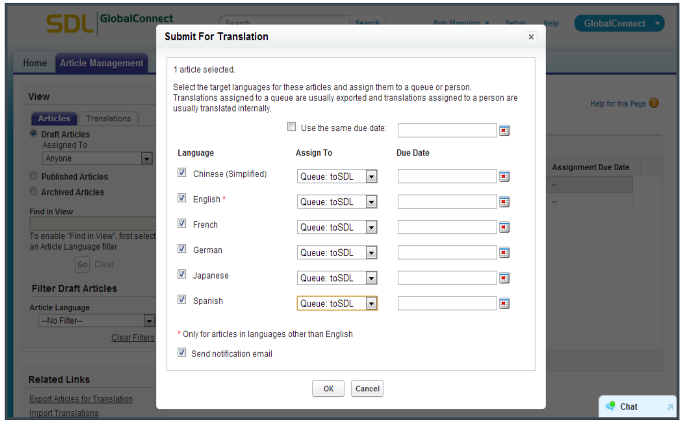


SDL

Knowledgebase Content Translation

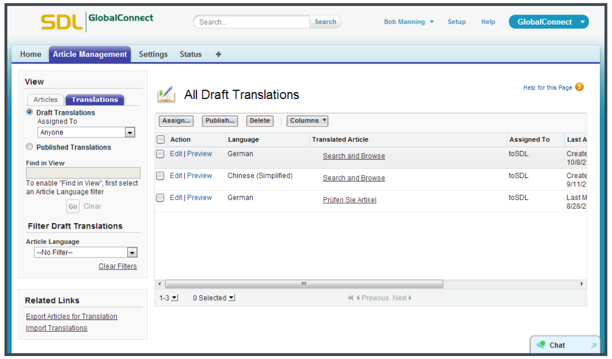
Description

- Translation of knowledge base content for local language customers of technical solutions



Benefits

- Reduces customer support costs and activity level by allowing remote language customers to directly access solutions
- Increases customer satisfaction by providing solutions in their native language



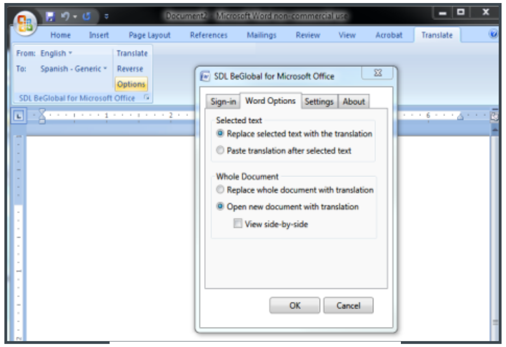
Microsoft Office (Word, Outlook, PowerPoint, Excel)

Description

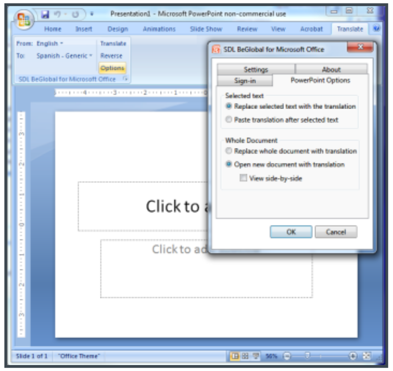
- Easy end-user translation of common Microsoft Office documents (Word, PowerPoint, Excel) and email messages (Outlook)

Benefits

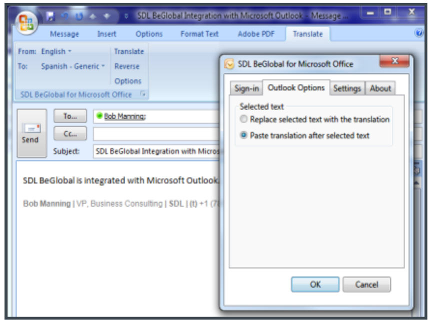
- Support for both baseline translation engines and custom-trained translation engines
- Term and Brand support



Microsoft Word



Microsoft PowerPoint

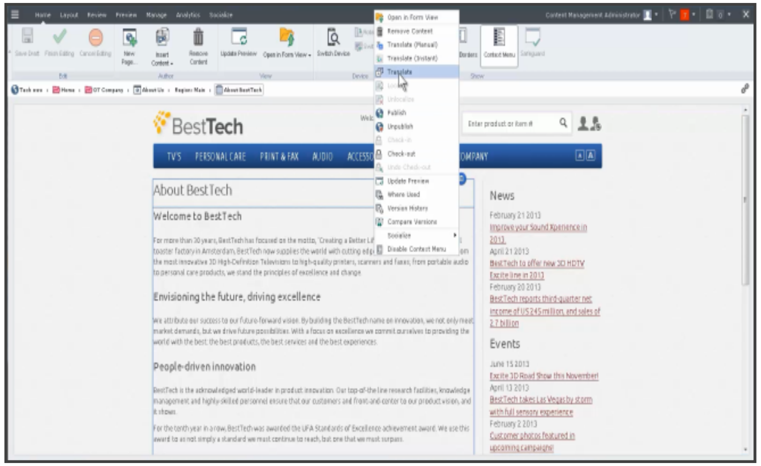


Microsoft Outlook

Web Content Translation

Description

- Integrate with web content management system to translate web site
- Embedding MT into the web site to support translation “on demand”



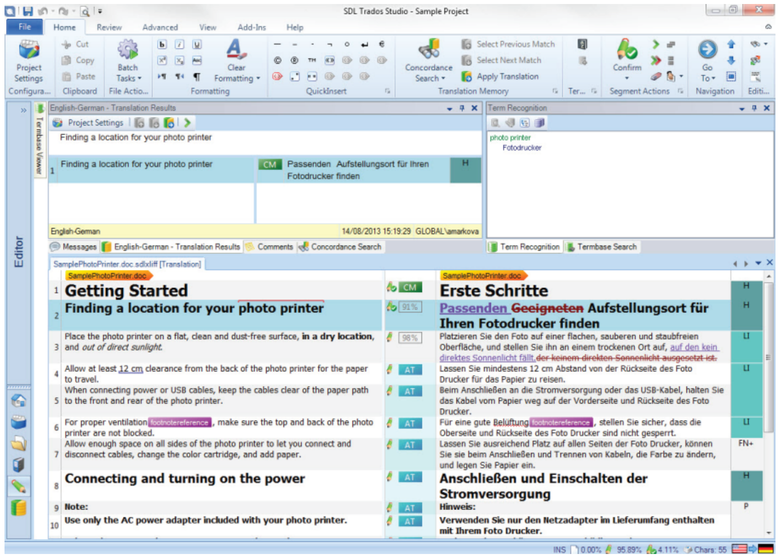
Benefits

- Ability to translate large volumes of web content that would not otherwise be translated because of cost
- Real-time translation can facilitate support for multi-lingual content with minimal changes to the development and storage of the source content

Translator Productivity

Description

- Direct access to machine translation from SDL Trados Studio



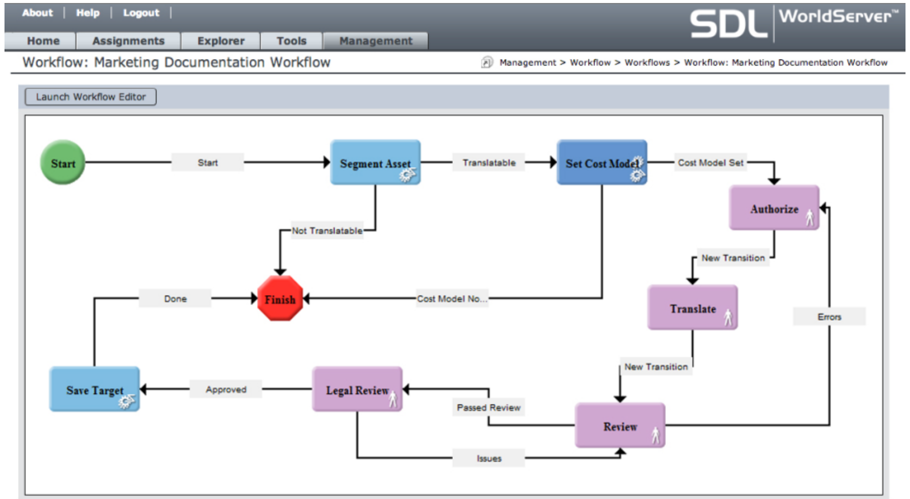
Benefits

- Improve the efficiency of translators by providing results of machine translation to them for segments that do not match entries in translation memory
- Adapt and personalize machine translation to each translator in real time

Enterprise Translation Productivity

Description

- Integration of machine translation to translation workflows in SDL WorldServer / SDL TMS



Benefits

- If a preexisting match is not found, the segment can be submitted to MT
- Improve the efficiency of the translators by providing results of machine translation to the translator for segments that do not match entries in translation memory

Text Analytics

Description

- Gain global perspectives by using MT to convert all multilingual text to English before analysis

Benefits

- Enables analysis of non-English content with little or no new development work
- SDL APIs allow the translation step to be easily integrated into the overall analysis process
- Term and Brand support can be valuable to improve the consistency of the translation results

Machine Translation Options

What is a Baseline System?

- A generic machine translation engine developed for use across a broad range of subject matter. SDL Language Technologies currently has nearly 100 "baseline systems" available for use today

★ GOOD

For Example:
- From eng to sp OR sp to eng

What is Term & Brand Management?

- A mechanism for enforcing the proper translation of brand and product terms specific to your company and its products

★★★ BETTER

For Example:
- "Apple", the computer company, is NOT "Pomme" when translated to French
- In your company, when translating from Spanish to English, "libro" always translates to "book" NEVER "tome" or "document"

What is a Trained System?

- An engine adapted to translate specific subject matter where accurate terminology is important. The adaptation process uses previous translations or other relevant domain

★★★★★ BEST

For Example:
- For particular customer terminology
- For a particular use case (chat, kb, email, etc.)
- For a particular product

SDL

Commercially-driven MT Research

SDL Research

- Formerly Language Weaver
  - founded by Daniel Marcu and Kevin Knight (USC ISI)
  - over a decade of leading expertise in SMT
  - major contributions (papers/patents) in phrase-based, string-to-tree and hierarchical MT, adaptive machine translation, quality estimation, tuning, evaluation...
- Research labs in Los Angeles (USA) and Cambridge (UK)
- Team members have published >100 on SMT and related technology
  - Current team includes: Bill Byrne, Samad Echihabi, Gonzalo Iglesias, Dragos Munteanu, Steve DeNeefe, Jonathan Graehl, Rory Waite, Wes Feely, Yuanzhe Dong…
  - Long experience in implementing, improving and deploying MT engines, big data processing, natural language processing, machine learning
  - Most recent papers: user feedback adaptation, optimization, neural networks pre-ordering, speed-constrained tuning

SDL Research

- Strong links with academia
  - University of Cambridge, USC ISI
- Summer internship program
  - 2012: morphologically-rich languages (Jan Botha, Rory Waite)
  - 2013: feature-rich pre-ordering (Laura Jehl), online adaptation (Felix Hieber)
  - 2014: decoding with target-side dependency LMs (Patrick Simianer)
  - 2015: bayesian optimization for speed-constrained tuning (Daniel Beck)
- Participation in research projects
  - DARPA GALE, DARPA BOLT, TSWG, etc… (translation quality, informal language…)
  - EC FP7: FAUST (adaptation to user feedback)
  - current research funded by products

Research focus

Bring MT research results to the products/services that customers use

- Factors for the adoption of MT

Quality of the models

Decoding speed in line with real MT user expectations

Approaches that work for many language pairs

Decoders must be able to run on premise and in the cloud

Respect terminology (source-to-target constraints)

Stability, simplicity, flexibility

Technology that adapts to user domain, style, and feedback (Adaptive MT)

Confidence metrics (TrustScore)

Controllable memory and disk footprint

Robustness to mis-spellings

Research tasks

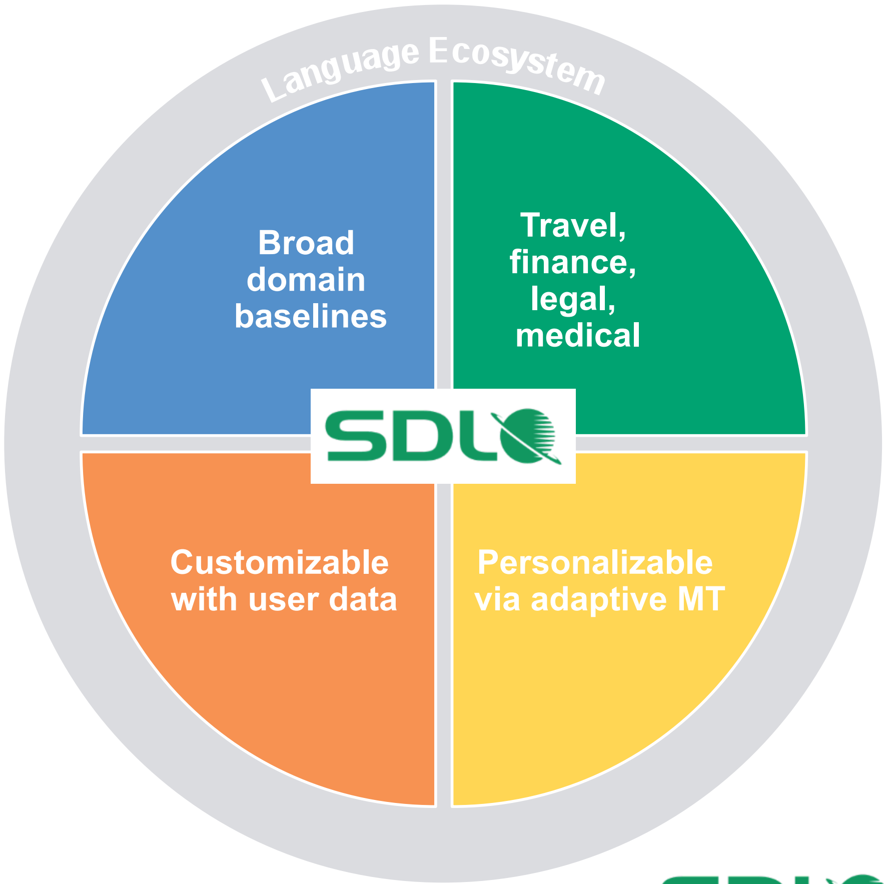
Many interesting research tasks emerge from the previous set of factors:

- incorporating the latest coolest technology (and confirming that it works)
- making it run at user-desired speed
- making it comply with varying customer constraints
- automating (and speeding-up) training to get optimal systems with minimal intervention
- running experiments over dozens of language pairs, domains, etc...
- experimenting with real-life post-editing data

Adaptive Machine Translation

Adaptive Technology is crucial

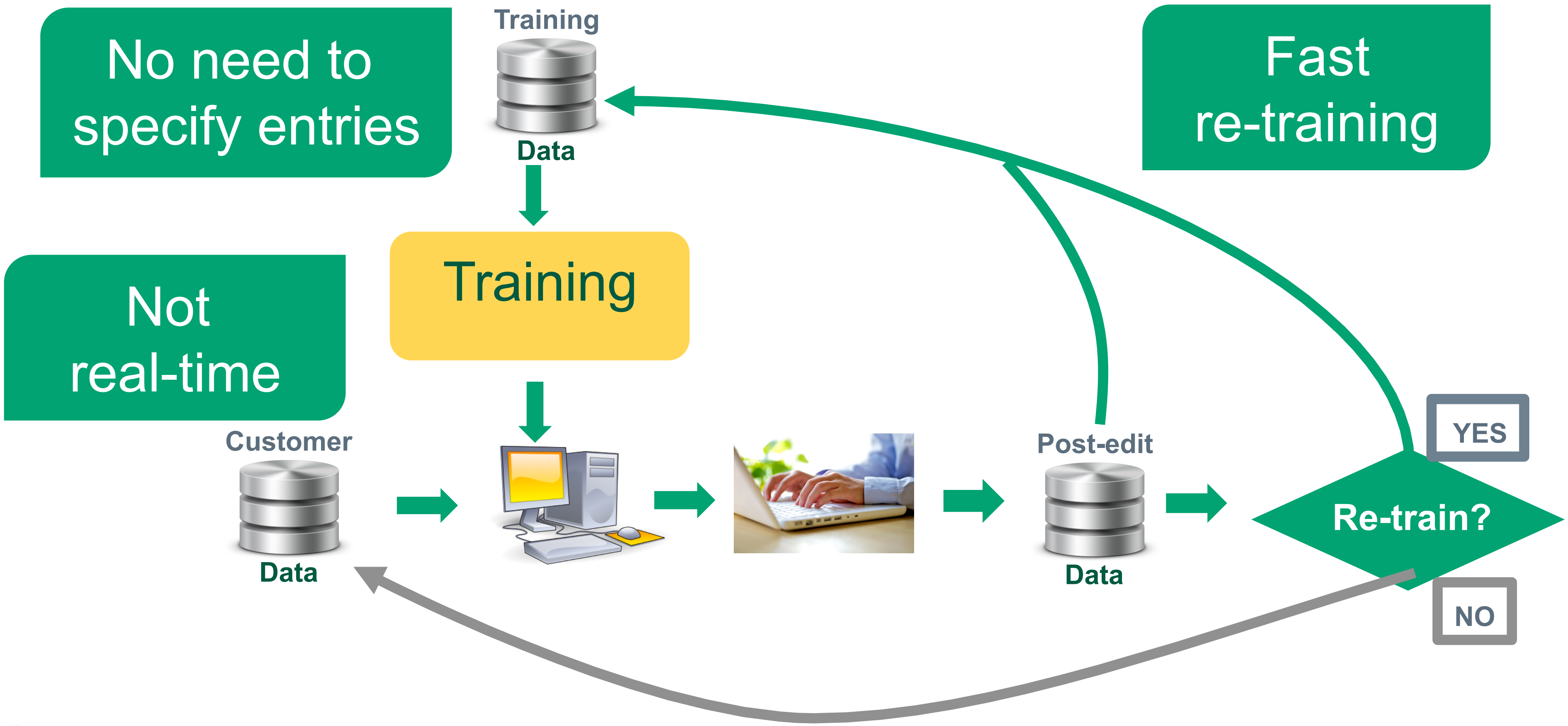
- Over 100K clients
- Trillions of in-house data
- Customer data
- Post-editing data
- Adaptive MT enables optimal use of available data sources to create customized and personalized MT engines for each user
  - Hundreds of trainings per month
  - Infrastructure for Big Data and MT at scale



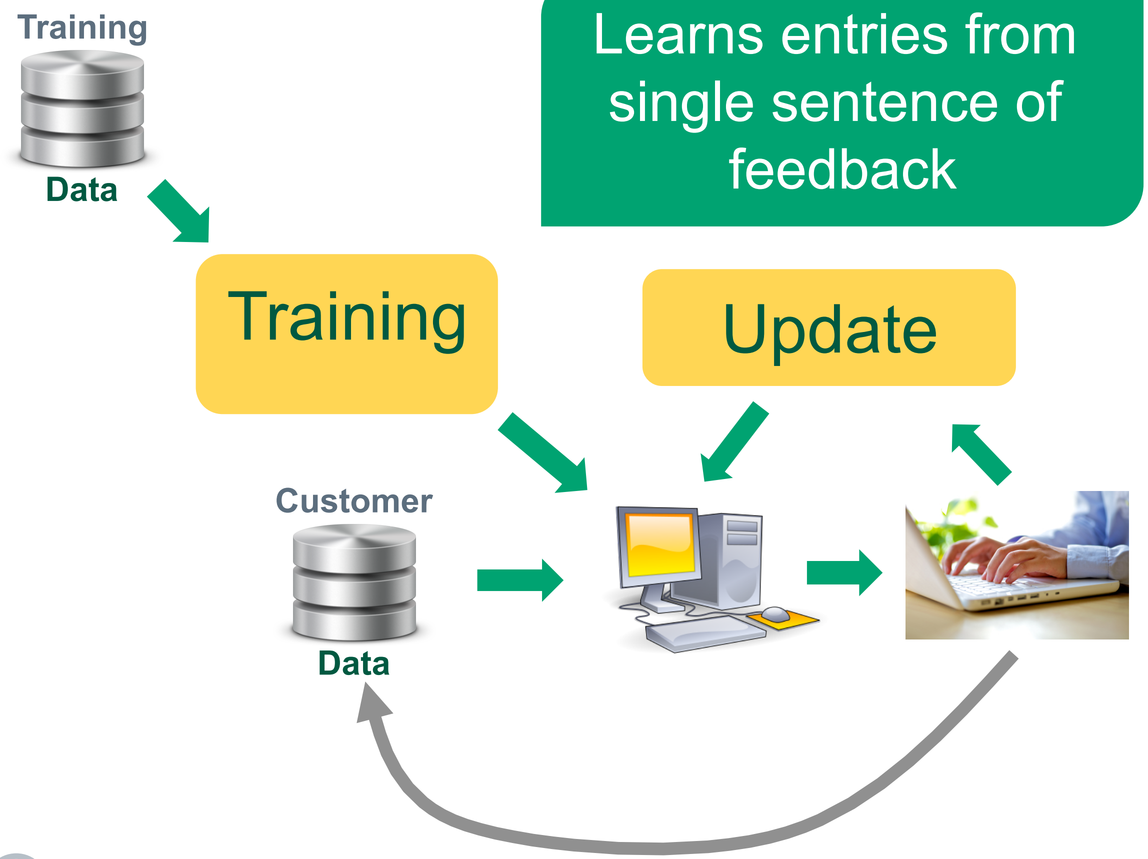
Robust and Automatic MT Adaptation

- Building customer specific engines
  - Merge of baseline and customer-specific models
  - Integration with translation memories
  - Customers benefit from super large gains in quality
- Enabling real time adaptation of MT engines
  - Integrating seamlessly user post-editing feedback
  - MT engines continuously learn over time
  - Improvement in both Translation quality and productivity
- Building domain-specific MT engines
  - Travel, Finance, Computer Software, etc…
- Meeting customer requirements
  - Model size (disk and memory footprint)
  - Offering expected speed, without sacrificing quality

Offline MT Adaptation



Online MT Adaptation



Learns entries from single sentence of feedback

Lightweight, personalized integration (like Dictionary)

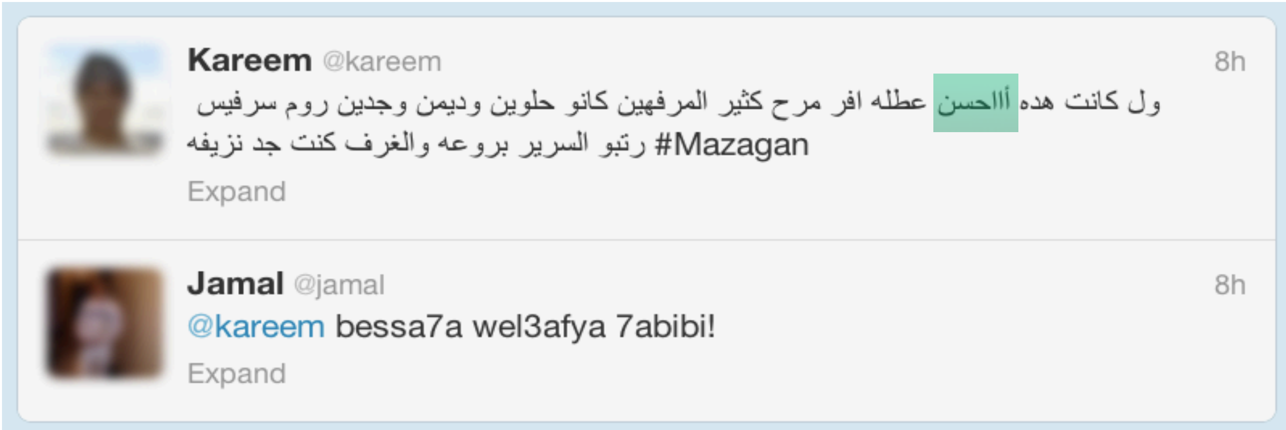
No manual intervention (like Domain Adaptation)

Translating Informal Language

- May, J. et al, AMTA 2014

Informal language

- Motivation: Informal language poses new challenges to machine translation

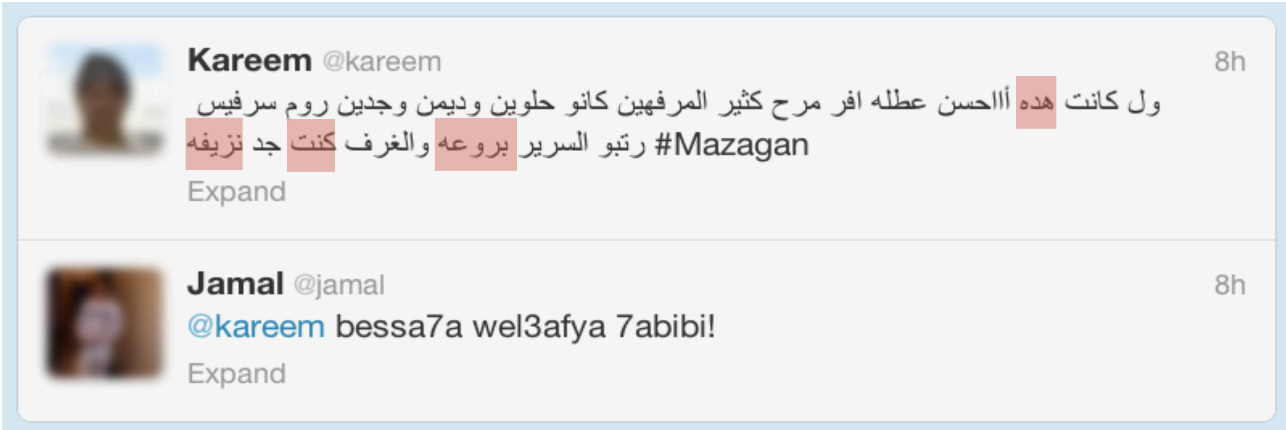


Character Repetition

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

Informal language

- Motivation: Informal language poses new challenges to machine translation



Spelling Errors

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

Informal language

- Motivation: Informal language poses new challenges to machine translation

Dialect

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

Informal language

- Motivation: Informal language poses new challenges to machine translation

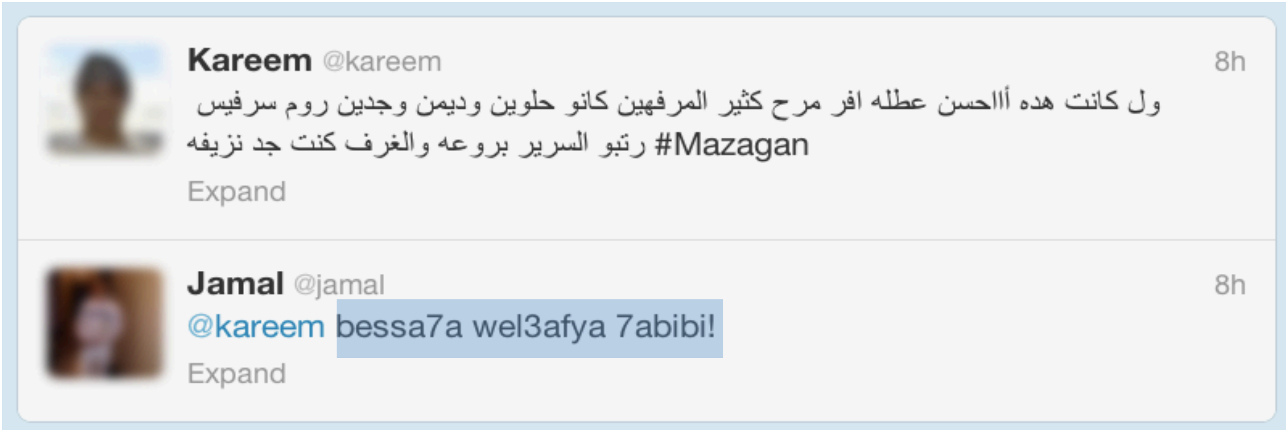
Morphology

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

SDL

Informal language

- Motivation: Informal language poses new challenges to machine translation

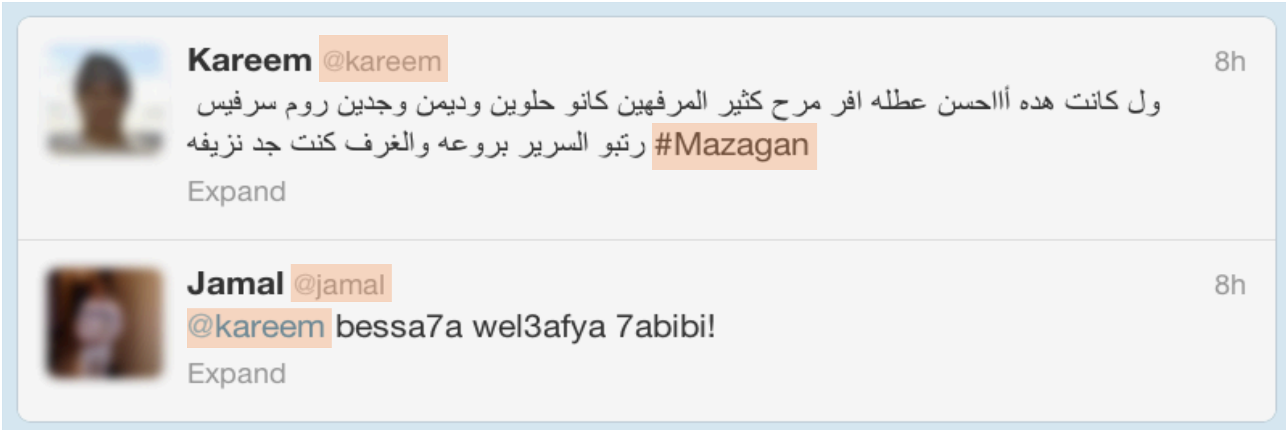


Romanization

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

Informal language

- Motivation: Informal language poses new challenges to machine translation



Metadata

- Goal: Improve SMT technology to better handle informal language using new techniques and algorithms

Social Media Example

@bothainakamel1 هو في حد اعترض على المسيحيين في دول الخليييييج؟
machrou3 achouf #البطل_أحمد_الجعبري موقع يهتم بالدعوة إلى الله #ساهم_في_نشر_الخير #رتويت_ليعم_الخير.
اهل السعوديه هم اهلنا من لحمنا ودمنا.. عسى Allaaaaaaah يرحمهم ويتقبلهم شهضاء.

Baseline system

@1 is an end in objected to the Christians in the states of alkhlyyyyyyj?
3 #albtl_ahmd_aljaibra site on the invitation to God #sahm_fy_nshr_alkhyr #rtwyt_lyaim_alkhyr.
The people of Saudi Arabia are our of hurts, and our blood..maybe mercy shhdha'a mechanical process.

Social Media Example

@bothainakamel1 هو في حد اعترض على المسيحيين في دول الخليييييج؟
machrou3 achouf #البطل_أحمد_الجعبري موقع يهتم بالدعوة إلى الله #ساهم_في_نشر_الخير
#رتويت_ليعم_الخير. اهل السعوديه هم اهلنا من لحمنا ودمنا.. عسى Allaaaaaaah يرحمهم ويتقبلهم شهضاء.

Metadata | Dialect | Normalization | Deromanization | Morphology | Spelling

Recent SDL system

@bothainakamel1 Is there anyone who objected to the Christians in the Gulf states? I see a project #of_the_hero_Ahmed_Jabari site dedicated to the call to Allah, #contribute_to_spreading_the_good #retweet_them_good. The people of Saudi Arabia are our people from when our blood. May God have mercy on them and accept them as martyrs.

Speed-constrained tuning

- Beck, D. et al, NAACL 2016

I want a good system that runs at...

Meeting customer requirements

- Model size (disk and memory footprint)
- Offering expected speed, without sacrificing quality

- Typical SMT system tuning
  - feature weight optimization (MERT, PRO, ...)
  - based on BLEU scores (or other similar metrics)
  - ignores decoding speed considerations

- How to make it fast?
  - modify training parameters → data filtering (rules, sentences...)
  - modify decoding parameters → increase pruning

Modify decoding parameters

- We are interested in:

Automatically finding the parameter values that maximise BLEU while keeping decoding speed above a certain threshold

- This is hard:
  - Expensive, cannot apply algorithms that rely on n-best/lattices (MERT, PRO...)
  - potentially big loss in translation quality
  - speed measurements can be noisy
  - avoid manual intervention?

Bayesian Optimization (BO)

- Goal: $$\theta_\star = \arg\max_{\theta \in \Theta} f(\theta)$$, where θ is the parameter vector

- BO approaches this by:
  - Defining a prior model over f and evaluating it sequentially
  - Choosing the evaluation points to maximise the utility of the measurement (*acquisition function*)
  - Trading off exploration of uncertain regions of Θ and exploitation of promising regions, based on known measurements
- Particularly useful when f is non-convex, non-differentiable and costly to evaluate

Bayesian Optimization (BO)

- Goal: $\theta_\star = \arg\max_{\theta \in \Theta} f(\theta)$, where θ is the parameter vector

BLEU score

decoding parameters

For phrase-based MT:

- Stack limit
- Distortion limit
- Number of rules per src

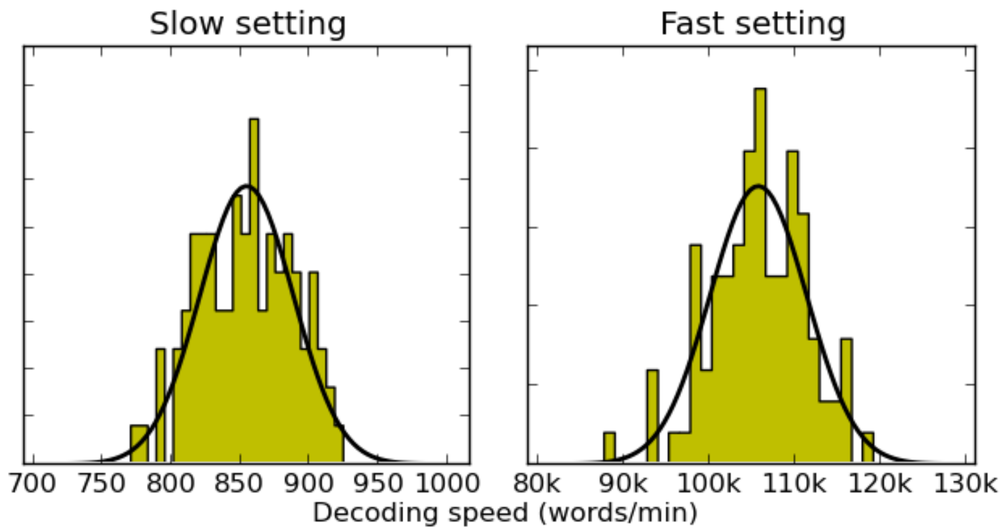
- BO approac
  - Defining a r f a
  - Choosing oint measurer
  - Trading o unc on of promising regions, based on k
- Particularly useful when f is n and

and decoding speed?

Constrained Bayesian Optimization (BO)

- Goal: $$\theta_\star = \operatorname*{arg\,max}_{\theta \in \Theta} f(\theta) \quad \text{s.t. } c(\theta) > t$$

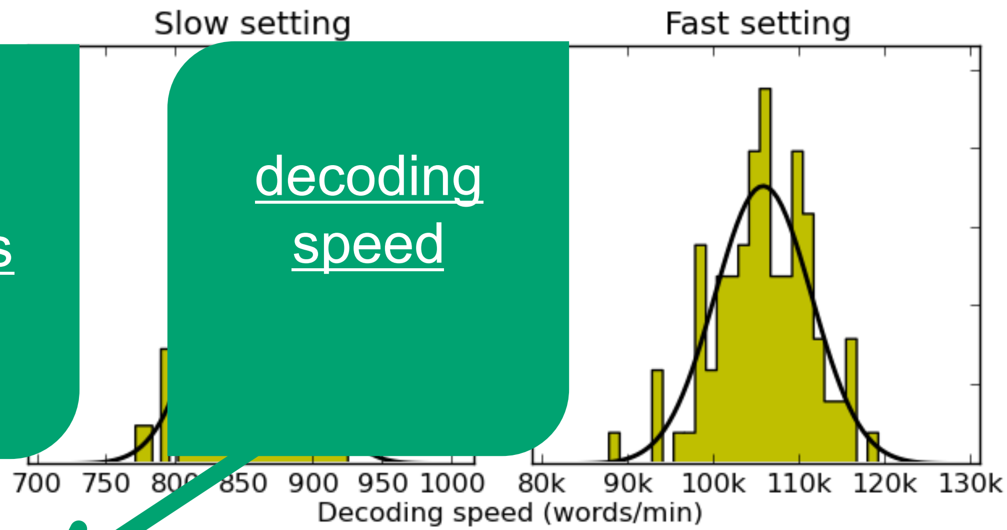
- $c(\theta)$ is the decoding speed achieved with params θ
- Noisy…
- … but we can take *c* as a probabilistic function too



$$\theta_\star = \operatorname*{arg\,max}_{\theta \in \Theta} f(\theta) \quad \text{s.t. } p(c(\theta) > t) \geq 1 - \delta$$

Constrained Bayesian Optimization (BO)

- Goal: $\theta_\star = \arg\max_{\theta \in \Theta} f(\theta) \quad \text{s.t. } c(\theta) > t$



BLEU score

decoding parameters

decoding speed

$$\theta_\star = \arg\max_{\theta \in \Theta} f(\theta) \quad \text{s.t. } p(c(\theta) > t) \geq 1 - \delta$$

Detailed Setup

- Priors: Gaussian Processes over $f(\theta)$ and $c(\theta)$ (Snoek et al. 2012)
- Acquisition function: Predictive Entropy Search (Hernández-Lobato, 2015)
  - Maximises the information around the global optimum θ^*
  - Empirically shown to give better results in constrained BO
  - Allows constraint decoupling
- SpearMint implementation (many others are available)

- Many other choices are possible
- We did not experiment with these here

Experimental Setup

Parameter ranges:

- Distortion limit: [0, 10]
- Stack size: [1, 500] (log scale)
- Number of translations: [1, 100] (log scale)

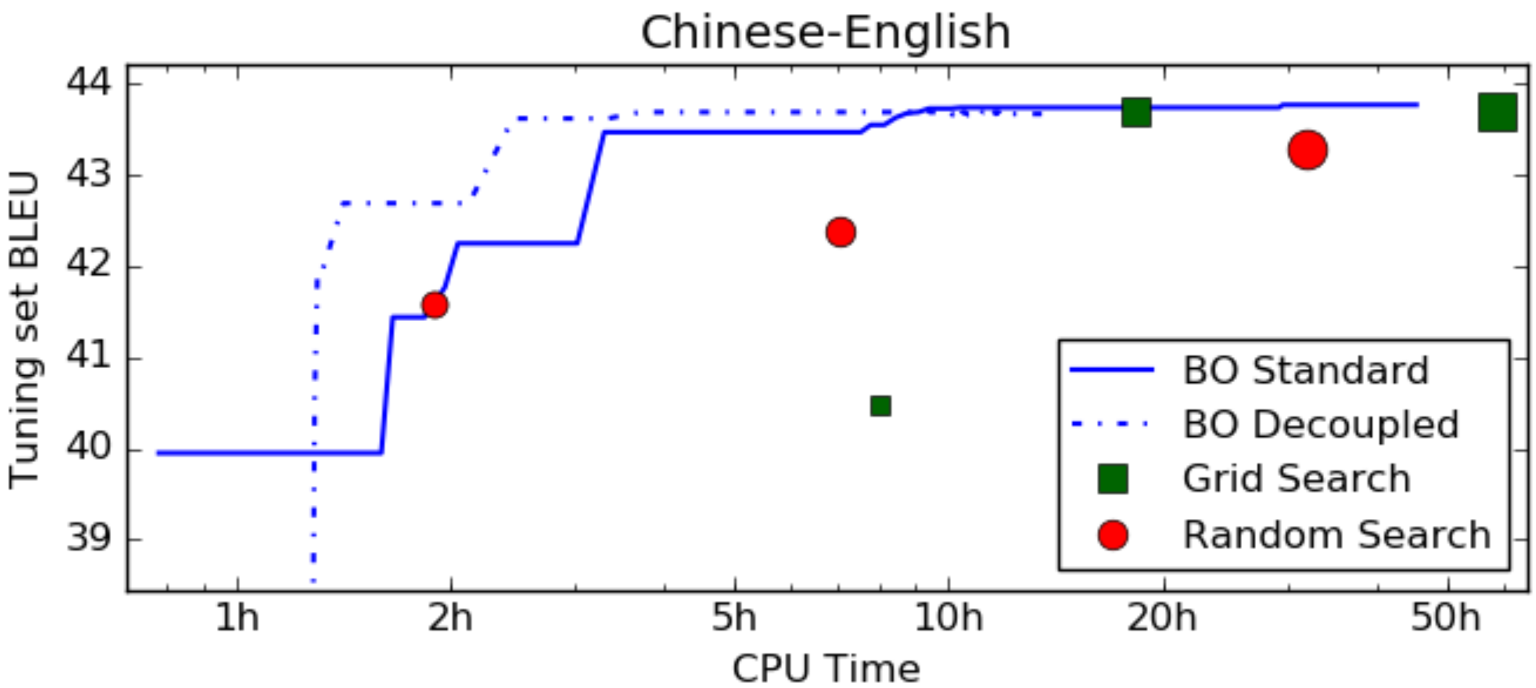
BO settings:

- Standard: $\delta = 0.01$, 125 max iterations.
- Decoupled: $\delta = 0.05$, 250 max iterations, speed measurements taken from a smaller subset of 50 sentences.

Baselines:

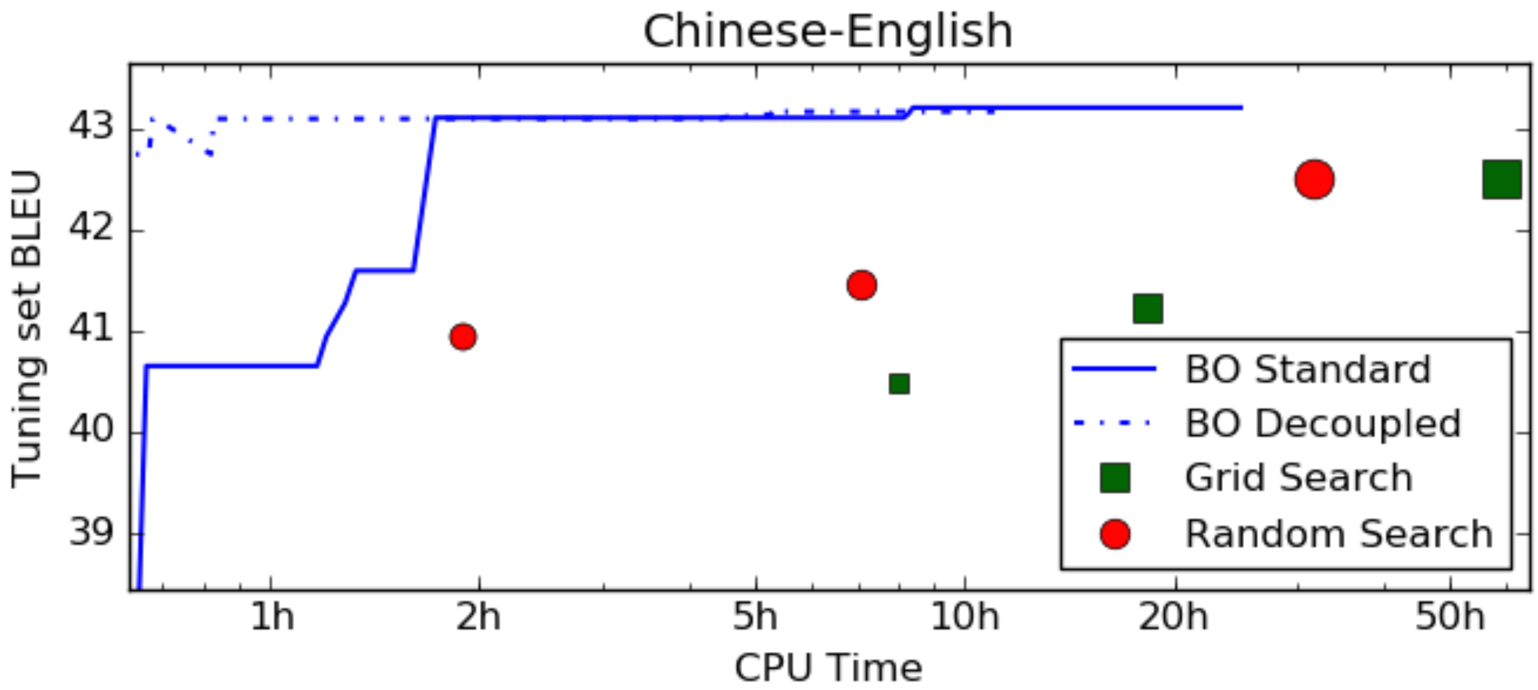
- Grid Search
- Random Search [Bergstra and Bengio, 2012]
- 125 max iterations for both.

Constraint: 2000 words-per-minute



Grid and Random = 8, 27 and 125 decodings

Constraint: 5000 words-per-minute



Grid and Random = 8, 27 and 125 decodings

Feature Weight Re-Optimization

- Chi-Eng at 2000 wpm:

	Tuning		Test		θ		
	BLEU	speed	BLEU	speed	d	s	n
MERT	44.3	–	42.5	51	10	1K	500
BO-S	43.8	2.0K	41.9	1.9K	10	25	100
MERT-flat	43.8	2.0K	41.4	1.9K	10	25	100
MERT-opt	**44.3**	**2.0K**	**42.4**	**1.9K**	10	25	100

Comments

- BO proves faster than grid and random at optimizing parameters for BLEU and speed

But…

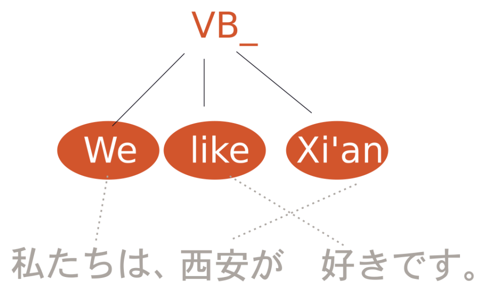
- Number of parameters must be small
  - Cost of acquisition function calculation grows exponentially
- Other priors may be more suitable
- Can we apply it to even more expensive functions?
  - Training parameters
  - Neural network parameters

Pre-ordering for long-distant language pairs

- Jehl, L. et al, EACL 2014
- De Gispert, A. et al, NAACL 2015

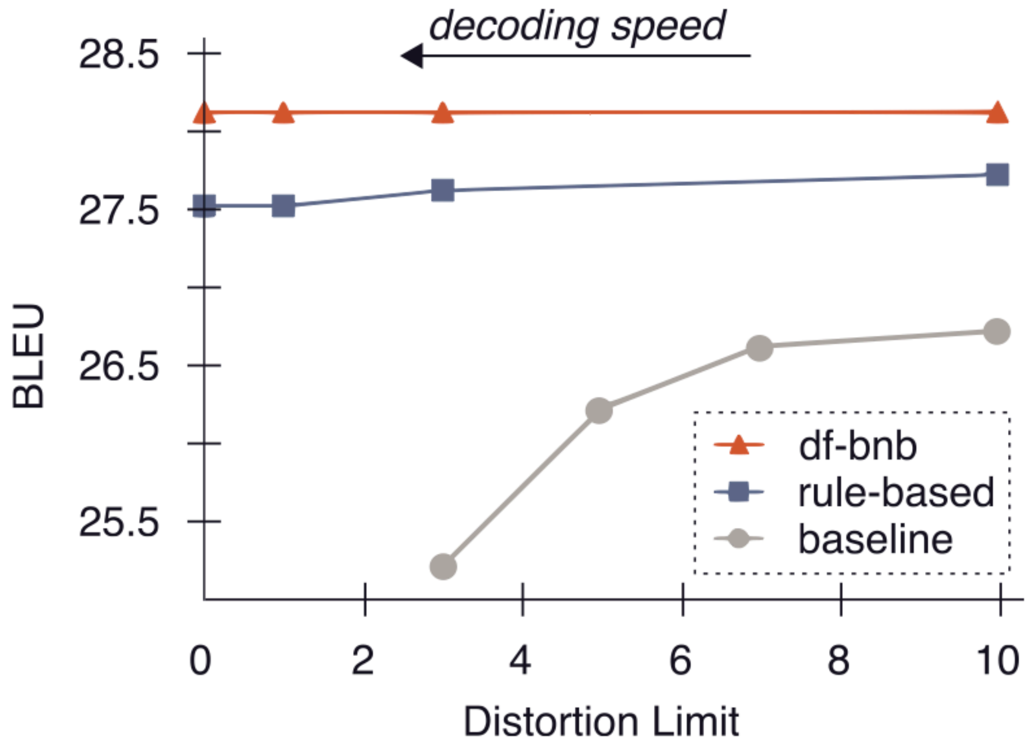
Preordering using Logistic Regression

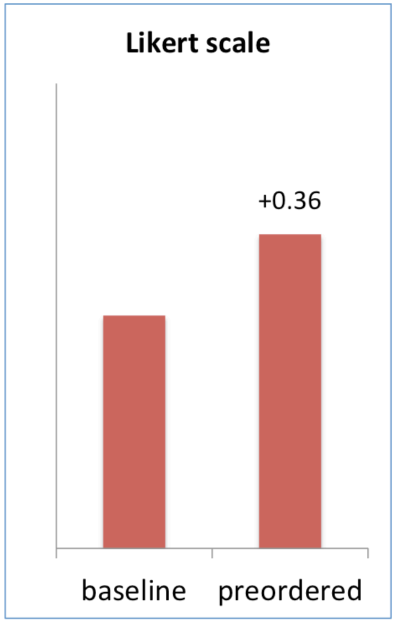
- Goal: order English sentences in Japanese/Korean order before translating
  - Done for test but also train data
  - Good for distant language pairs
- Train a Logistic Regression model to predict the probability to swap two sibling nodes
  - Use dependency parse and lexical info
  - Best when modeled via feed-forward neural nets
- Use these probabilities to find a global ordering
  - Efficient depth-first branch-and-bound search




```
(We,like): 0.4
(like,Xi'an):0.8
(We,Xi'an):0.3
```

Faster and better





Better results AND faster decoding!

Thank you for your attention !

Global Customer Experience Management