

# Practical Experience with Statistical Machine Translation

Daniel Hardt  
language**e**ns

Nordisk arbeidsseminar i oversettingsteknologi  
24 October 2008  
Språkrådet, Oslo, Norge

# Statistical Machine Translation System

In use at Lingtech

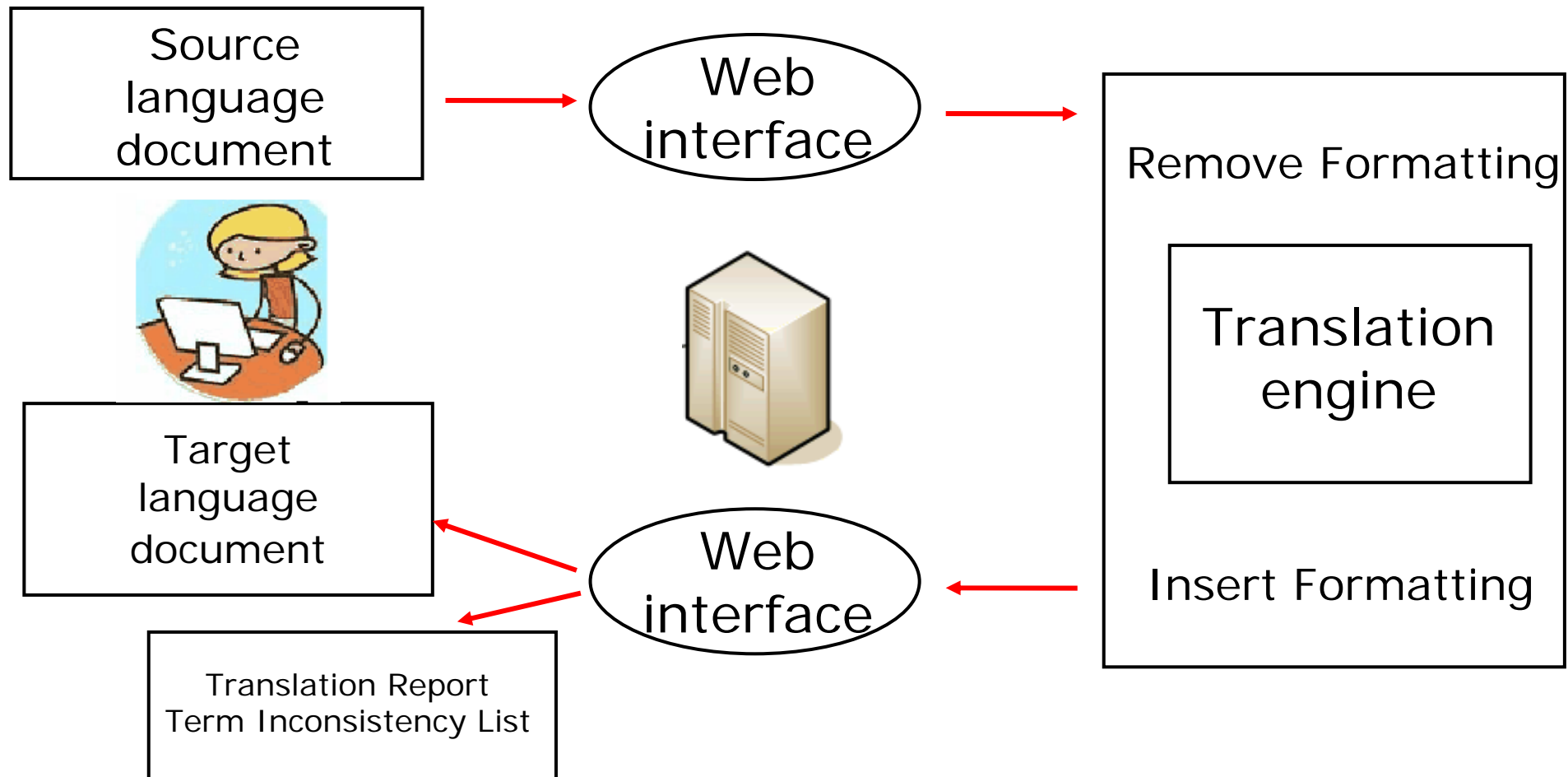
- Over 5 million words translated since Sept 2007

Reported Savings of 40-80% of translation work

Customized, Domain-Specific System

- System can be automatically set up for different domains and language pairs in a matter of days

## Using the System



# Overview

## The Past

why MT “should be” impossible

## The Present

a look under the hood: how it works

## The Future

The Coming Breakthrough

# **The Past**

## **Why MT “should be” Impossible**

Translation requires understanding of

*Grammar*

*Logic/Meaning*

*World Knowledge*

*Stylistics*

*Terminology*

**None of these have been successfully automated!**

# The Present

## How it Works

Search for *most likely translation* based on experiences of previous translations

- ***Translation Table***
- ***Context Model***

# **The Present**

## *Why it Works*

### Computing power

- Can now process millions or even billions of words of translation data

### Improved techniques

- Machine learning
- Heuristic search

# Building System from Data

***Translation Table*** built from large collection of domain-specific translation data

- Includes multiword entries
- Dozens or hundreds of choices per entry
- Each choice with probability

## Build Translation Table

### Translation Table

<b>in</b>	<b>i</b>	<b>63.00%</b>
in	ind	0.03%
in	ind for	0.001%
in	...	...
in conclusion	afslutningsvis	55.00%
...	...	...

Translation data



Sample Translation Table Entry: **we have***Total of 1581 entries*

Source	Target	Probability
we have	vi har	0.515146
we have	har vi	0.147763
we have	, vi har	0.0328003
we have	vi er	0.0328003
we have	er er	0.0150469
we have	må vi	0.0134826
we have	det er	0.013284
we have	vi	0.0102548

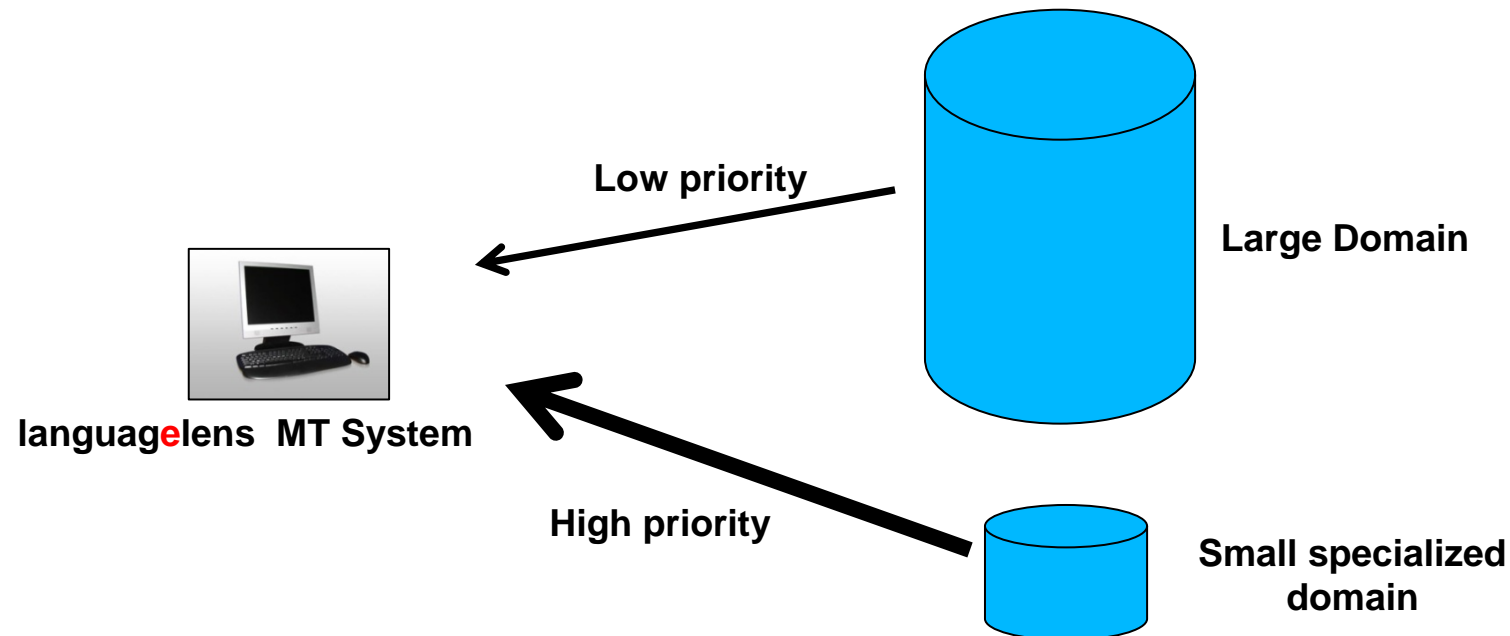
## Translation example

***Context Model*** guides choices among translation options

In	conclusion	,	we	have	several	points	.
I	konklusion	,	vi	har	<b>flere</b> →	<b>punkter</b> →	.
<b>Afslutningsvis</b>			vi har		adskillige	bemærkninger	.
På	afslutning		<b>har vi</b>		flere af	point	.

# Future Developments

## Customizing with Prioritized Domains



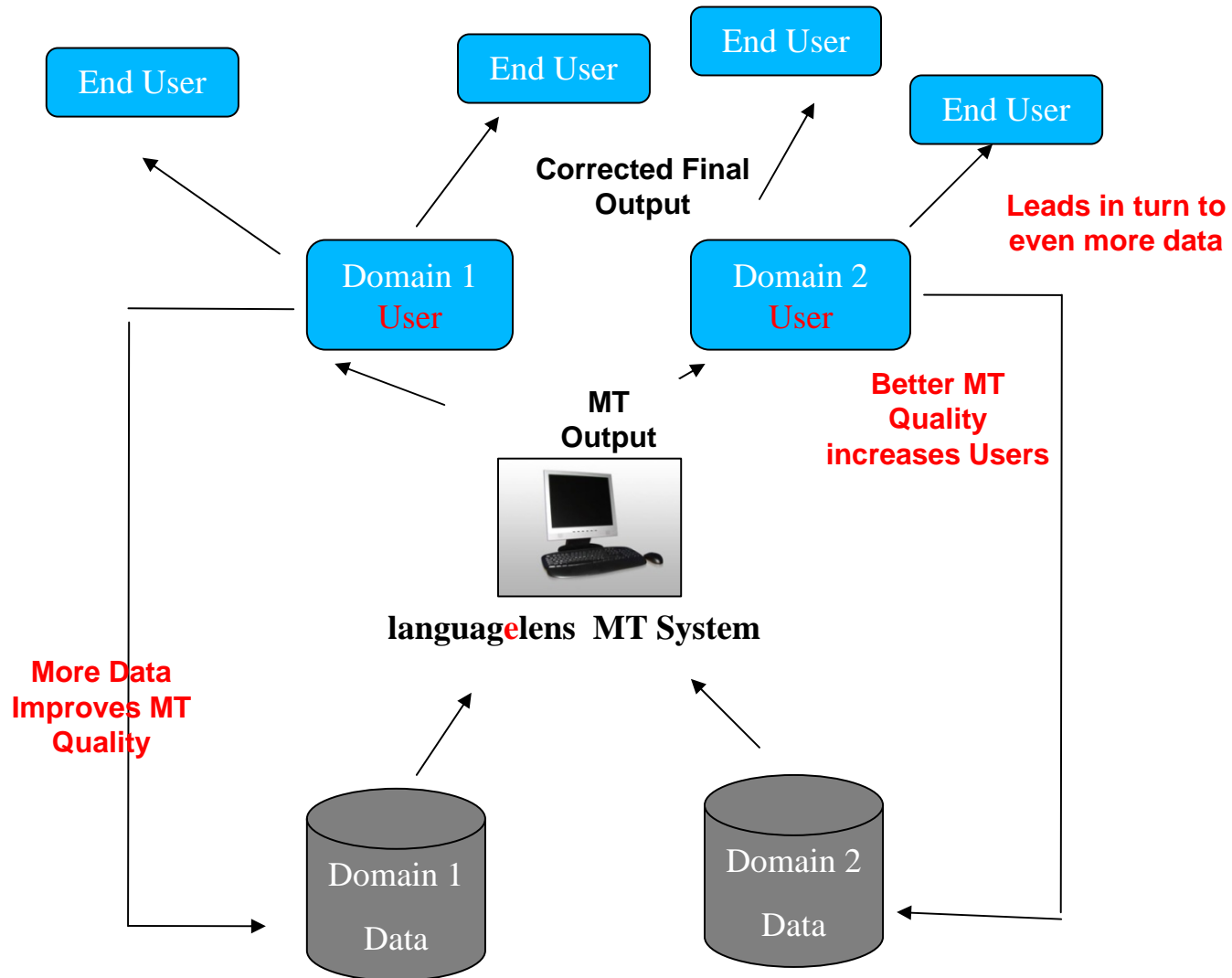
# The Future

## The Coming Breakthrough

- Lingtech's results for patent translation can be duplicated in other domains for which data is available
- The data is out there!
- As the system is used, output is post-edited, making more translation data available
- The system automatically improves with use.

# language**e**lens

automated translation for professionals



# language<sup>e</sup>lens

---

automated translation for professionals