Multilingual Systems for Career Development and Organizational Intelligence

Sisay Adugna Chala

Institute of Knowledge Based Systems and Knowledge Management
Department of Electrical Engineering and Computer Science
University of Siegen

July 24, 2017

UNIVERSITÄT SIEGEN

KBS KM
Knowledge Based Systems & Knowledge Management

Outline

Outline

What are Multi-lingual Systems?

Definition

Multilingual System is "any dynamic system that can support a variety of languages or locales" [1]

[1] Neustein, Amy, and Judith A. Markowitz. Where Humans Meet Machines. Springer, 2013.

What are Multi-lingual Systems?

Definition

Multilingual System is "any dynamic system that can support a variety of languages or locales" [1]

Multilingual Systems

- are computer programs which allow user interaction with the computer in one or more languages
- enable language selection dynamically, either at the time of invocation of the program or subsequently during its execution

[1] Neustein, Amy, and Judith A. Markowitz. Where Humans Meet Machines. Springer, 2013.

Why Do We Need ML Systems?

We need ML systems because

"Multilingual information access is critical for the acquisition, dissemination, exchange, and understanding of knowledge in the global information society." [2]

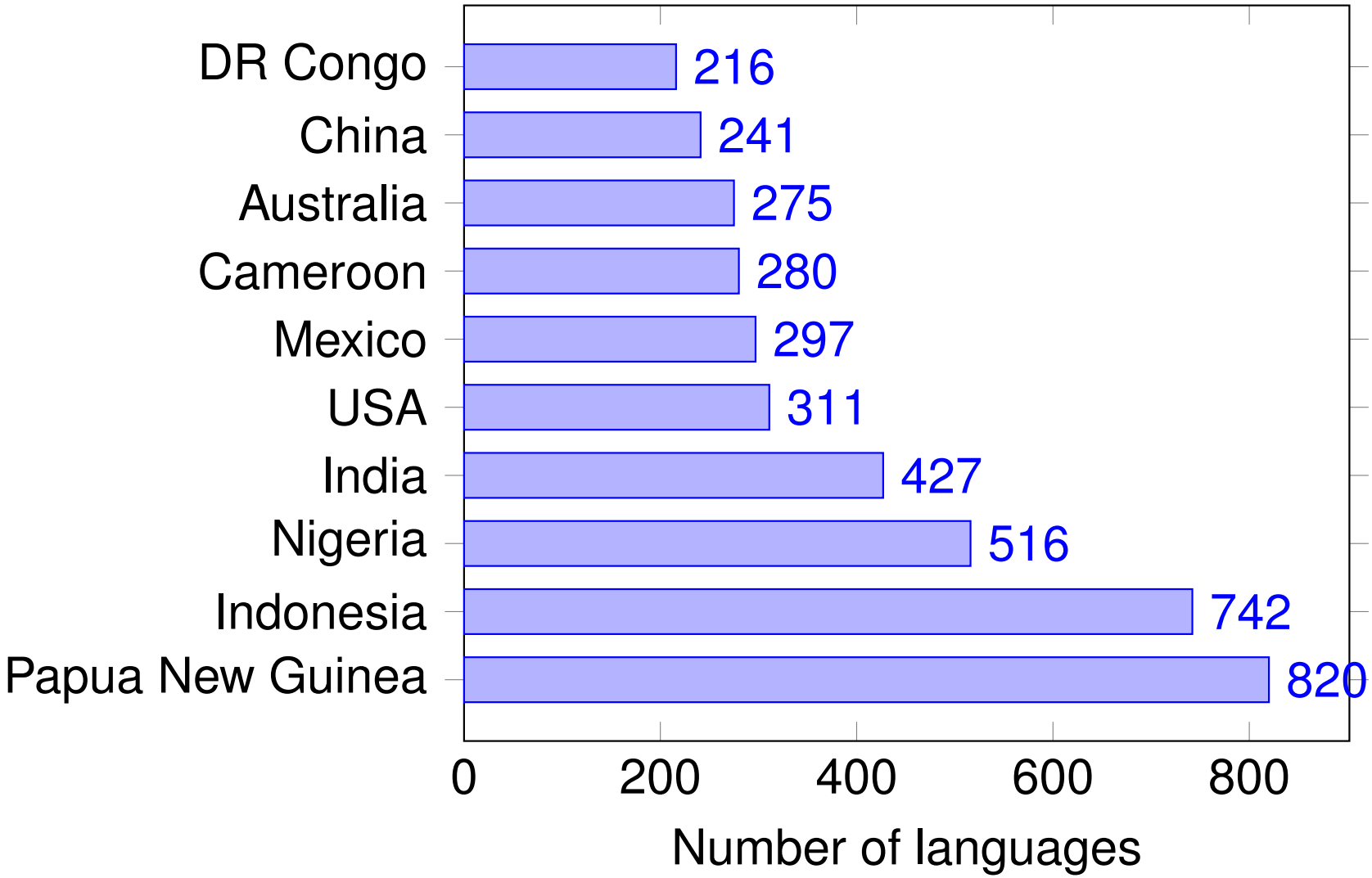
[2] https://www.microsoft.com/en-us/research/group/multilingual-systems/

Why Do We Need ML Systems? ...

Multilingual information access is increasingly becoming critical need due to:

- the accelerated growth in the size, content and reach of Internet
- the diversity of user demographics
- the skewness in the availability of information across languages

Why Do We Need ML Systems? ...



(Source: Ethnologue: Languages of the World, 15th ed.)

Why Do We Need to Study ML Systems Design?

- Offering content in several languages can add many new layers of complexity to web design.
- Translating, structuring and maintaining consistency of a multilingual website is not an easy task.
- Setting up and running a multilingual system is an expensive endeavor
- Though multilingual access to information is crucial in everyday activities, it is grossly overlooked by academic community with a few exceptions like AI lab of Uni Arisona [3]

[3] https://ai.arizona.edu/research

Issues to Consider Multilingual System Design

When building multilingual systems, the following questions need to be considered:

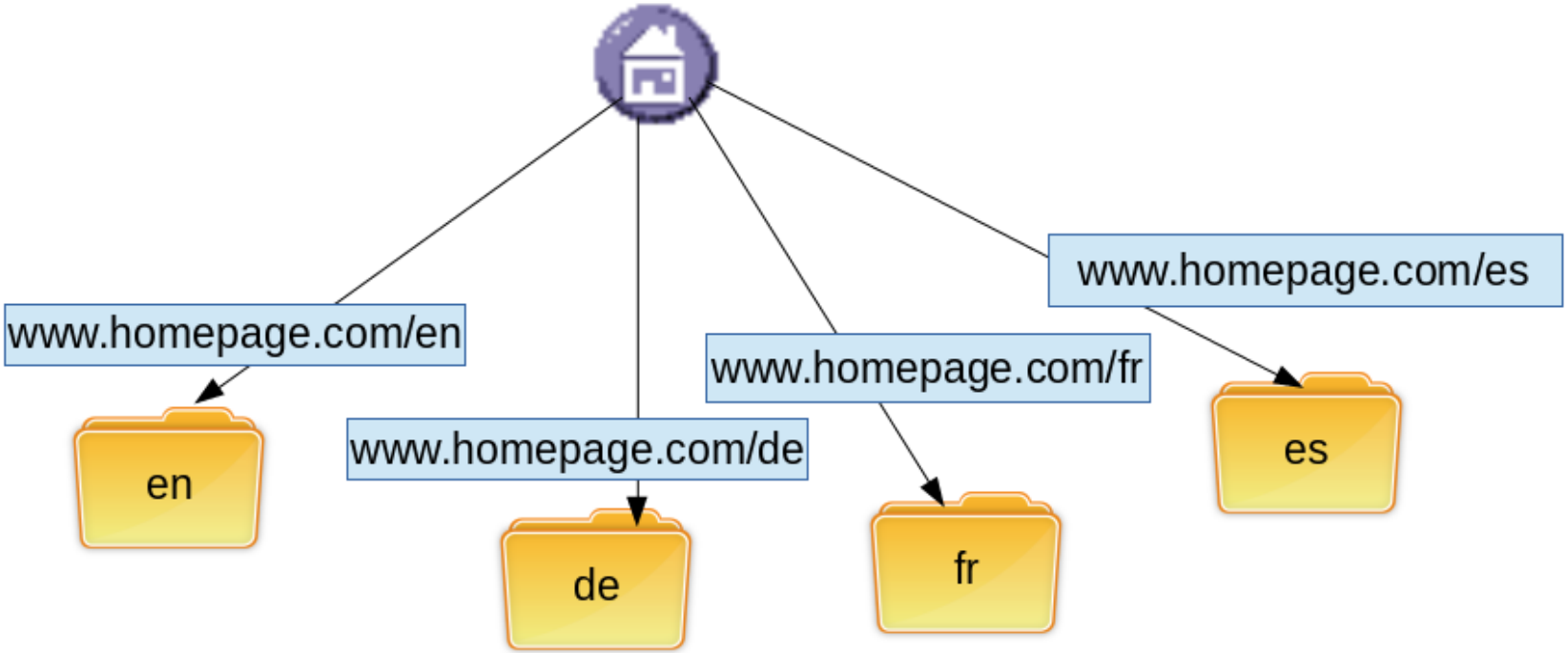
- Do we need the multilingual system for display only?
- Do we need the multilingual system to also include interaction messages? e.g, information, confirmation, alert
- Do we need data to be collected in multiple languages?

How do we Design Multilingual Systems?

1. Multiple versions fo translated static pages
2. Localization/Internationalization
3. Multilingual support to both database and interface templates

1. Multiple Versions of Translated Static Pages

Creating **separate folders** for each language and **linking** to the respective folder up on **language switching**



1. Multiple Versions of Translated Static Pages ...

Pros

- Easy to set up
- Files related to every language is updated independently
- Effective for static content

1. Multiple Versions of Translated Static Pages ...

Pros

- Easy to set up
- Files related to every language is updated independently
- Effective for static content

Cons

- Inconsistency of versions across languages

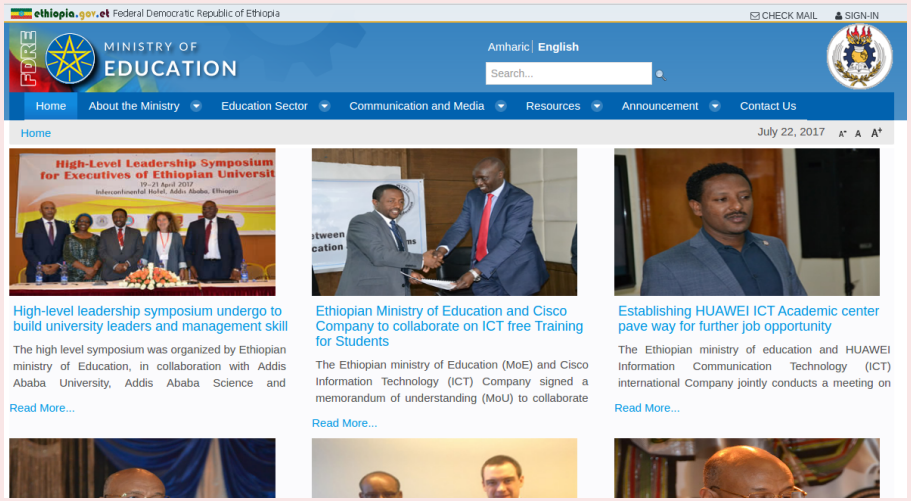
1. Multiple Versions of Translated Static Pages ...

Pros

- Easy to set up
- Files related to every language is updated independently
- Effective for static content

Cons

- Inconsistency of versions across languages



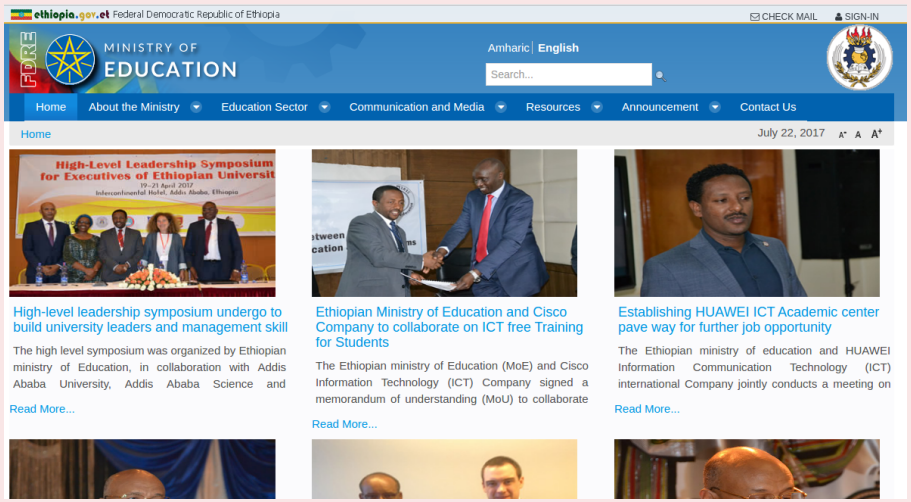
1. Multiple Versions of Translated Static Pages ...

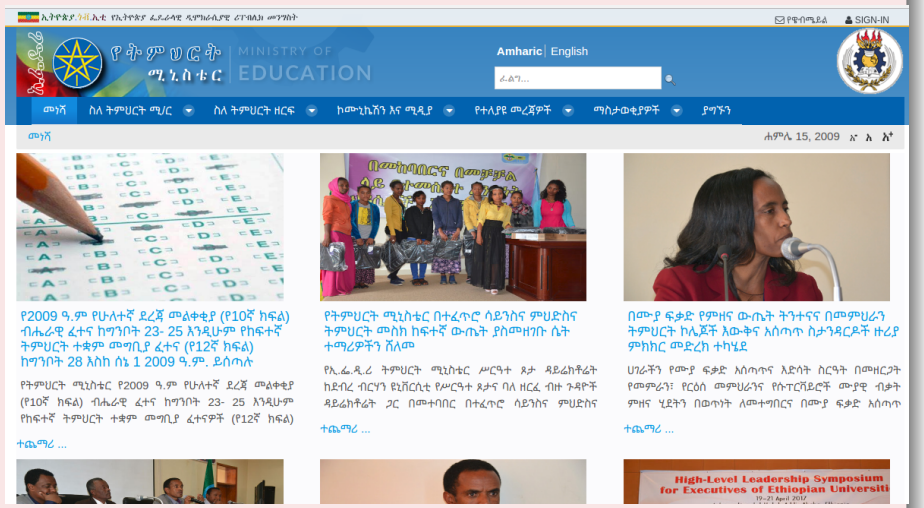
Pros

- Easy to set up
- Files related to every language is updated independently
- Effective for static content

Cons

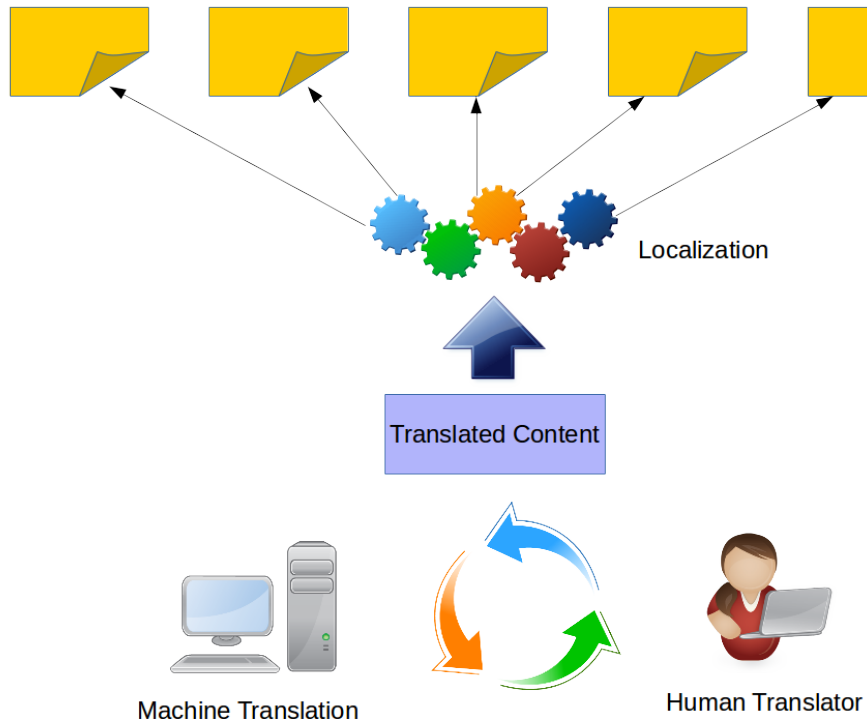
- Inconsistency of versions across languages





2. Localization/ Internationalization

- Using templates to create interface elements and interaction messages using Localization (L10n) or internationalization (i18n)
- Localization is the adaptation of systems to meet the needs of a particular language, culture or desired population's "look-and-feel."
- Language translation is a large part of localization



2. Localization/ Internationalization ...

Pros

- Relatively easy to maintain
- Consistent versions across languages

2. Localization/ Internationalization ...

Pros

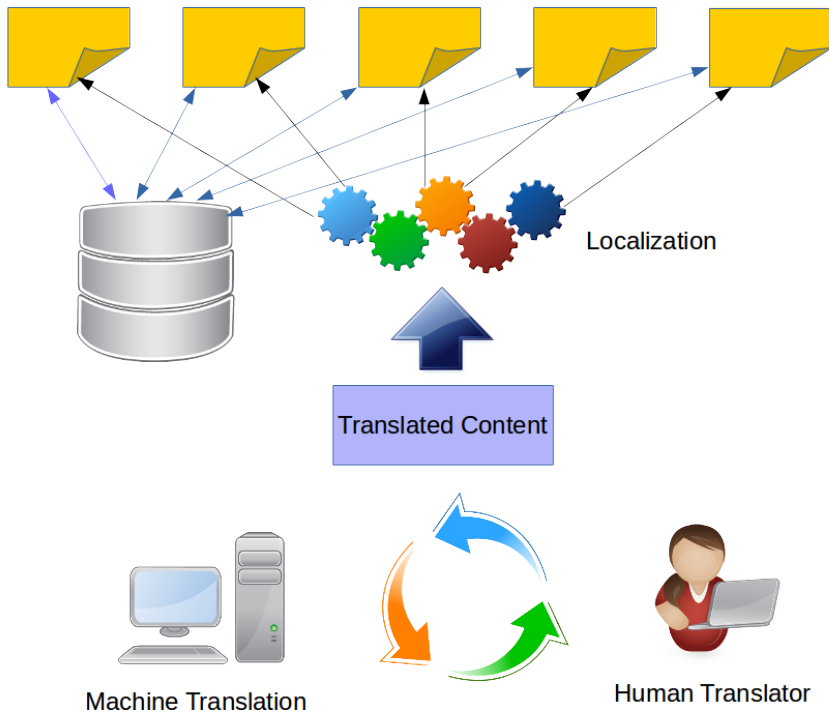
- Relatively easy to maintain
- Consistent versions across languages

Cons

database generated content is not translated

3. Multilingual Support to Both Database and Interface Templates

- Adding multilingual support to both database and website templates
- After switching languages, the system deploys interface elements, interaction messages and data entry in the chosen language



3. Multilingual Support to Both Database and Interface Templates ...

Pros

- Relatively easy to maintain
- Consistent versions across languages
- Interacts with users in multiple languages

3. Multilingual Support to Both Database and Interface Templates ...

Pros

- Relatively easy to maintain
- Consistent versions across languages
- Interacts with users in multiple languages

Cons

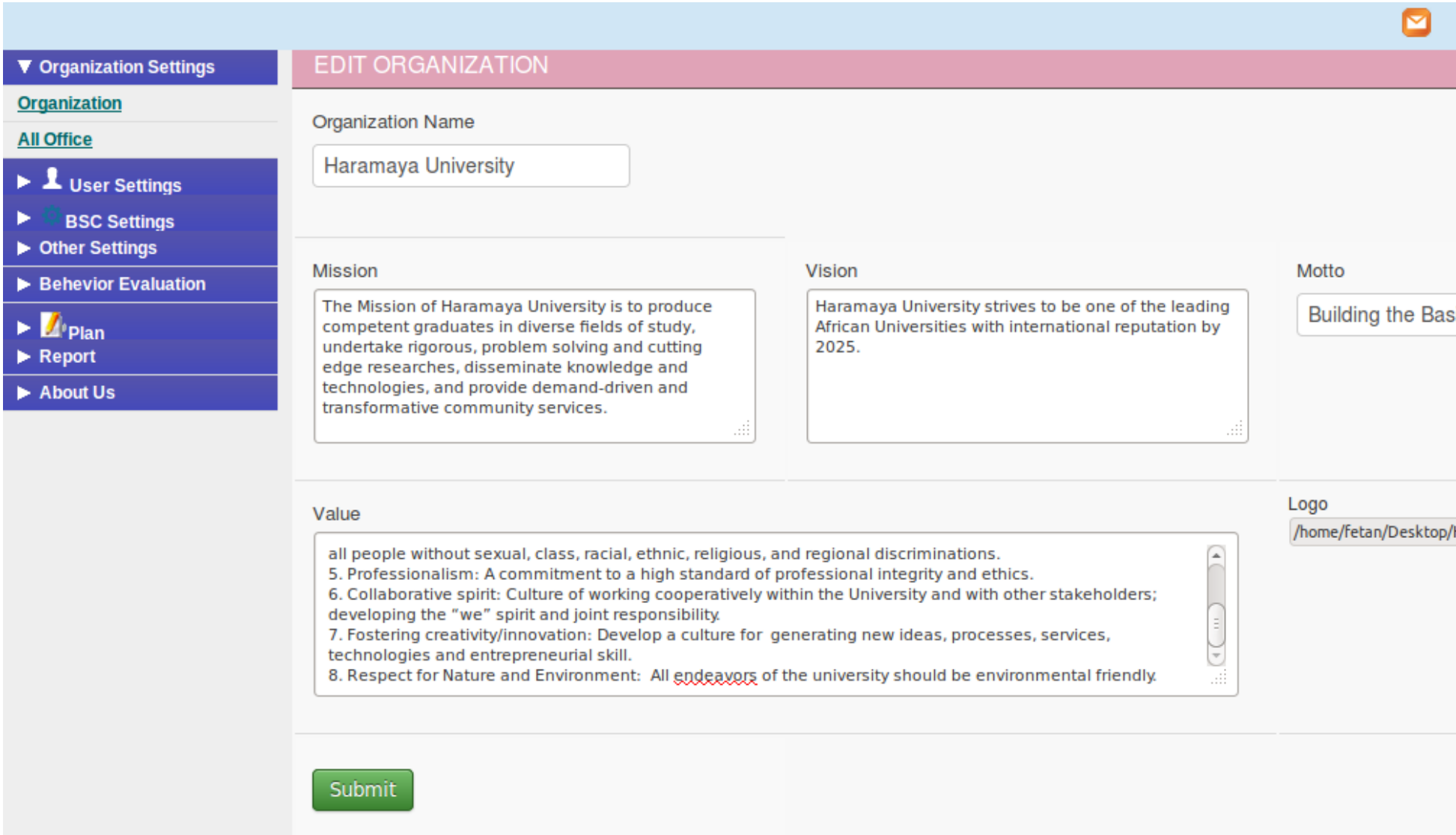
- Difficult to design
- Difficult to analyze the collected multilingual data

Outline

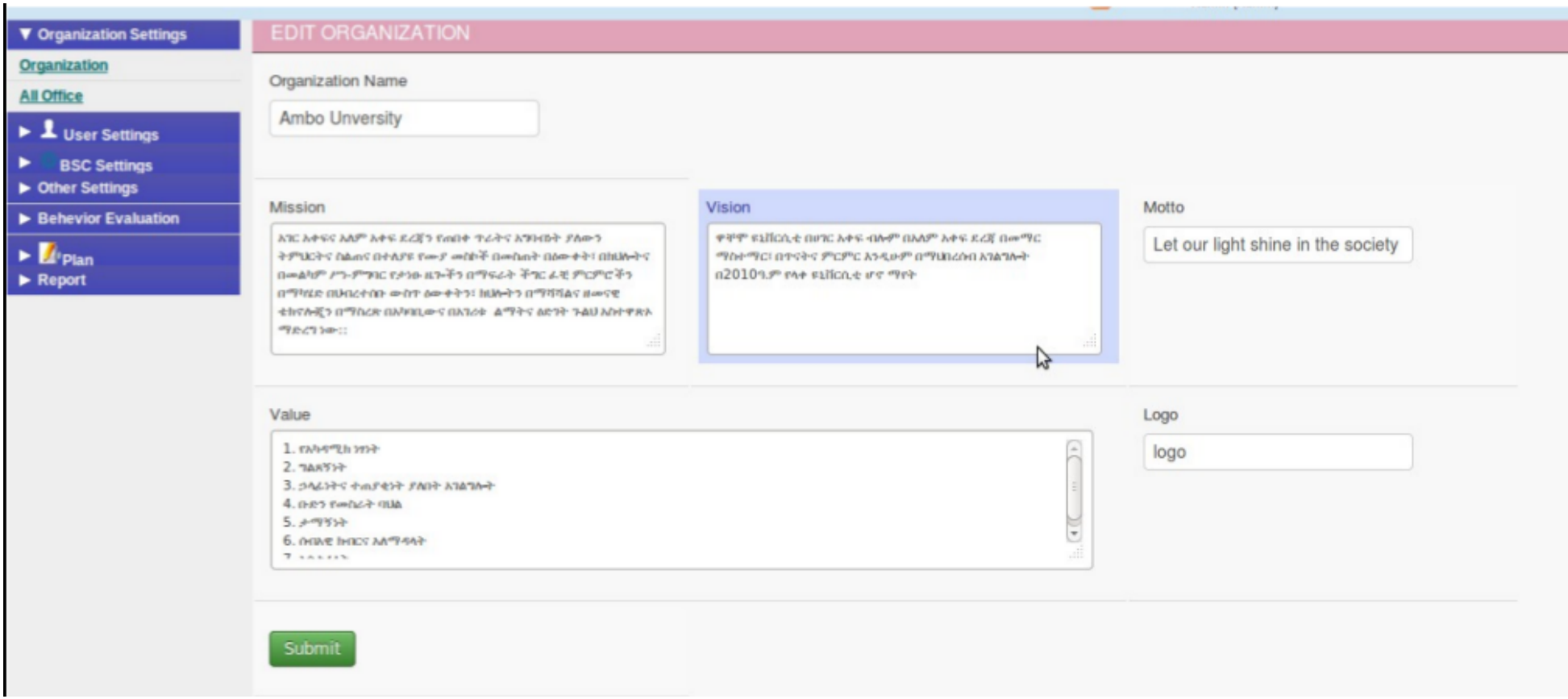
Multilingual Organizational Intelligence System

- Multilingual organizational intelligence system is crucial to communicate with employees and customers
- Application was developed in English
- Have the interface elements and user interaction messages into Amharic to produce this file
- The file is loaded by the application when users switch languages
- Can be extended to any number of language so long as this file is translated

Multilingual Organizational Intelligence System ...

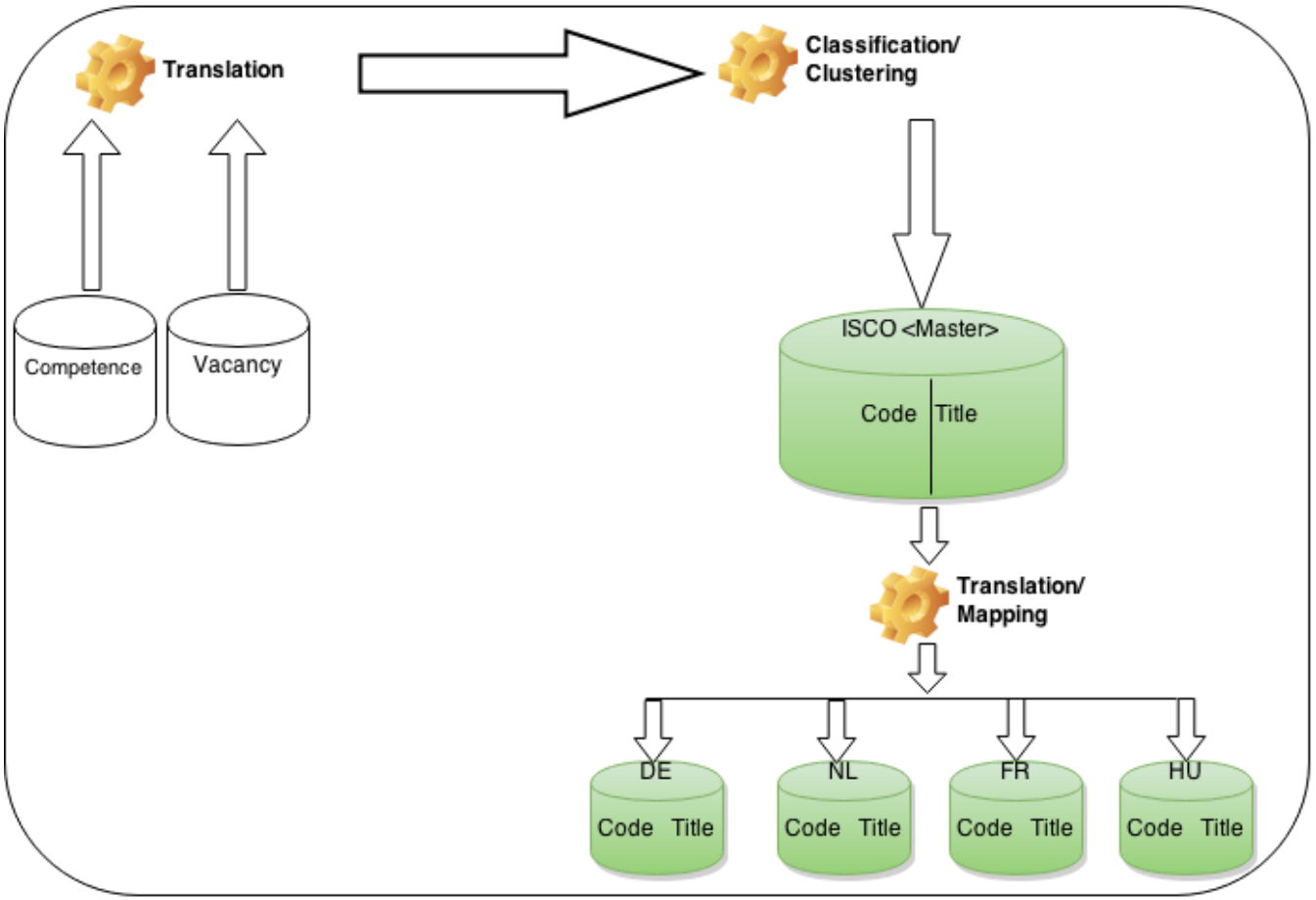


Multilingual Organizational Intelligence System ...



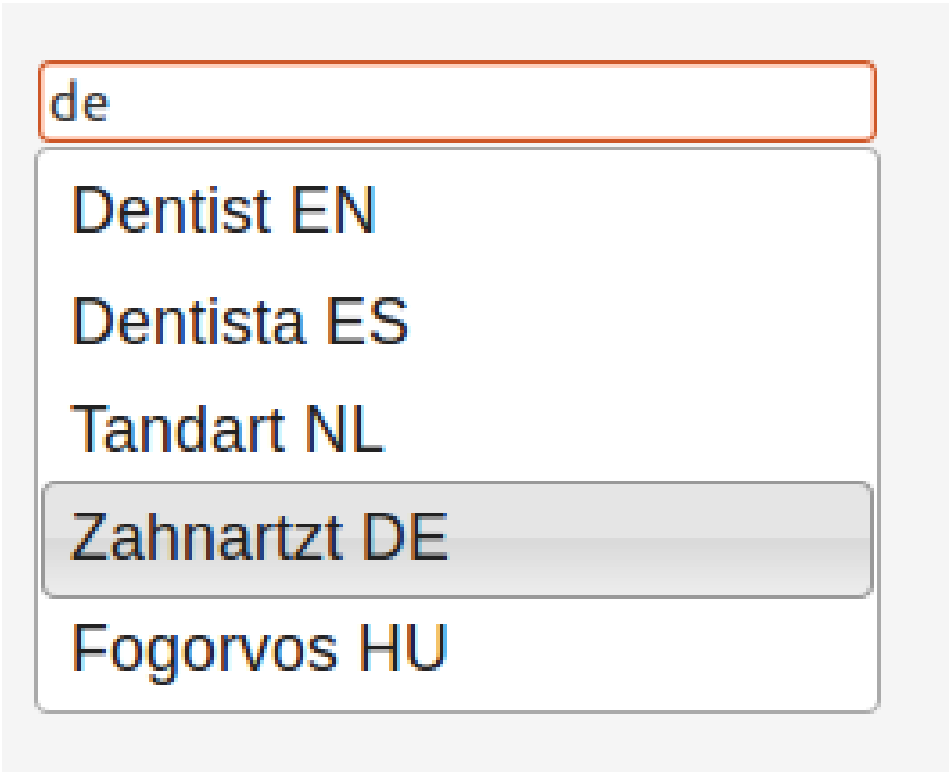
Multilingual Career Development System

Multilingual career development system



ML Career Development

- Assisting data collection using Dynamic Text Field
- Suggesting job titles in different languages



Outline

Machine Translation

- Machine translation is "translation carried out by a computer" [4]
- It learns from a collection of rules or data to preform automatic translation
- It is useful in dynamic environments [5]
- It can be used as a core component of multilingual systems development

[4] https://en.oxforddictionaries.com/definition/machine_translation
[5] Carlos Teixeira. 2013. "Where Humans Meet Machines"

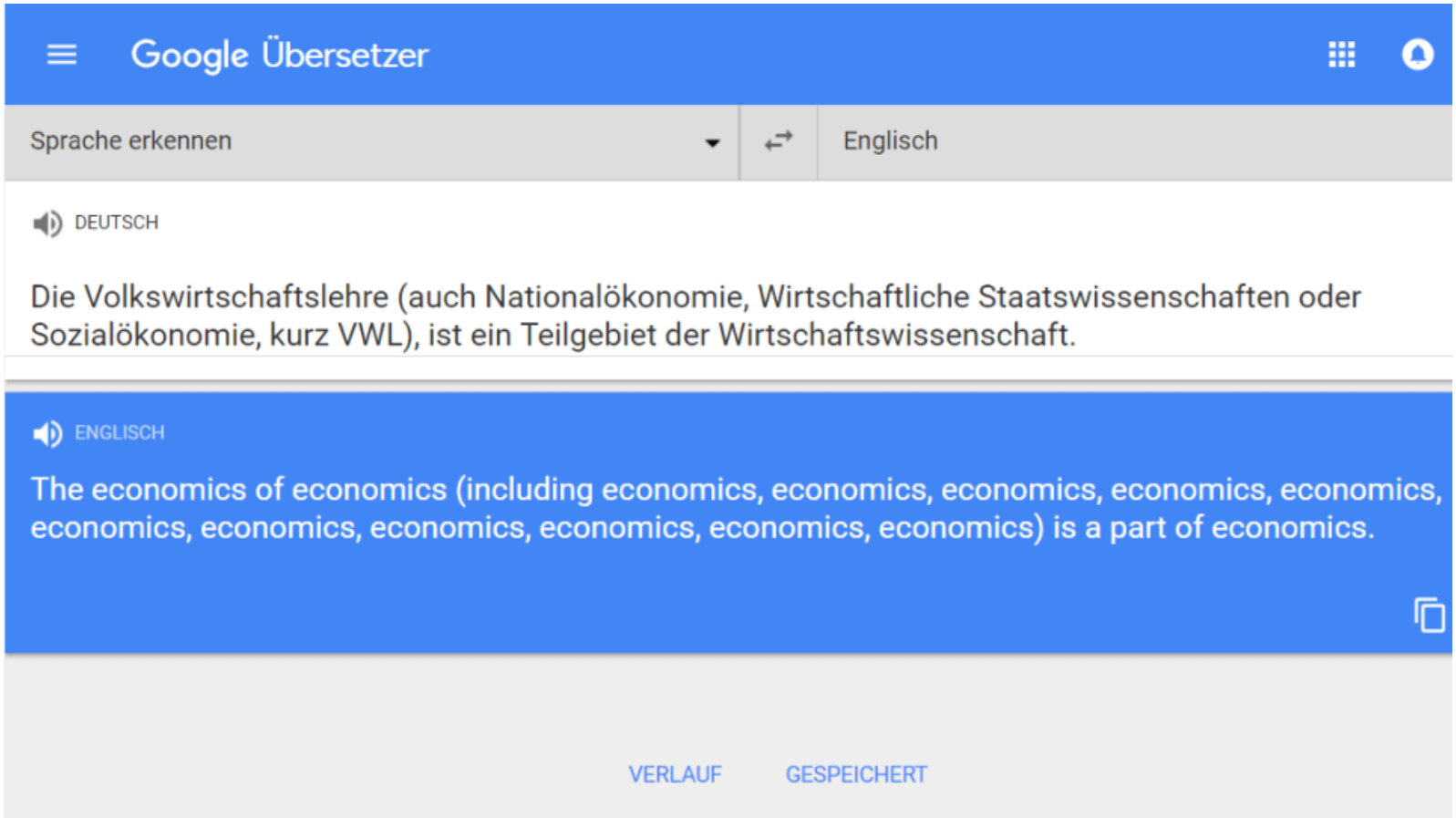
Manual Translation

- Slow
- Expensive
- Requires the translator to be a professional in specialized field.
- Repetitive, monotonous and thus boring
- Accurate and reliable

Machine Translation (MT)

- Minimize cost
- Maximize speed
- Aids human translators to do their job more efficiently
- Allows documents to be translated that would otherwise remain untranslated due to lack of resources
- Has not yet achieved the pinnacles of human translator's art

Machine Translation (MT) ...



(Source: https://www.googlewatchblog.de/2017/06/the-kuriose-google-translate/)

Types of MT Approaches

- The two most common machine translation approaches are: Rule-Based Machine Translation (RBMT), and Statistical Machine Translation (SMT)
- RBMT depends on large number of linguistic rules and millions of bilingual dictionaries for each language pair
- SMT uses computing power to build sophisticated data models generated from the analysis of monolingual and bilingual training data to translate one language into another

SMT	RBMT
Data	Data
Memory	Memory
CPU	CPU
Time	Time
Human Effort	Human Effort

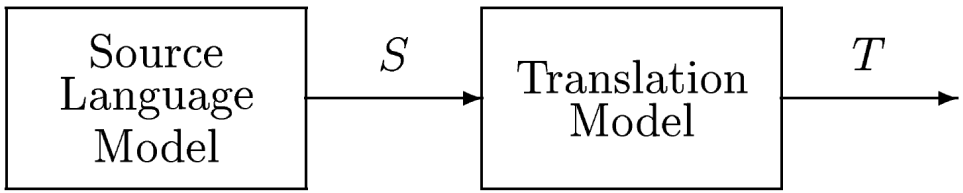
SMT	RBMT
Data	Data
Memory	Memory
CPU	CPU
Time	Time
Human Effort	Human Effort

- Technology is high enough, i.e., memory and CPU are no more big issues
- Data is becoming no more a problem
- **Time continues to be scarce and expensive**
- **Human effort continues to be scarce and expensive**

Statistical Machine Translation

S is the source language text and

T is the target language text

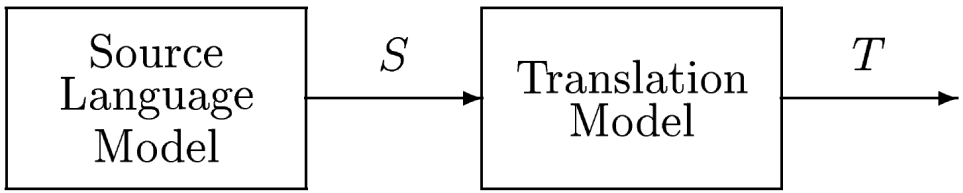


$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$

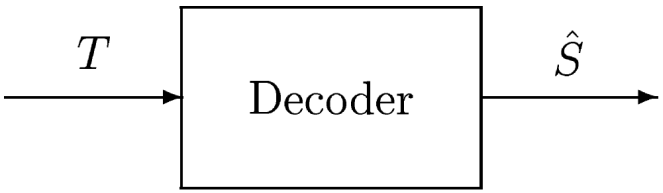
Statistical Machine Translation

S is the source language text and

T is the target language text



$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$

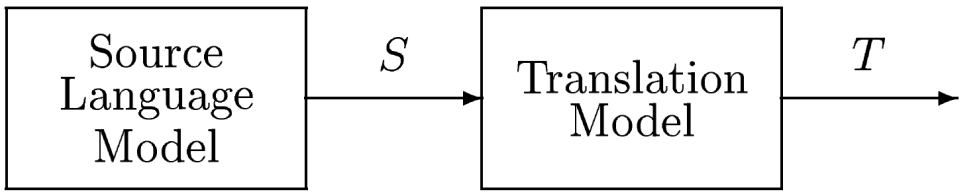


$$\hat{S} = \underset{S}{\operatorname{argmax}} \Pr(S \mid T) = \underset{S}{\operatorname{argmax}} \Pr(S, T)$$

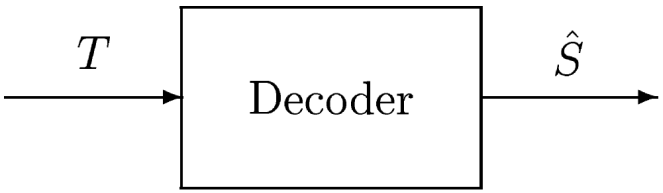
Statistical Machine Translation

S is the source language text and

T is the target language text



$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$



$$\hat{S} = \underset{S}{\operatorname{argmax}} \Pr(S \mid T) = \underset{S}{\operatorname{argmax}} \Pr(S, T)$$

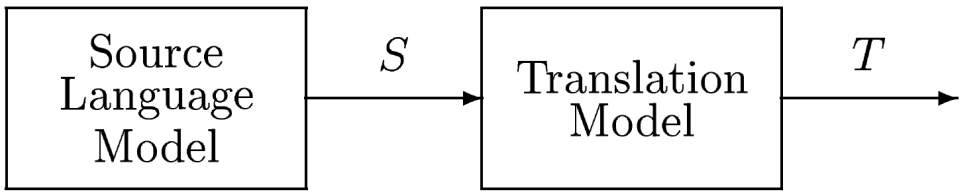
The cat drinks milk,

Die Katze trinkt Milch

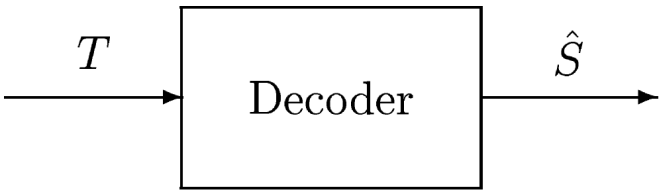
Statistical Machine Translation

S is the source language text and

T is the target language text



$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$



$$\hat{S} = \underset{S}{\operatorname{argmax}} \Pr(S \mid T) = \underset{S}{\operatorname{argmax}} \Pr(S, T)$$

The cat drinks milk,

Die Katze trinkt Milch

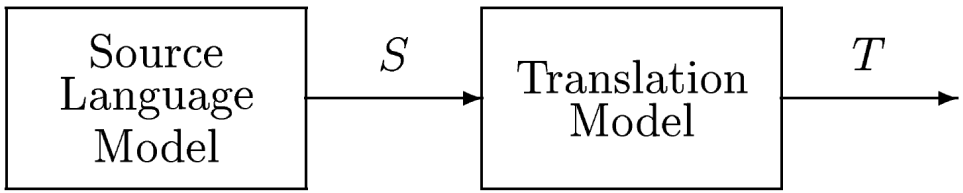
The dog **eats meat**,

Der Hund **frisst Fleich**

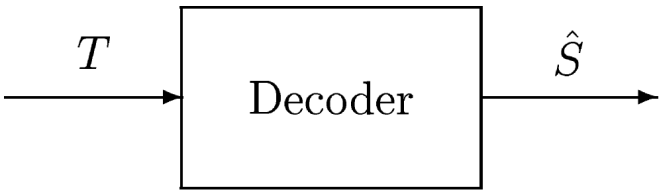
Statistical Machine Translation

S is the source language text and

T is the target language text



$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$



$$\hat{S} = \underset{S}{\operatorname{argmax}} \Pr(S \mid T) = \underset{S}{\operatorname{argmax}} \Pr(S, T)$$

The cat drinks milk,

Die Katze trinkt Milch

The dog **eats meat**,

Der Hund **frisst Fleich**

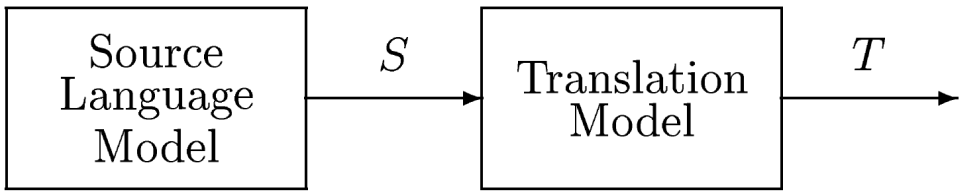
She has a **cat**,

Sie hat eine **Katze**

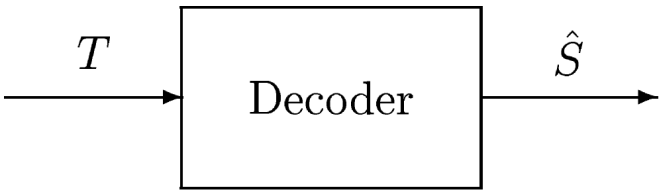
Statistical Machine Translation

S is the source language text and

T is the target language text



$$\Pr(S) \times \Pr(T \mid S) = \Pr(S, T)$$



$$\hat{S} = \underset{S}{\operatorname{argmax}} \Pr(S \mid T) = \underset{S}{\operatorname{argmax}} \Pr(S, T)$$

The cat drinks milk,

Die Katze trinkt Milch

The dog **eats meat**,

Der Hund **frisst Fleich**

She has a **cat**,

Sie hat eine **Katze**

The cat eats meat,

?????

Statistical Machine Translation

- Language Modeling
  - is analogous to tossing a coin having as many faces as the vocabulary of the language in question
  - Discourse is a sequence of tossing the coin
  - What is the probability that one face appears after one face is seen?
  - Given a sentence $s = w_1 w_2 \ldots w_{n-1} w_n$

$$p(S) = p(w1) \times p(w2|w1) \times p(w3|w1,w2) \times \ldots \times p(wn|w1,w2,\ldots wn-1)$$

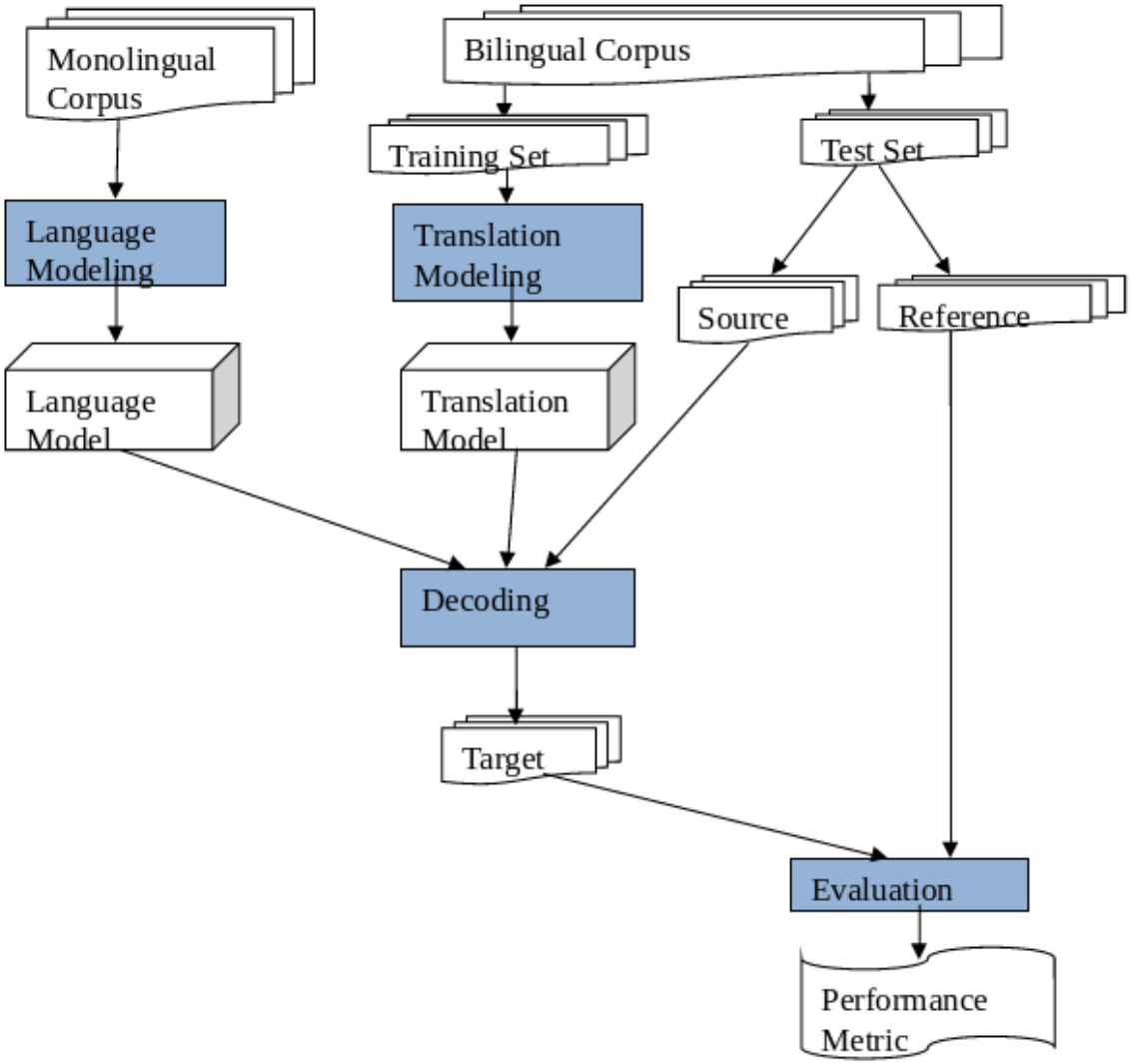
$$= \prod_{i=1}^{n} (\mathrm{p}(\mathrm{wi}|\mathrm{w1}..\mathrm{wi}-1)$$

Statistical Machine Translation

- Translation Modeling
  - TM - is the probability distribution P(T|S)
  - Given parallel sentences S and T,

$$P(T/S) = \frac{P(S/T)xP(T)}{P(S)}$$
$$= P(S/T)xP(T)$$

English - Oromo SMT ...



English - Oromo SMT

- Data
  - 20K Sentence pairs (EN, OM) or (300,000 words) for TM
  - 62K Sentences (OM) or (1,024,156 words) for LM

English - Oromo SMT ...

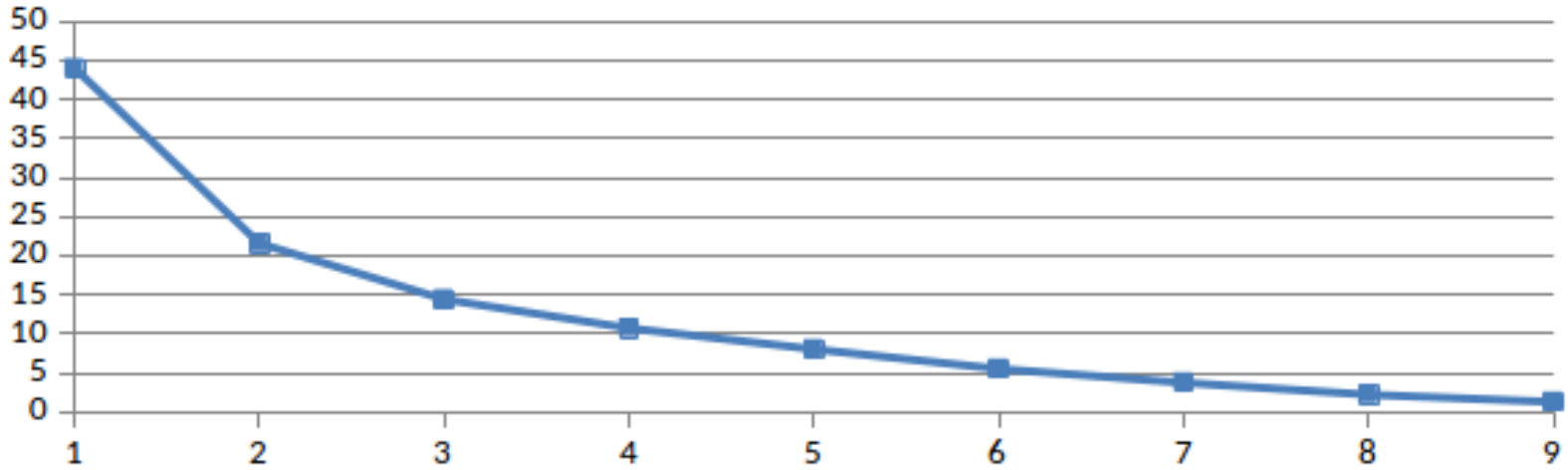
- Hardware Environment
  - Processors: 8 dual core processors operating at 2.8GHz
  - Memory: 128GB RAM

English - Oromo SMT ...

- Software Environment
  - Preprocessing : PERL and python scripts
  - Language Modeling: SRILM
  - Alignment: GIZA++
  - Phrase-based Translation Modeling: Moses
  - Decoding: Moses
  - Postprocessing: PERL scripts
  - Evaluation: PERL Script
  - Demonstration: Python Scripts

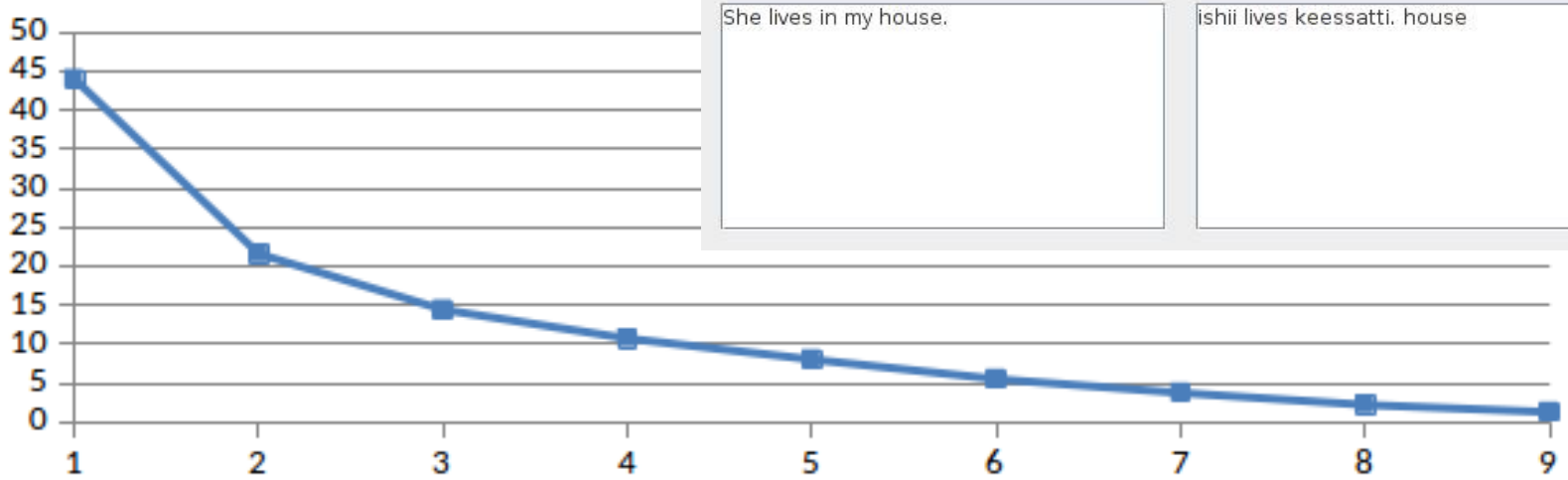
English - Oromo SMT ...

- Result



English - Oromo SMT ...

- Result



  - Average BLEU Score of 17.74
  - As N increases, accuracy decreases sharply

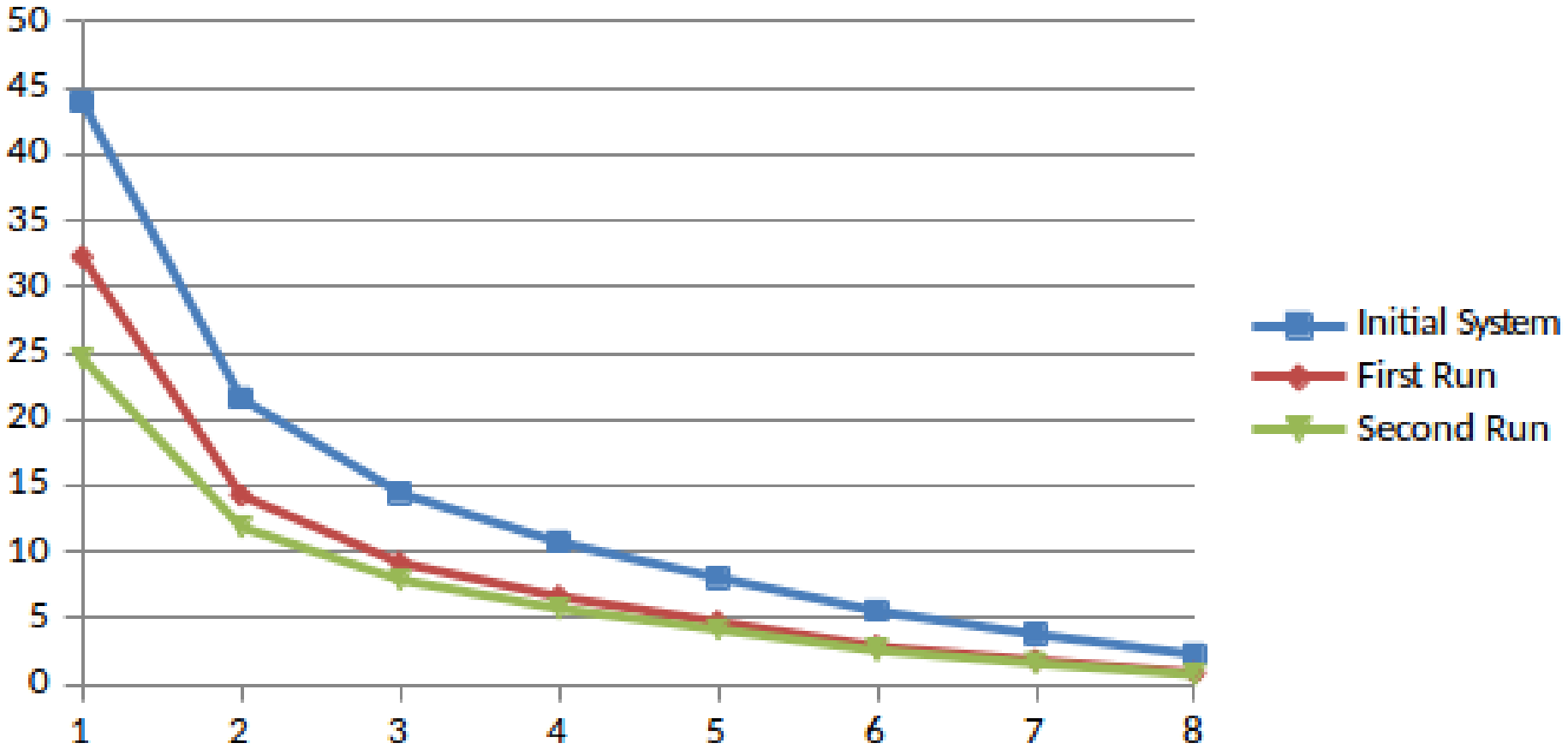
English - Oromo SMT ...

- Reasons for low score are due to
  - Poor data quality of the parallel data
  - Availability of single reference translation
  - Diversity in domain of the data
  - Limited size of the data

Sentence 34)	BLEU: 0.3672 (0.5455/0.4/0.3333/0.25)
Source	PART TWO Payment of Rural Land Use Payment And Income Tax.
Ref 0	2) Dhaabbileefi invastarootni lafa baadiyyaa seeraan kennameef hundumarratti kaffaltii itti fayyadama lafa baadiyyaa raawwachuu qabu.
Output 0	KUTAA LAMA kaffaltii kaffaltii itti fayyadama lafa baadiyyaa fi gibira galii.

English - Oromo SMT ...

Trying to use translation output as input

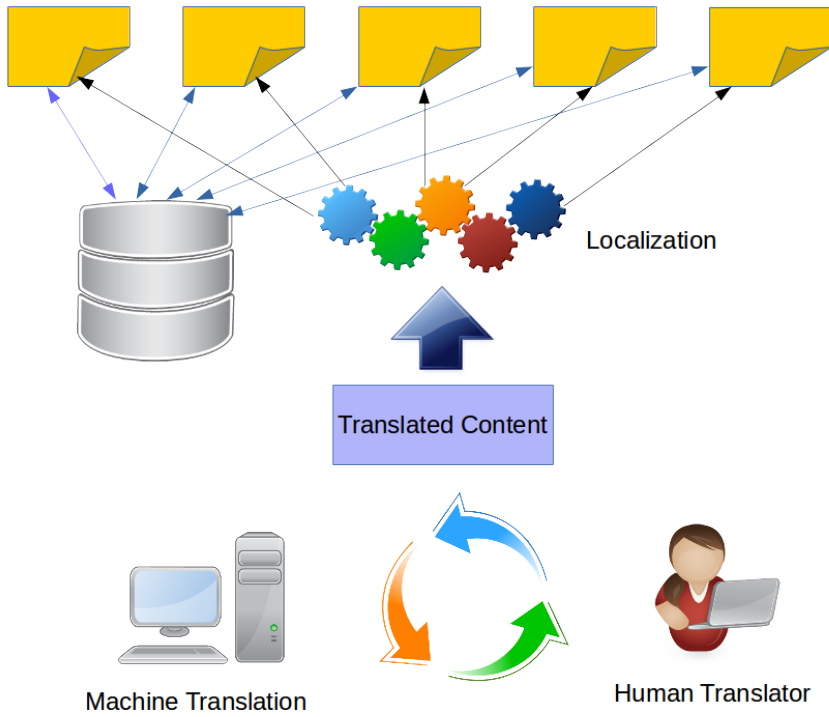


Conclusion

- Multilingual system are crucial in helping effective communication with users of diverse background
- Translation is the most important component of ML System
- Machine translation provides cost-effective translation with short turnaround time

Future Works

- Development of language resources for MT
- Integration of automatic translation into ML systems engineering
- Analysis of data collected through multilingual interfaces



Thank You