Facilitating cross-language retrieval and machine translation by multilingual domain ontologies

Petr Knoth, Trevor Collins, Elsa Sklavounou, Zdenek Zdrahal
Knowledge Media Institute, The Open University, UK
Systran, France

KMi

22th May 2010

Outline

Outline

Context and goals

- Multilingual access and retrieval of eLearning materials particularly important in domains that are quickly evolving:

Context and goals

- Multilingual access and retrieval of eLearning materials particularly important in domains that are quickly evolving:
  - Lecturers required to often change their materials (e.g. genetics, nanotechnology)
  - Lecturers usually required to produce and deliver teaching materials in their language.

Context and goals

- Multilingual access and retrieval of eLearning materials particularly important in domains that are quickly evolving:
  - Lecturers required to often change their materials (e.g. genetics, nanotechnology)
  - Lecturers usually required to produce and deliver teaching materials in their language.
- Typical problems of students: How to find the interesting materials? How to use them?

Context and goals

- Multilingual access and retrieval of eLearning materials particularly important in domains that are quickly evolving:
  - Lecturers required to often change their materials (e.g. genetics, nanotechnology)
  - Lecturers usually required to produce and deliver teaching materials in their language.
- Typical problems of students: How to find the interesting materials? How to use them?

Objective

To show how can ontologies be used to improve the multilingual access to domain specific information.

Eurogene

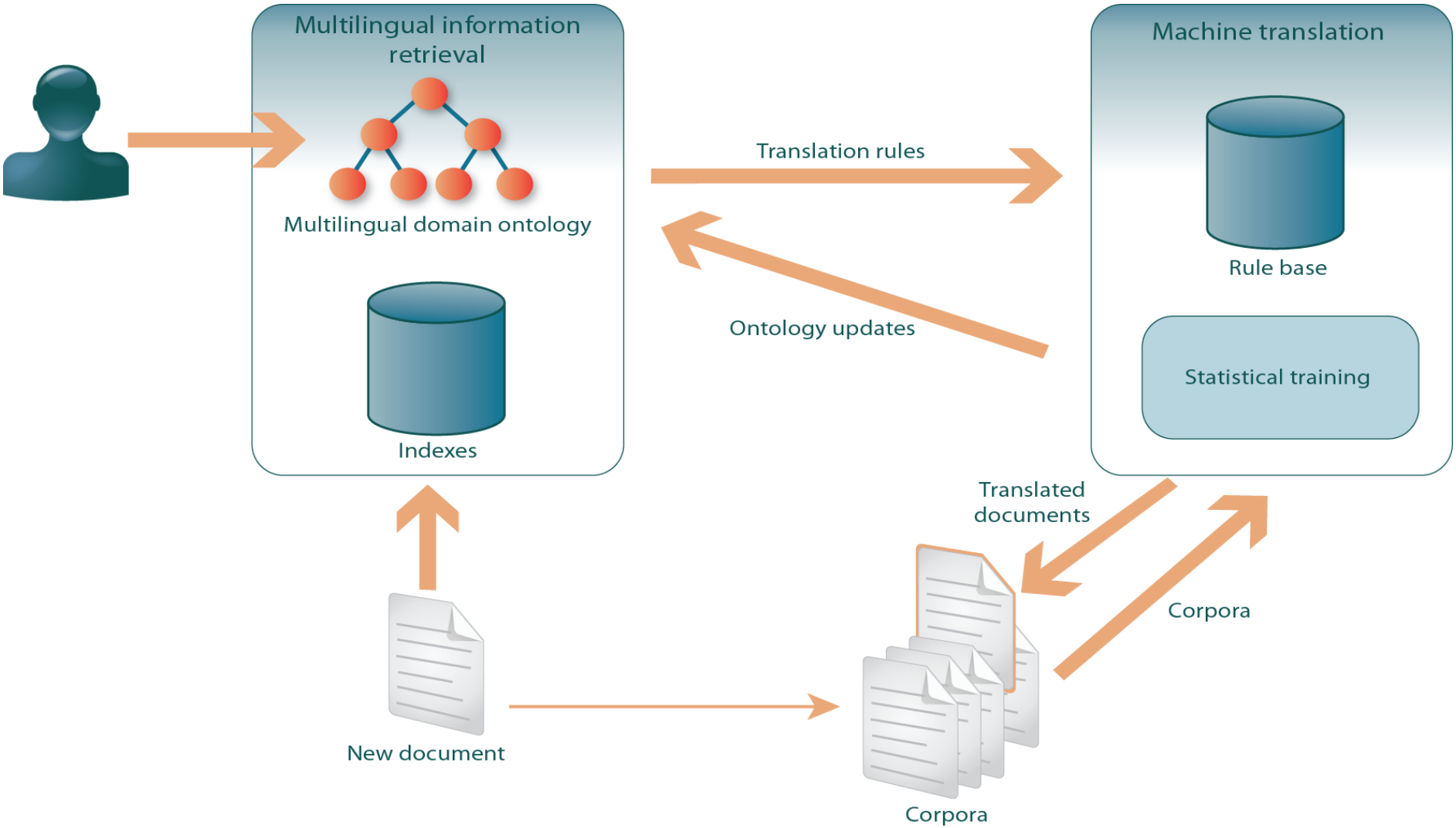
- A 3-year eContentplus supported project (18 content providers, 3 technical partners).

Eurogene

- A 3-year eContentplus supported project (18 content providers, 3 technical partners).
- Architecture for accessing and sharing multilingual resources is one of the project subgoals (KMI & Systran).
  - *Cross-language information retrieval (CLIR)*
  - *Machine translation (MT)*.
  - Both should be synchronized for terminology.

Outline

Architecture



Domain CLIR approaches

- *MT approach* - Query translated from the source language to the target language and submitted to the search system.

Domain CLIR approaches

- *MT approach* - Query translated from the source language to the target language and submitted to the search system.
  (a) MT system used to translate the query to all languages of interest.

Domain CLIR approaches

- *MT approach* - Query translated from the source language to the target language and submitted to the search system.
  (a) MT system used to translate the query to all languages of interest.
  (b) A multilingual ontology used to map the submitted query to different languages.

Domain CLIR approaches

- *MT approach* - Query translated from the source language to the target language and submitted to the search system.
  (a) MT system used to translate the query to all languages of interest.
  (b) A multilingual ontology used to map the submitted query to different languages.
- *Statistical approach* - The system trained on a collection of texts (typically parallel). Query is then mapped to a language independent document vector using approaches, such as LSI (Dumais, 1997).

Domain CLIR approaches

- *MT approach* - Query translated from the source language to the target language and submitted to the search system.
  (a) MT system used to translate the query to all languages of interest.
  (b) A multilingual ontology used to map the submitted query to different languages.
- *Statistical approach* - The system trained on a collection of texts (typically parallel). Query is then mapped to a language independent document vector using approaches, such as LSI (Dumais, 1997).

We are using approach 1(a) because:

- Multilingual ontologies are well-suited for domain CLIR.
- Multilingual ontologies can also be used to adapt MT to the target domain.
- Parallel corpora in specific domains often not initially available.

Synergy of CLIR and MT

Method has two phases:

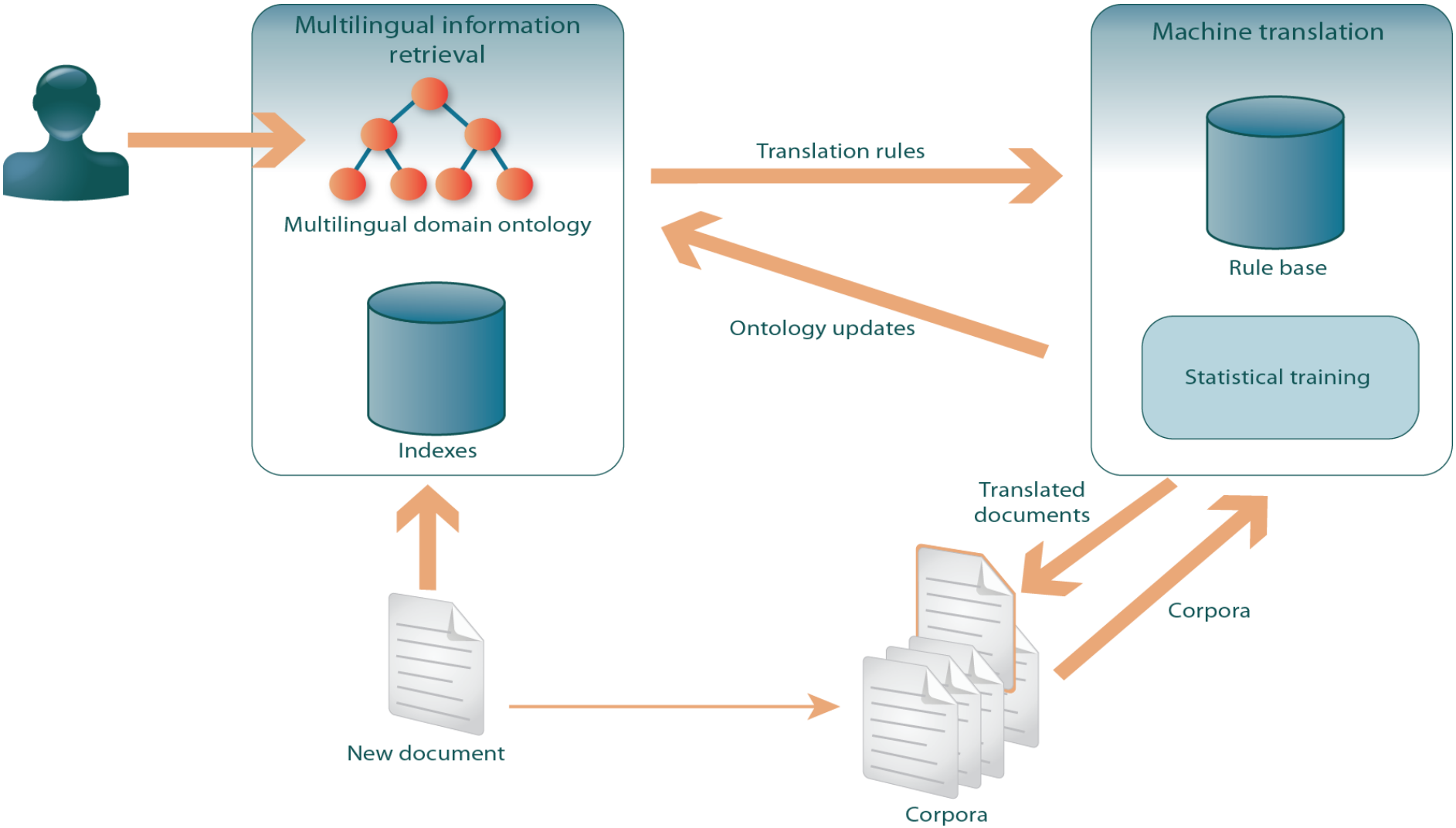
- *initialization* phase
  - Development of a *seed* monolingual ontology - reuse of an existing ontology or by using ontology learning methods (Cimiano and Völker, 2005, Sclano and Velardi 2007).
  - Extension of the ontology to multiple languages.

Synergy of CLIR and MT

Method has two phases:

- *initialization* phase
  - Development of a *seed* monolingual ontology - reuse of an existing ontology or by using ontology learning methods (Cimiano and Völker, 2005, Sclano and Velardi 2007).
  - Extension of the ontology to multiple languages.
- *bootstrapping* phase
  - Adaption of the MT dictionaries (hybrid MT system required).
  - Adaption of the multilingual ontology.

Architecture



Monolingual ontology

A monolingual ontology is a 4-tuple $O = \langle C, T, E, f \rangle$:

- C is a set of concepts.
- T is a set of terms (representations of concepts).
- E is a set of oriented relations (*is-a* relations), such that $\langle C, E \rangle$ is a directed acyclic graph.
- $f : T \rightarrow C$ is a surjective function from terms to concepts.

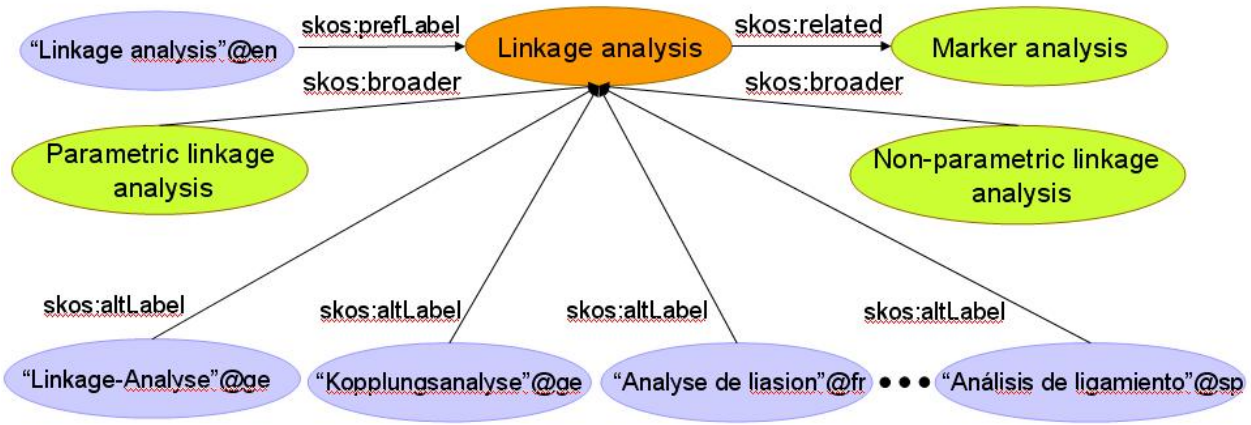
Multilingual ontology

A *multilingual ontology* is a 6-tuple $O' = \langle C, T, E, f, L, lang \rangle$

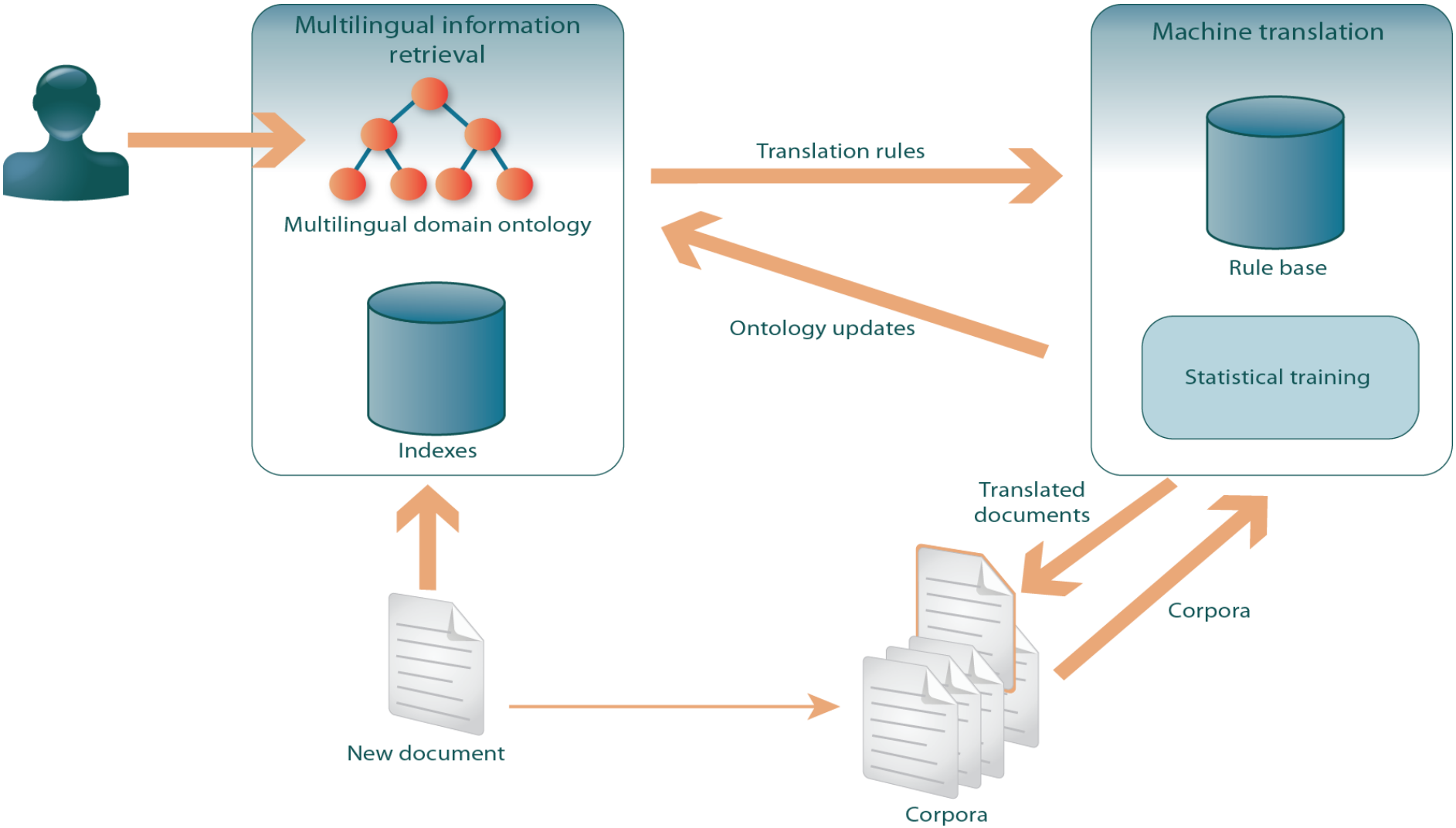
- A monolingual ontology $O = \langle C, T, E, f \rangle$
- L is the set of languages.
- $lang : T \rightarrow L$ is a mapping from terms to languages.

Multilingual ontology - example

- Lightweight ontology
- SKOS-like representation



Architecture



Bootstrapping phase - MT improvement step

The MT system is adapted to a specific domain:

- Using bilingual substitution rules of form: $t_{L_1} \rightarrow t_{L_2}$ extracted from the multilingual ontology.
- Rules satisfy the condition $f(t_{L_1}) = f(t_{L_2})$, where $t_{L_1} \in T_{L_1}, t_{L_2} \in T_{L_2}$ and T_{L_n} is defined as $T_{L_n} = \{t | lang(t) = L_n\}$.
- Flattening the ontological structure and deriving pairs for all supported combinations.

Bootstrapping phase - MT improvement step

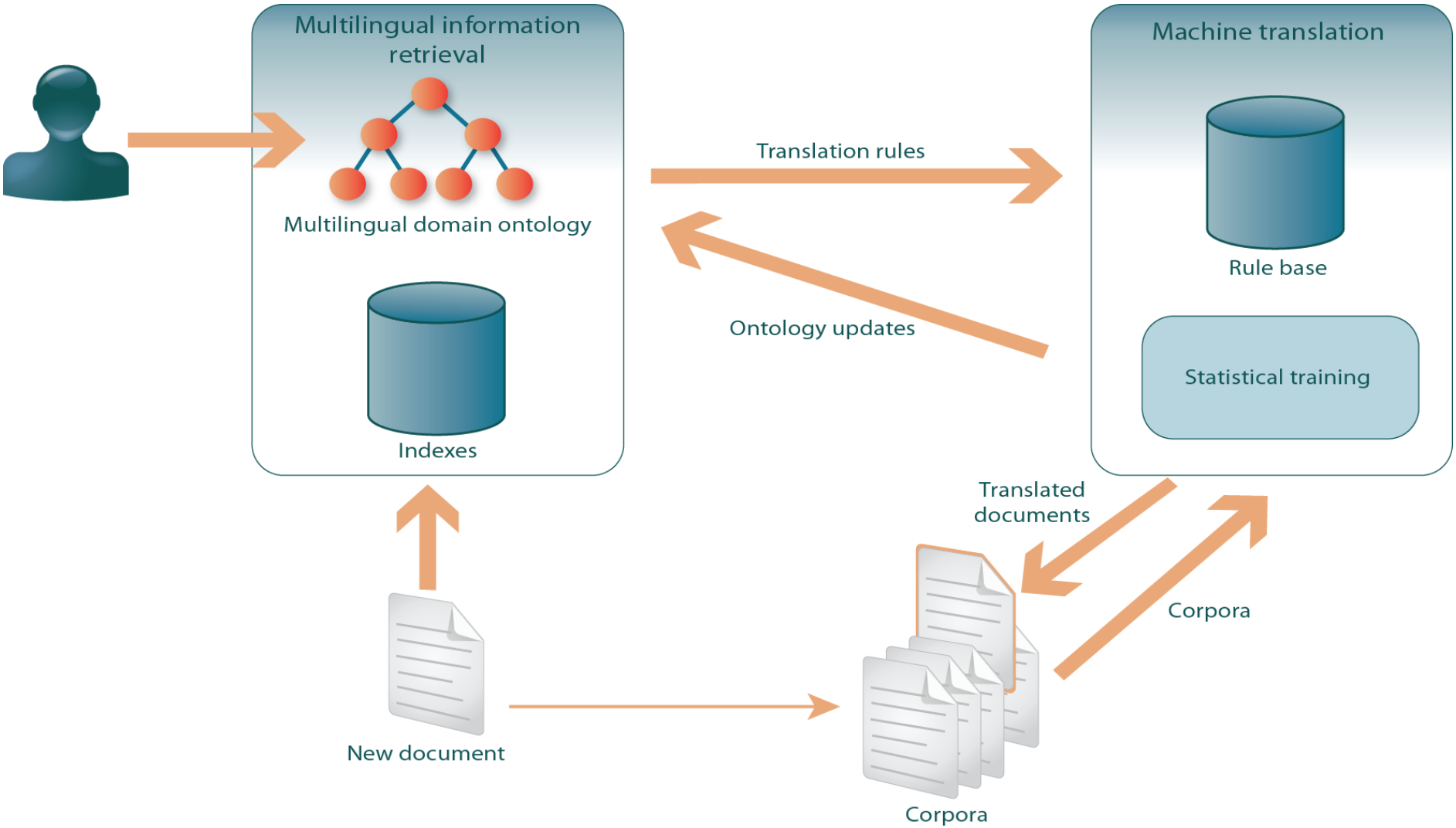
The MT system is adapted to a specific domain:

- Using bilingual substitution rules of form: $t_{L_1} \rightarrow t_{L_2}$ extracted from the multilingual ontology.
- Rules satisfy the condition $f(t_{L_1}) = f(t_{L_2})$, where $t_{L_1} \in T_{L_1}, t_{L_2} \in T_{L_2}$ and T_{L_n} is defined as $T_{L_n} = \{t | lang(t) = L_n\}$.
- Flattening the ontological structure and deriving pairs for all supported combinations.

Example

linkage analysis$_{en}$ $\rightarrow$ Kopplunganalyse$_{de}$
analyse de liasion$_{fr}$ $\rightarrow$ Analisis de ligamiento$_{sp}$

Architecture



Bootstrapping phase - Ontology refinement

- Content grows over time.
- New parallel texts can be automatically recognized (Resnik, 2003) and used by the machine translation system for training.
- If new pairs of text are discovered, statistical training is performed to improve the MT language model.

Bootstrapping phase - Ontology refinement

- Content grows over time.
- New parallel texts can be automatically recognized (Resnik, 2003) and used by the machine translation system for training.
- If new pairs of text are discovered, statistical training is performed to improve the MT language model.

The ontology is adapted by rules of form $(t_{L_1}, t_{L_2}, conf, lang_q)$ produced as an output of the statistical phase.

- *conf* is the confidence measure of translating term t_{L_1} to t_{L_2} estimated from text.
- $lang_q : T \rightarrow L$ is a mapping from terms to languages.

Bootstrapping phase - Ontology refinement

- Content grows over time.
- New parallel texts can be automatically recognized (Resnik, 2003) and used by the machine translation system for training.
- If new pairs of text are discovered, statistical training is performed to improve the MT language model.

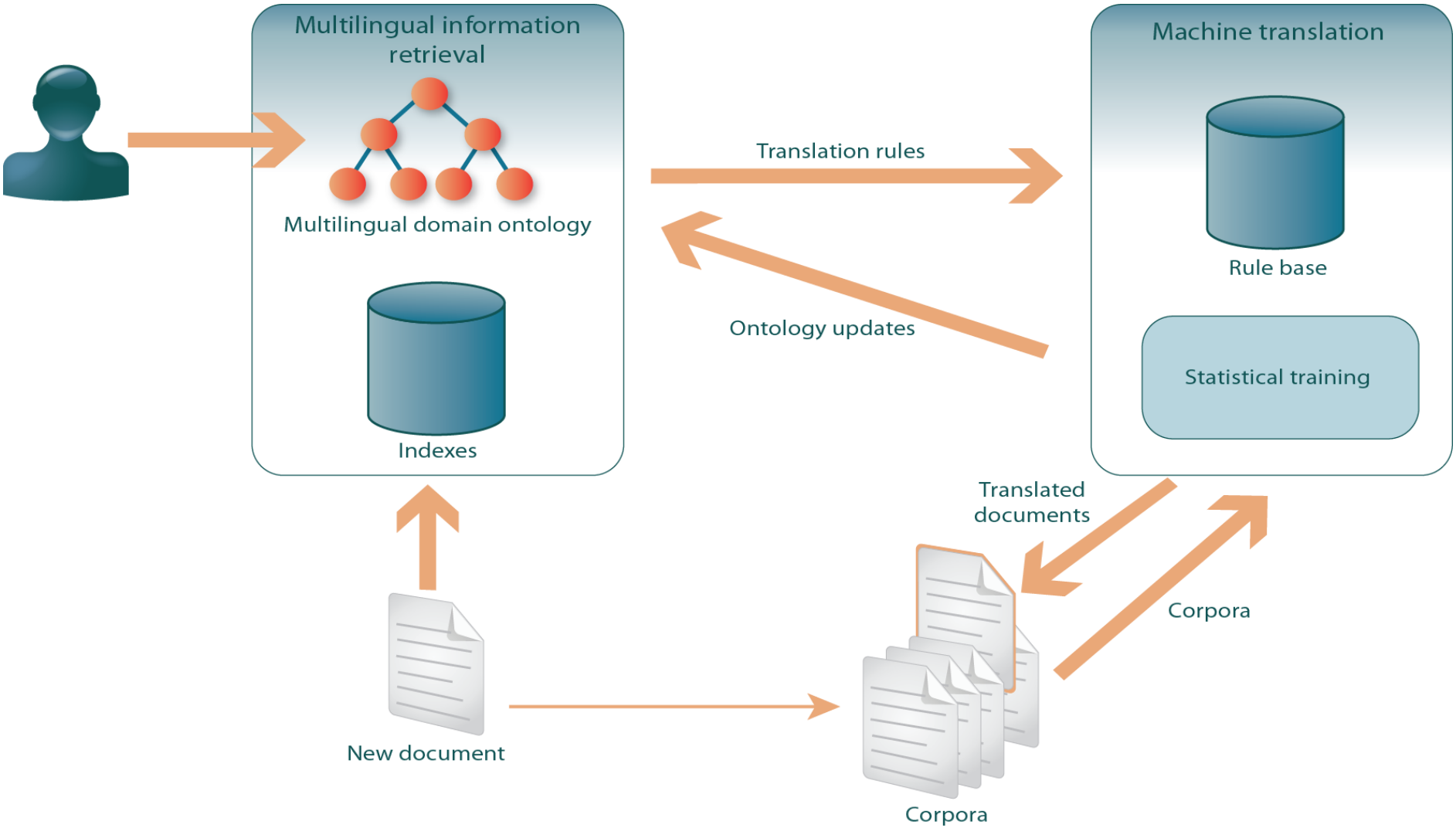
The ontology is adapted by rules of form $(t_{L_1}, t_{L_2}, conf, lang_q)$ produced as an output of the statistical phase.

- *conf* is the confidence measure of translating term t_{L_1} to t_{L_2} estimated from text.
- $lang_q : T \rightarrow L$ is a mapping from terms to languages.

Example

$\langle \text{indirekte DNA-Analyse}_{de}, \text{linkage analyse}_{en}, 0.85 \rangle$

Architecture



Outline

Eurogene portal

- Ontologies used for:
  - Annotation
  - CLIR
  - Query expansion
  - Navigation across content in multiple languages (semantic similarity)
  - MT

Eurogene portal

- Ontologies used for:
  - Annotation
  - CLIR
  - Query expansion
  - Navigation across content in multiple languages (semantic similarity)
  - MT
- Statistics:
  - About 20,000 files (papers, presentations, videos, images)
  - About 15,000 ontological terms.
  - Nine languages (English, French, Spanish, German, Greek, Italian, Dutch, Czech, Lithuanian).
- `http://eurogene.open.ac.uk/`

Outline

Properties of the approach

- Performance of both CLIR and MT should never decrease as a result of any bootstrapping iteration.
- Two steps where an error may be introduced:
  - The update of the MT rule base.
  - The update of the multilingual ontology.

Properties of the approach

- Performance of both CLIR and MT should never decrease as a result of any bootstrapping iteration.
- Two steps where an error may be introduced:
  - The update of the MT rule base.
  - The update of the multilingual ontology.
- Evaluation - many components involved:
  - Coverage and specificity of the ontology.
  - Amount of domain corpora available.
  - Performance of the statistical training.
  - Validity of human judgements.
  - Other factors ...

Outline

Summary

- Multilingual ontologies suitable for domains with rich terminology.

Summary

- Multilingual ontologies suitable for domains with rich terminology.
- Can be used as a synchronization component for domain adaption of CLIR and MT systems.

Summary

- Multilingual ontologies suitable for domains with rich terminology.
- Can be used as a synchronization component for domain adaption of CLIR and MT systems.
- The solution is easily readable and adjustable by humans.

Summary

- Multilingual ontologies suitable for domains with rich terminology.
- Can be used as a synchronization component for domain adaption of CLIR and MT systems.
- The solution is easily readable and adjustable by humans.
- Publishing of multilingual ontologies on the Web in a standard format may allow an application to decide which domain ontology to use for query expansion and for adaption of the MT system based on the context of the query.

Discussion

Thank you for attention !