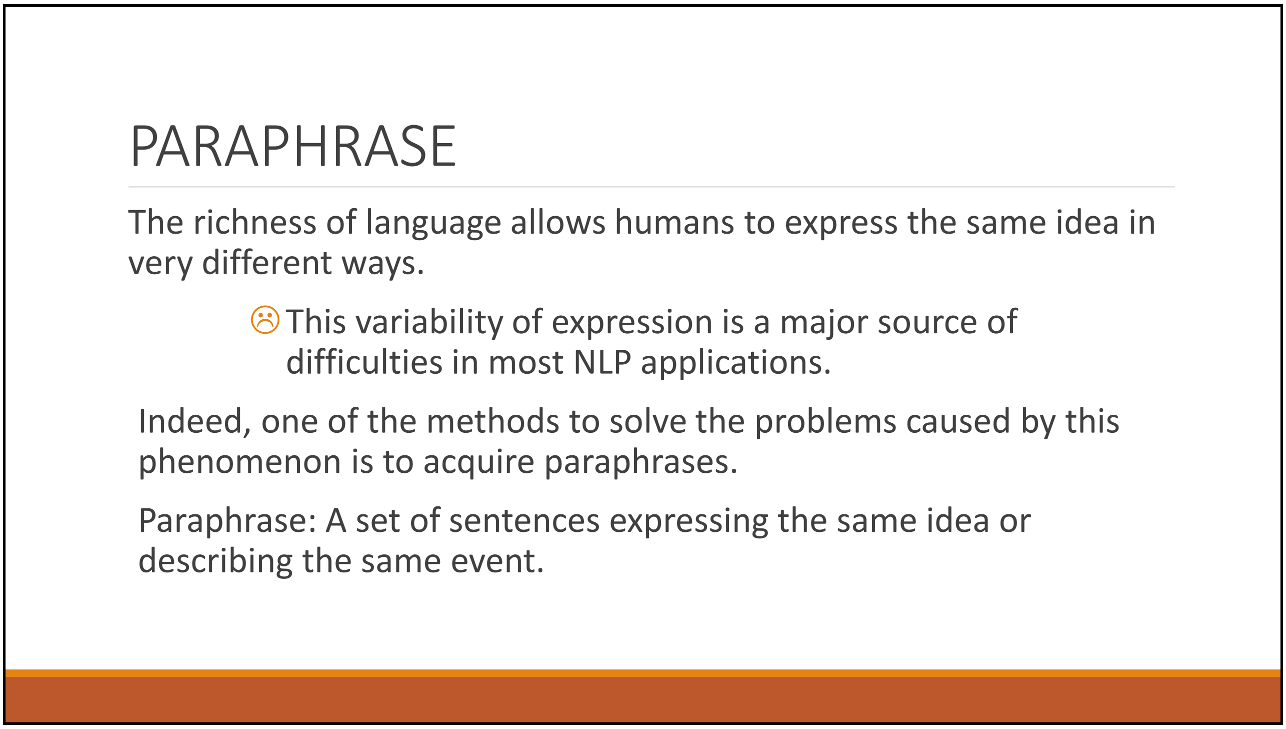
NATURAL LANGUAGE PROCESSING

LESSON 13: PARAPHRASING / ONTOLOGY MAPPING

OUTLINE

- Paraphrase
  - Methods
    - Linguistic resources
    - Corpus based)
- Ontology Mapping
  - Monolingual Ontology Mapping
  - Cross Lingual Ontology Mapping
  - CLOM Approaches



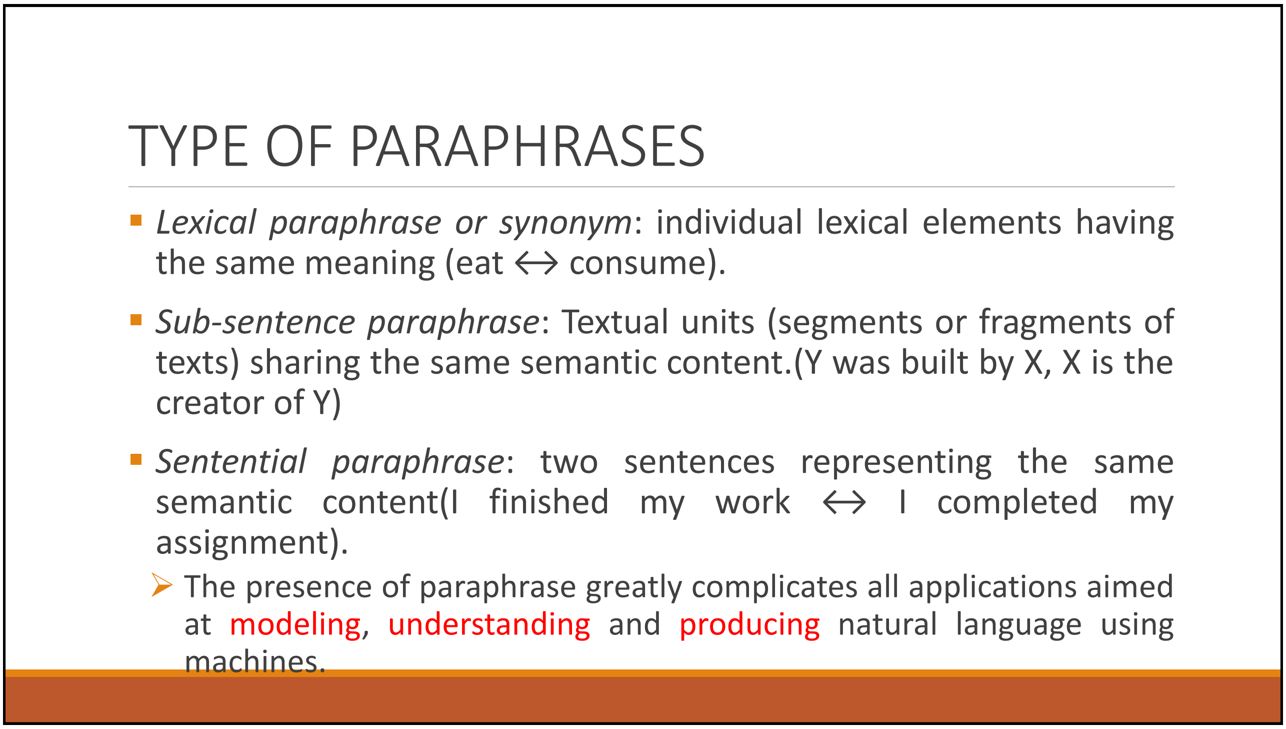
PARAPHRASE

The richness of language allows humans to express the same idea in very different ways.

☹ This variability of expression is a major source of difficulties in most NLP applications.

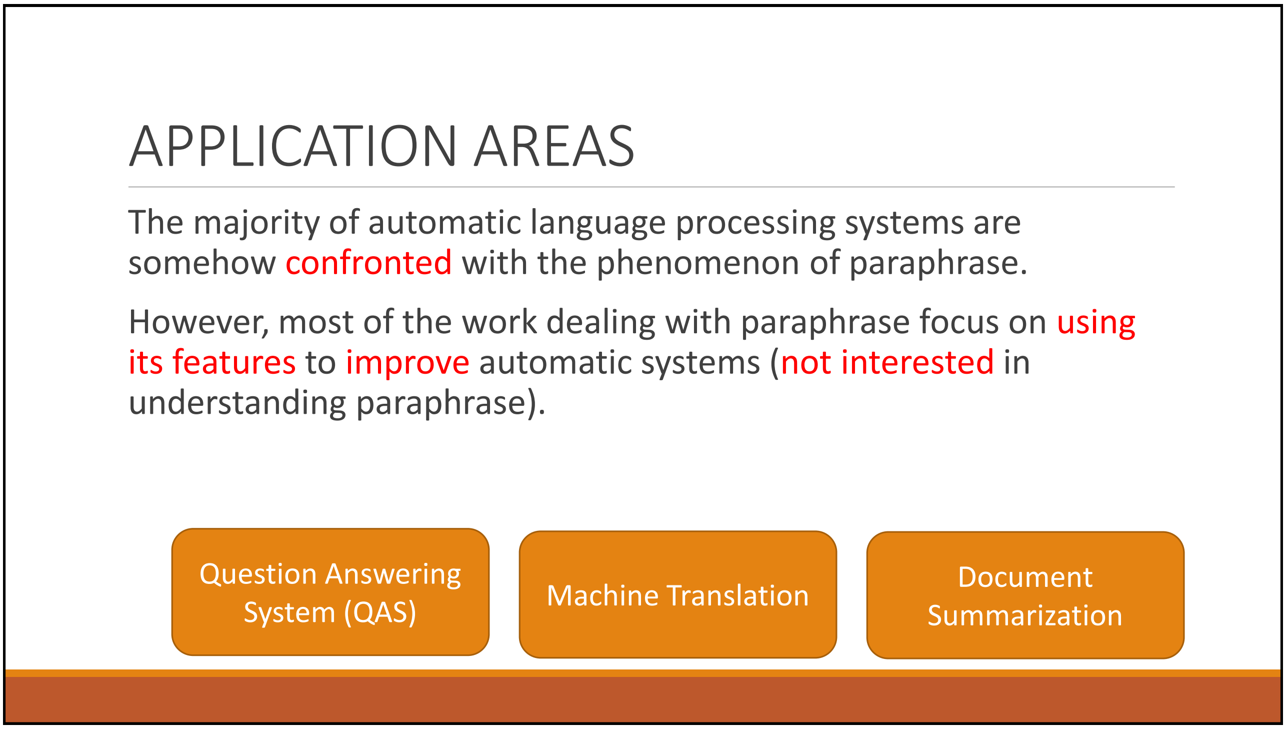
Indeed, one of the methods to solve the problems caused by this phenomenon is to acquire paraphrases.

Paraphrase: A set of sentences expressing the same idea or describing the same event.



TYPE OF PARAPHRASES

- *Lexical paraphrase or synonym*: individual lexical elements having the same meaning (eat ↔ consume).
- *Sub-sentence paraphrase*: Textual units (segments or fragments of texts) sharing the same semantic content.(Y was built by X, X is the creator of Y)
- *Sentential paraphrase*: two sentences representing the same semantic content(I finished my work ↔ I completed my assignment).
  - The presence of paraphrase greatly complicates all applications aimed at modeling, understanding and producing natural language using machines.



APPLICATION AREAS

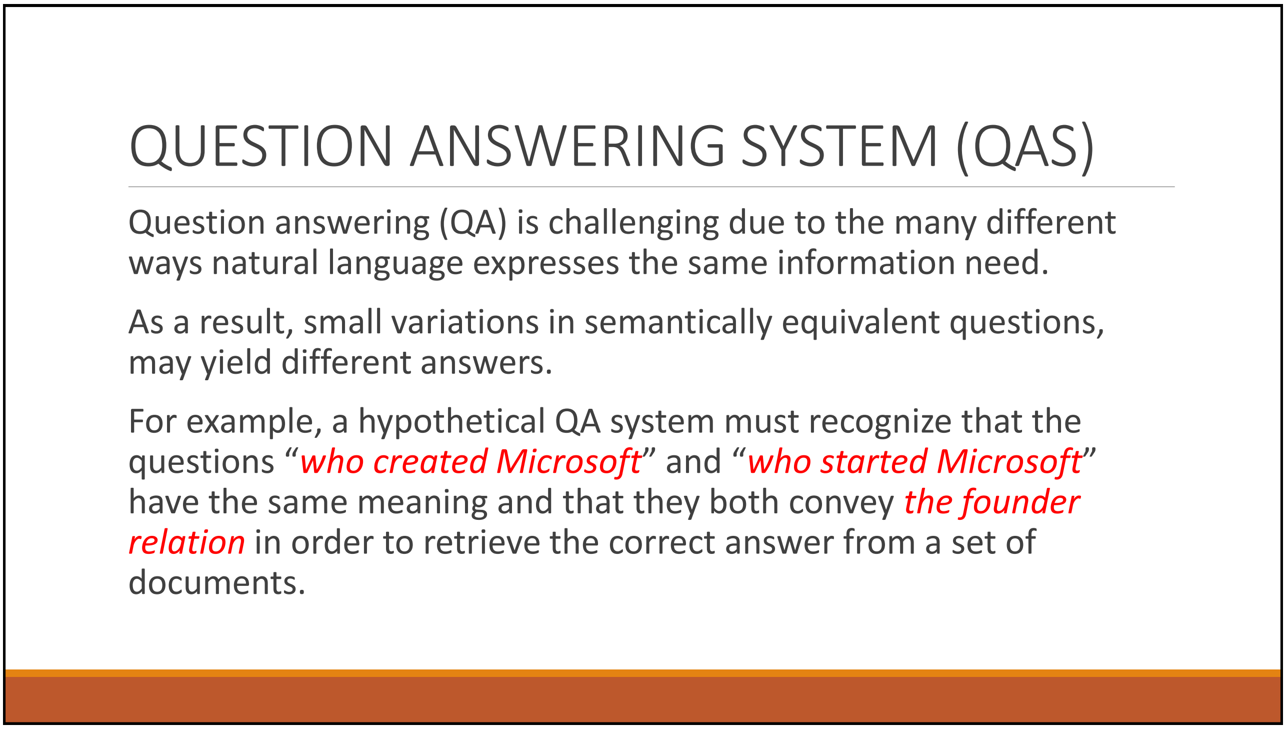
The majority of automatic language processing systems are somehow confronted with the phenomenon of paraphrase.

However, most of the work dealing with paraphrase focus on using its features to improve automatic systems (not interested in understanding paraphrase).

Question Answering System (QAS)

Machine Translation

Document Summarization



QUESTION ANSWERING SYSTEM (QAS)

Question answering (QA) is challenging due to the many different ways natural language expresses the same information need.

As a result, small variations in semantically equivalent questions, may yield different answers.

For example, a hypothetical QA system must recognize that the questions "*who created Microsoft*" and "*who started Microsoft*" have the same meaning and that they both convey *the founder relation* in order to retrieve the correct answer from a set of documents.

MACHINE TRANSLATION

The hypotheses produced by a system are evaluated by measuring their similarity to reference translations created by humans.

These similarity measures are essentially based on the number of groups of common words in the two sentences.

- However, it is impossible to identify the different formulations of the same semantic content with a single reference translation.

☹This can penalize the hypotheses of translation conveying the same meaning, but using expressions different from those present in the reference.

çizgi filmleri görmek istiyorum ↔ *I would like to watch cartoons (ref)*

Sys 1 - I want to see cartoons

Sys 2 - I would like to watch movies

DOCUMENT SUMMARIZATION

In automatic summarization, the identification of paraphrases can condense the information contained in several documents and improve the quality of automatic summaries.

Producing a paraphrase shorter than an original sentence can condense a text, an essential step in automatic summary.

The sentence " *She hates apple, orange, pear*." is summarized as "*She hates fruits*"

PREVİOUS WORKS

In the last years, several works have been concerned with the processing of paraphrase.

The extraction of paraphrases can be achieved two main methods:

Methods exploiting linguistic resources
Corpus-based methods.

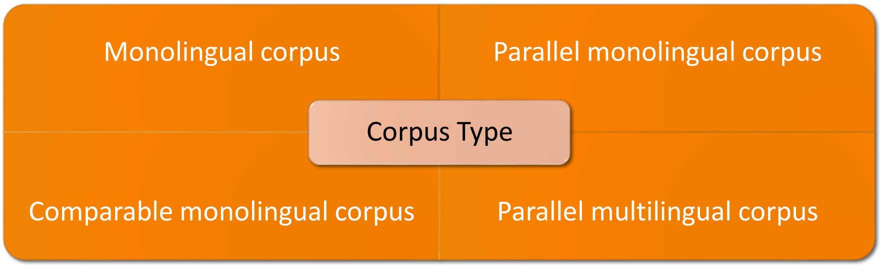
METHODS EXPLOITING LINGUISTIC RESOURCES

For a source segment, a paraphrase is obtained by replacing certain words with their synonyms.

1. Extract synonyms for the terms to be substituted from a semantic network such as Wordnet.
2. Choose the synonym most adapted to the context of appearance of each term.

CORPUS-BASED METHODS

The techniques used to extract paraphrases are generally very dependent on the types of corpora on which they were developed.


Monolingual corpus
Parallel monolingual corpus
Corpus Type
Comparable monolingual corpus
Parallel multilingual corpus

MONOLINGUAL CORPUS

A corpus of similar documents from the Web.

For example, the automatic recognition of paraphrases is done from the revisions of WIKIPEDIA (It is a free online encyclopedia, created and edited by volunteers around the World).

Hubert Beuve-Méry

He founded the French-speaking [newspaper → daily paper] "Le Monde" in 1944.

COMPARABLE MONOLINGUAL CORPUS

It is composed of associated text pairs based on a measure of textual similarity possibly, such as newspaper articles published in the same time interval.

CNN - Bush says he'll *helps* NY with $20 billion

Washington Post - Bush *Reassures* New York of $20 Billion

PARALLEL MONOLINGUAL CORPUS

It consists of pairs of equivalent meaning statements aligned in a supervised manner, such as

- multiple translations of books
  - Emma burst into tears and he tried to comfort her, saying things to make her smile.
  - Emma cried, and he tried to console her, adorning his words with puns.
- or groups of questions having the same answer
  - How many ounces are there in a pound ?
  - What's the number of ounces per pound ?

PARALLEL MULTILINGUAL CORPUS

It consists of pairs of sentences available in two or more languages (such as transcripts of European parliamentary debates).

Bannard and Callison-Burch (2005) propose a pivotal approach where segments aligned with the same terms in the pivot language are considered potential paraphrases.

Example of German English corpus:

in check↔Unterkontrolle↔under control.

ONTOLOGY

In recent years, with the important evolution of the World Wide Web (WWW), the sources of information become more and more multiform (article, wiki, video, photo, library, etc.).

These sources of information are represented in forms useful to the users but difficult for automatic processing by a computer.

Indeed, several computer applications such as information retrieval , summarizing or machine translation, require an increasing development of tools able to manage the knowledge expressed in natural language.

ONTOLOGY

Such systems generally require intelligent processing of the **textual content** of the information sources available on the web.

The crucial problem to solve is that of the **polysemy of words**.

Many efforts have been made in this field, with the aim of enabling the machine to understand the information and to extract its meaning from the words, in order to facilitate their use in automatic processing.

As a result, the implementation of techniques and tools for automatic pre-processing of information sources becomes a necessity.

ONTOLOGY

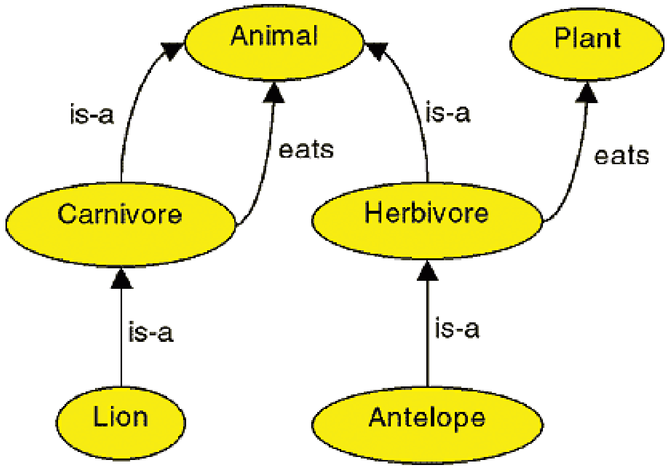
Ontologies are among the tools that allow **the semantic representation** of information sources in order to make them interpretable by machine.

In particular, ontologies are tools that allow to **represent** a corpus of knowledge in a form **usable by machine**.

They aim to provide **shared and common knowledge** on an domain to facilitate knowledge sharing and reuse.

This knowledge is represented as a structured set of **concepts** which are organized in the form of a graph whose **relations** can be semantic relations.

ONTOLOGY



ONTOLOGY

Concretely, in the context of NLP, the use of an ontology aims to **improve** the quality and generality of a system.

Indeed, they make it possible to obtain a representation of the text deeper, more abstract and independent of language.

MONOLINGUAL ONTOLOGY MAPPING

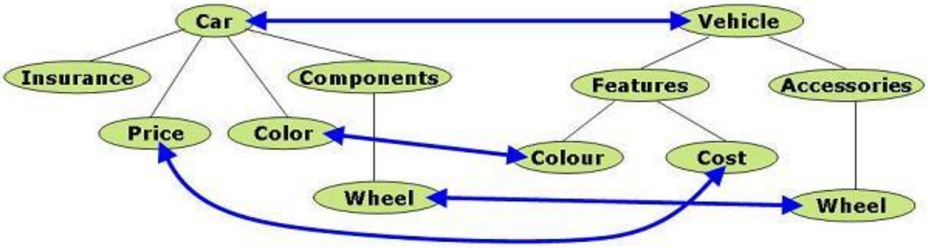
The **heterogeneity** issue occurs when ontologies are authored by different actors like database management problem, where database administrators use different terms to store the same information in different database systems.

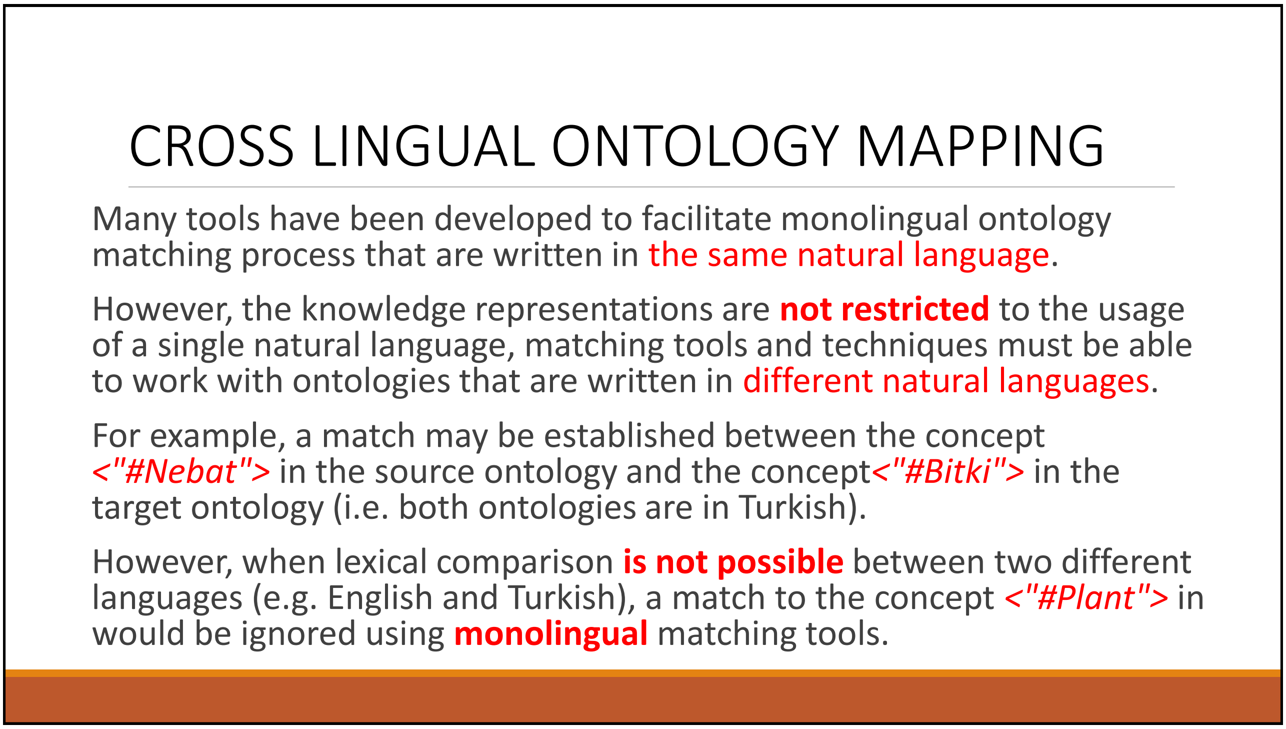
This means that the views on the same domains of interest will differ from one person to the next, depending on their conceptual model and background knowledge.

To address the heterogeneity issue arising from ontologies, ontology mapping has become an important research field.

MONOLINGUAL ONTOLOGY MAPPING

Ontology mapping is viewed as a two-step process, whereby the first step involves the generation of candidate correspondences (i.e. pre-evaluation) and the second step involves the generation of validated correspondences (i.e. post-evaluation).





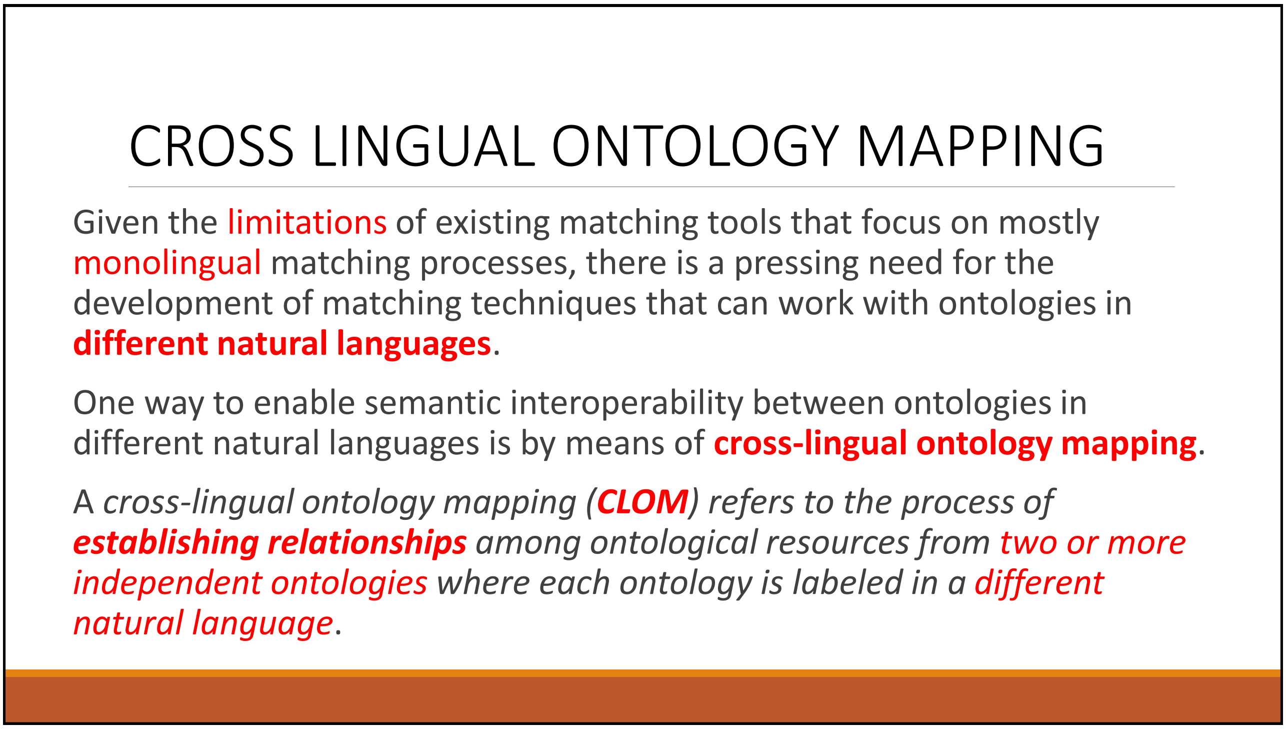
CROSS LINGUAL ONTOLOGY MAPPING

Many tools have been developed to facilitate monolingual ontology matching process that are written in the same natural language.

However, the knowledge representations are **not restricted** to the usage of a single natural language, matching tools and techniques must be able to work with ontologies that are written in different natural languages.

For example, a match may be established between the concept *<"#Nebat">* in the source ontology and the concept*<"#Bitki">* in the target ontology (i.e. both ontologies are in Turkish).

However, when lexical comparison **is not possible** between two different languages (e.g. English and Turkish), a match to the concept *<"#Plant">* in would be ignored using **monolingual** matching tools.



CROSS LINGUAL ONTOLOGY MAPPING

Given the limitations of existing matching tools that focus on mostly monolingual matching processes, there is a pressing need for the development of matching techniques that can work with ontologies in **different natural languages**.

One way to enable semantic interoperability between ontologies in different natural languages is by means of **cross-lingual ontology mapping**.

*A cross-lingual ontology mapping (**CLOM**) refers to the process of **establishing relationships** among ontological resources from two or more independent ontologies where each ontology is labeled in a different natural language.*

CATEGORIES OF CLOM APPROACHES

Current approaches to CLOM can be grouped into five categories:

- Manual CLOM
- Corpus-based CLOM
- CLOM via linguistic enrichment
- CLOM via indirect alignment
- Translation-based CLOM

MANUAL CLOM

Manual CLOM refers to those approaches that rely **only** on human experts whereby mappings are generated **by hand**.

An example of manual CLOM: an *English thesaurus*: AGROVOC (developed by the FAO containing a set of agricultural vocabularies) is mapped to a *Chinese thesaurus*: CAT (Chinese Agricultural Ontology, developed by the Chinese Academy of Agricultural Science) by hand.

The thesauri are assigned to groups of terminologists to generate mappings. These manually generated mappings are reviewed and stored.

The **advantage** of this approach is that the mappings generated are likely to be accurate and reliable. However, given large and complex ontologies, this can be a **time-consuming**.

CORPUS BASED CLOM

Corpus-based CLOM refers to those approaches that require the assistance of **bilingual corpora** when generating mappings.

Such an example is presented in [Ngai et al., 2002]. Ngai et al. use a bilingual corpus (newspaper) to align WordNet (in English) and HowNet (in Chinese).

The **advantage** of this approach is that the corpora don't need to be parallel, which makes the construction process easier.

However, a **disadvantage** of using corpora is that the construction could be a costly process for domain-specific ontologies.

CLOM VIA LINGUISTIC ENRICHMENT

Pazienza & Stellato [2005] developed an interface which allows to add synonyms (e.g. extracted from WordNet) during the ontology development.

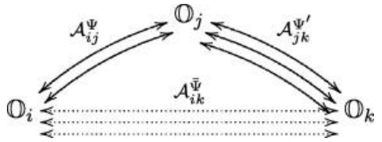
Linguistic enrichment of ontological resources will offer strong evidence in the process of mapping generation.

However, this enrichment process is currently un-standardized.

As a result, it can be difficult to build CLOM algorithms based upon these linguistically enriched ontologies.

CLOM VIA INDIRECT ALIGNMENT

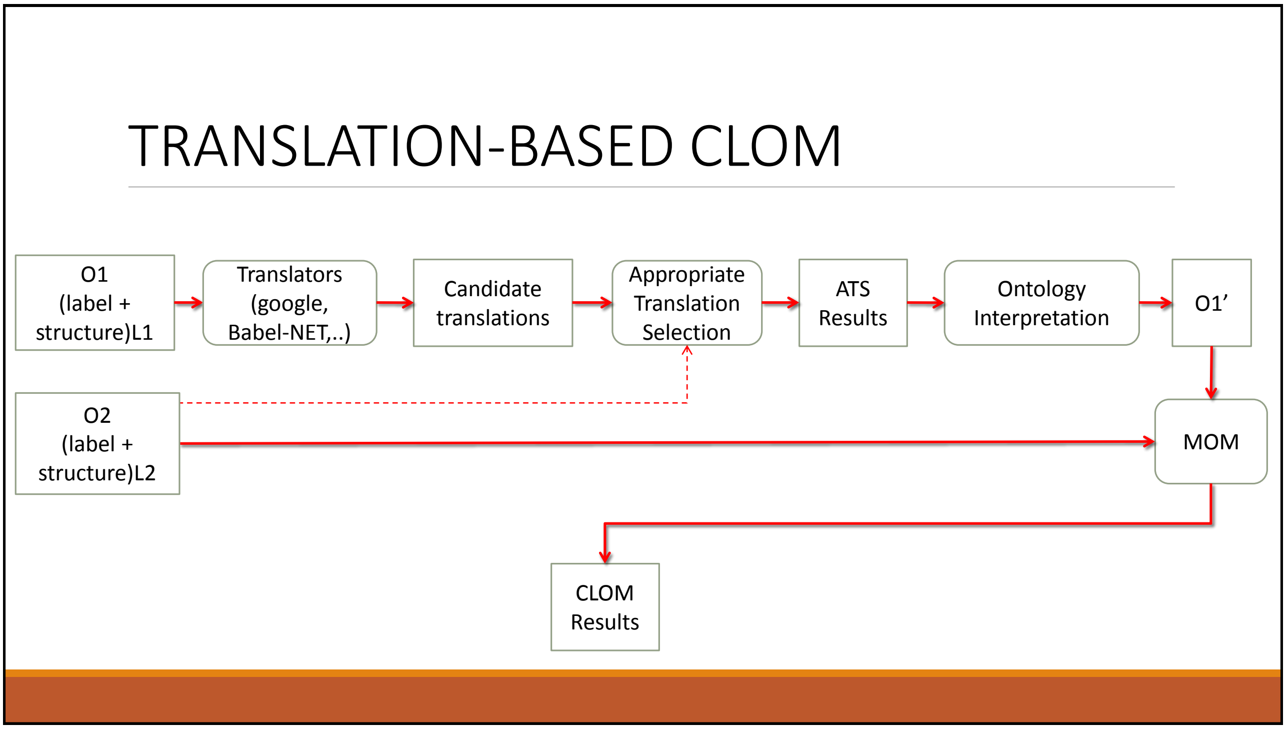
It refers to the process of **generating new CLOM** results **using pre-existing CLOM** results. Such an example [Jung et al., 2009]. They present indirect alignment among ontologies in English, Korean and Swedish, given alignment ***A*** which is generated between ontology **O_i** (e.g. in Korean) and **O_j** (e.g. in English), and alignment ***A'*** which is generated between ontology **O_j** and **O_k** (e.g. in Swedish). Then mappings between **O_i** and **O_k** can be generated by reusing alignment ***A*** and ***A'*** since they both concern one common ontology **O_j**.



TRANSLATION-BASED CLOM

Here the CLOM problem is converted to a MOM problem first, which is then solved using MOM techniques. It can be summarized as follows: given ontologies **O1** and O2 that are labeled in different natural languages, the labels of **O1** are first translated into the language used by O2. As both ontologies are now labeled in the same natural language, the mappings between them can then be created by simply applying monolingual ontology matching techniques.

The outcome of the mapping process is **conditioned** on the translations selected for the given ontology resources. In order to generate quality mapping results, translations must be **selected appropriately**.



TRANSLATION-BASED CLOM

O1
(label +
structure)L1

Translators
(google,
Babel-NET,..)

Candidate
translations

Appropriate
Translation
Selection

ATS
Results

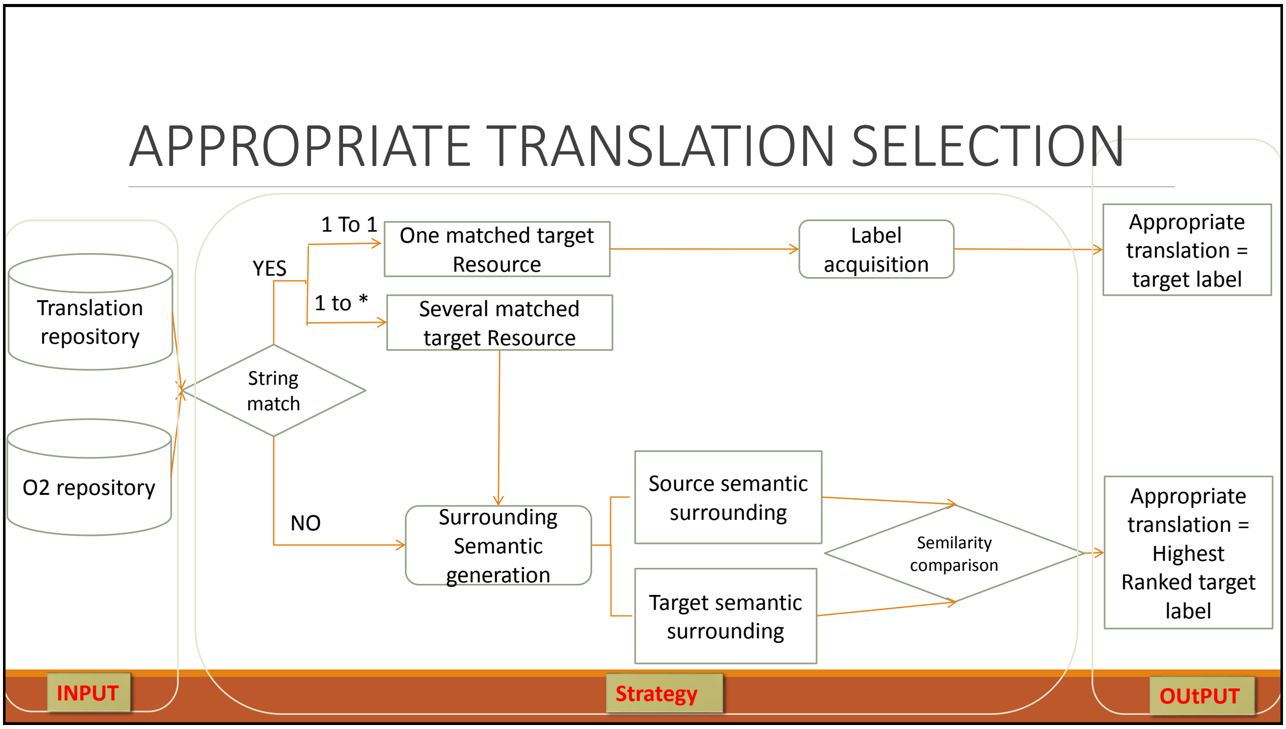
Ontology
Interpretation

O1’

O2
(label +
structure)L2

MOM

CLOM
Results



APPROPRIATE TRANSLATION SELECTION

Translation
repository

O2 repository

String
match

YES

NO

1 To 1

1 to *

One matched target
Resource

Several matched
target Resource

Label
acquisition

Surrounding
Semantic
generation

Source semantic
surrounding

Target semantic
surrounding

Semilarity
comparison

Appropriate
translation =
target label

Appropriate
translation =
Highest
Ranked target
label

INPUT

Strategy

OUtPUT