# Cross-Lingual Ontology Enrichment Based on Multi-Agent Architecture

Mohamed Ali, Said Fathallaa, Shimaa Ibrahim,
**Mohamed Kholief** and Yasser Hassan

SEMANTiCS18 - 14th International Conference on Semantic Systems
Vienna, 10 - 13 September 2018

# Motivation

- Most of the literature on ontology creation focuses on learning ontologies from **monolingual** data sources specially English language.
- Increasing the amount of **multilingual** data on the web and the consequent development and enrichment of ontologies in different natural languages.
- There are many existing well-formed ontologies in English language, however it is difficult to find those ontologies in other languages.
- It is necessary to develop innovative algorithms that are capable of digesting multilingual data, as well as enriching an ontology using a well-formed one in a different natural language (cross-lingual enrichment).

# Objective

Develop a technique to enrich existing ontologies by learning from multilingual sources and/or existing ontologies in other languages

# Proposed Approach

Cross-lingual ontology enrichment (**CLOE**) approach based on a multi-agent architecture in order to enrich ontologies from a multilingual text or ontology.

- Two novel algorithms: **simultaneous ontology enrichment** and **agents communication**.
- The most prominent features of **CLOE are:**
  - agents could learn from each other, using a predefined communication scheme, to get the benefit of already-learned concepts found in the input ontologies. In addition,
  - the enrichment of an ontology is performed using another ontology or text in different languages.

# CLOE Architecture

- **Input:** text that might be written in different natural languages and ontologies to be enriched.

  **Output:** enriched ontologies.

- CLOE Comprises three phases:

1. Text pre-processing and concept extraction,
2. Candidate sentence selection,
3. Multi-agent based enrichment.

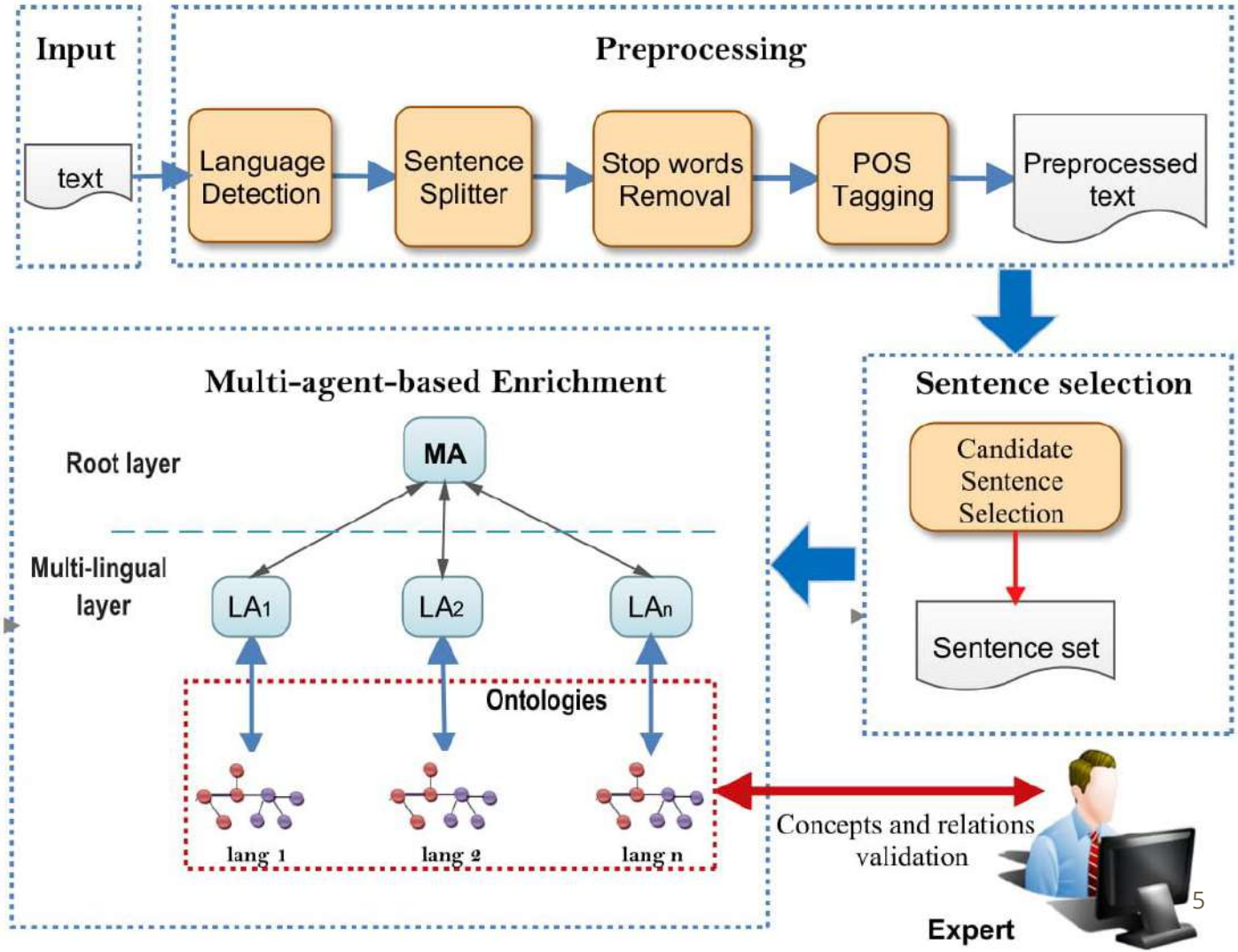
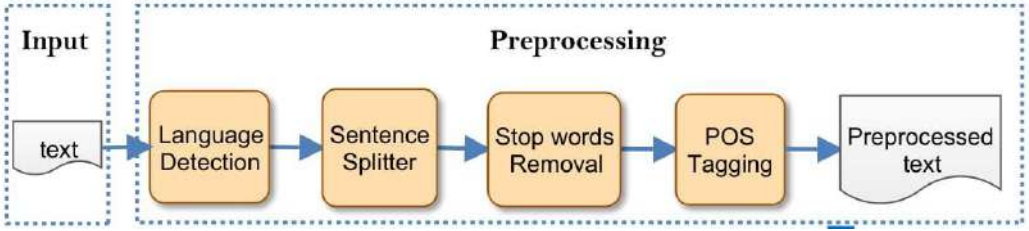


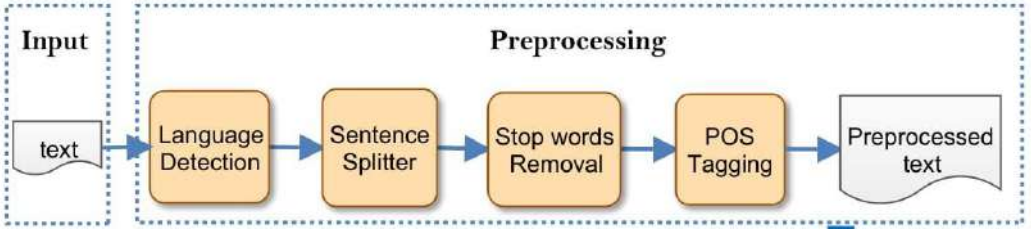

Figure 1: CLOE architecture

# Pre-processing




- Language identification:
  - Example: given an input text T = " مشغل الأقراص لا يعملyour drive may not be found"
    T will be divided into two sentences:
    - T1 = "مشغل الأقراص لا يعمل" with label L1 = Ar
    - T2 = "your drive may not be found" with label L2 = En
  - Depending on the language of the text, appropriate techniques are used for the next steps.
- Sentence splitter:
  - Splits the text into a list of sentences. We keep each sentence as a block of text in the further preprocessing tasks.

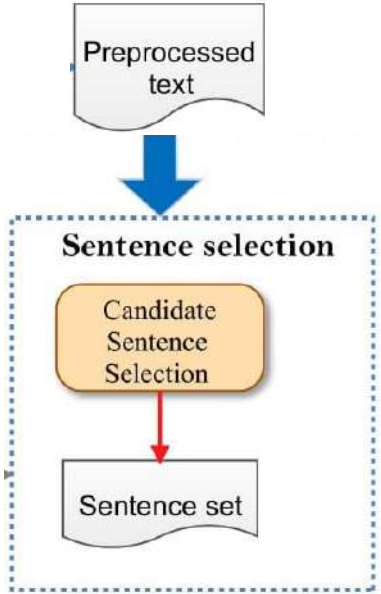
# Pre-processing




- Tokenization:
  - Divides each sentence to a set of tokens. Tokens separated by delimiters such as white-space characters.
- Stop words and unnecessary words filter:
  - Remove words with high frequencies of occurrence, but have no contribution to the subject of text such as pronouns, prepositions, conjunctions, and unnecessary words such as “very”, “really”, and non-alphanumeric contents.
- POS tagger:
  - Analyzes the text to find out phrase structures and group words according to their syntactic and semantic property. There are nine default tags: Adjective (ADJ), Adverb (ADV), Conjunction (CONJ), Determiner (DET), Noun (N), Number (NO), Preposition (PREP), Pronoun (PRO), and Verb (V). Token verbs are identified to represent a relation. At this stage, the identified verbs in each sentence are counted as a two-argument relation with nouns in the same sentence.

# Candidate Sentence Selection

- Filtering the pre-processed sentences in order to eliminate unnecessary/useless sentences.
- Iterate on sentences to count the number of nouns and noun phrases in a sentence.



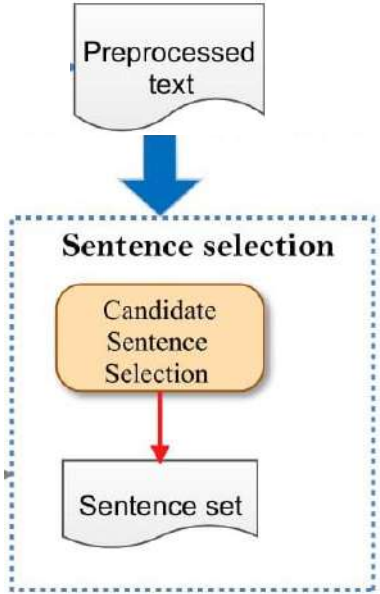
# Candidate Sentence Selection

- Candidate sentences selected according to the following rules:
  - **Rule 1:** Single-concept sentences are ignored.
    - **Example:** “Cydia has been installed successfully”.

      This sentence has only "Cydia" as a noun and does not has any useful information to be kept, therefore it will be ignored.

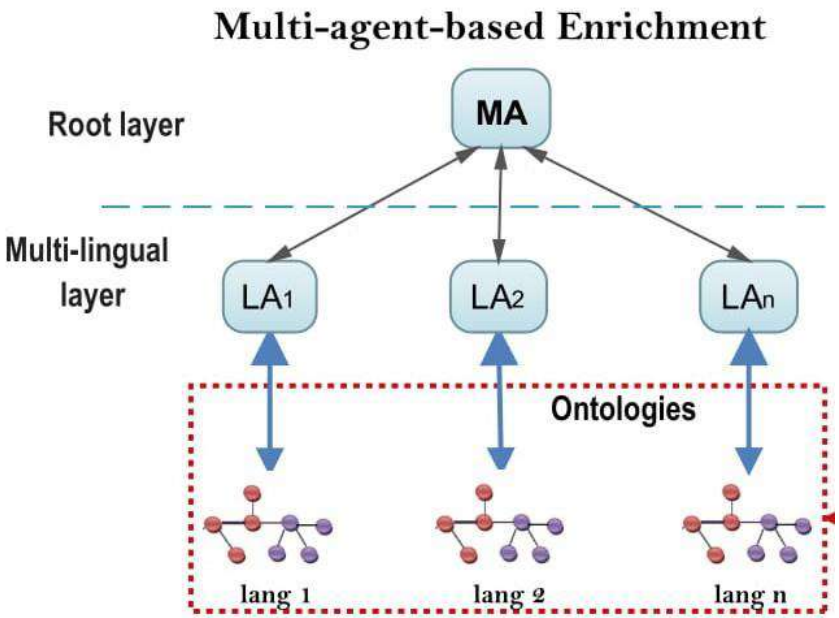
  - **Rule 2:** If the sentence has two POS or more as nouns or noun phrases, then keep the sentence, otherwise discard it.
    - **Example:** "Cydia has been installed successfully on the iOS device".

      This sentence has two nouns: "Cydia" and "iOS device", therefore it will be kept. This is required to extract the relations between concepts.



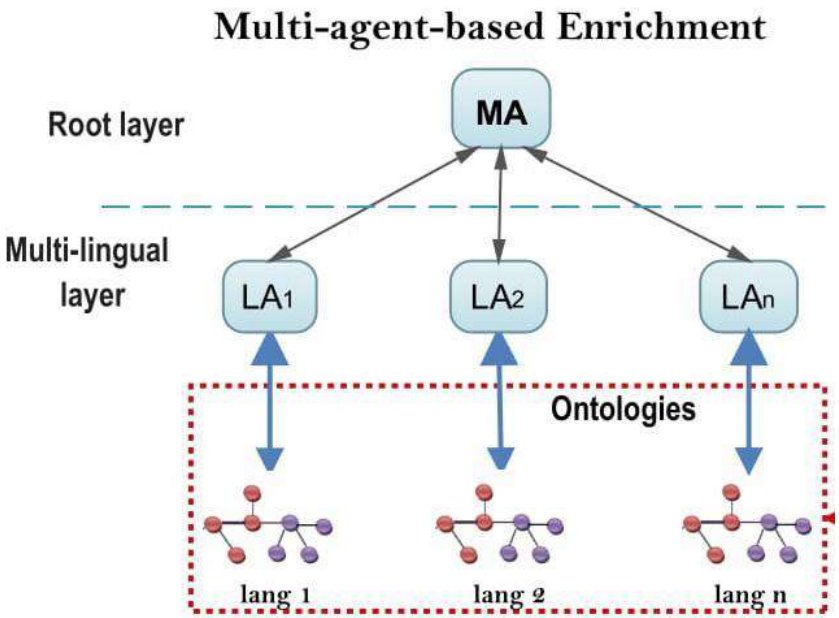
# Multi-agent Based Enrichment

- Two types of agents:
  - Master Agent (MA): responsible for managing a set of n language agents
  - Language Agent (LA): responsible for learning concepts, relations, and instances for a particular ontology language.



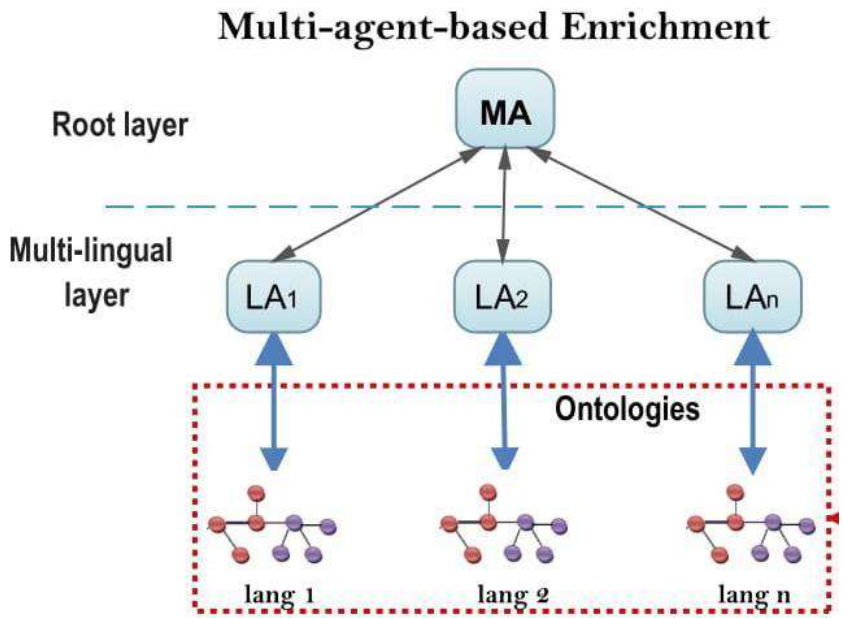
# Multi-agent Based Enrichment

- Agents Registration:
  - LAs register themselves at MA by sending the language of the associated ontology.
  - MA creates the look-up table LT and gives each one of them a unique id.
  - The look-up table consists of tuples of < id, Language> which is used by MA to decide which LA is responsible for learning the contents of each incoming text based on the language of the text.



# Multi-agent Based Enrichment

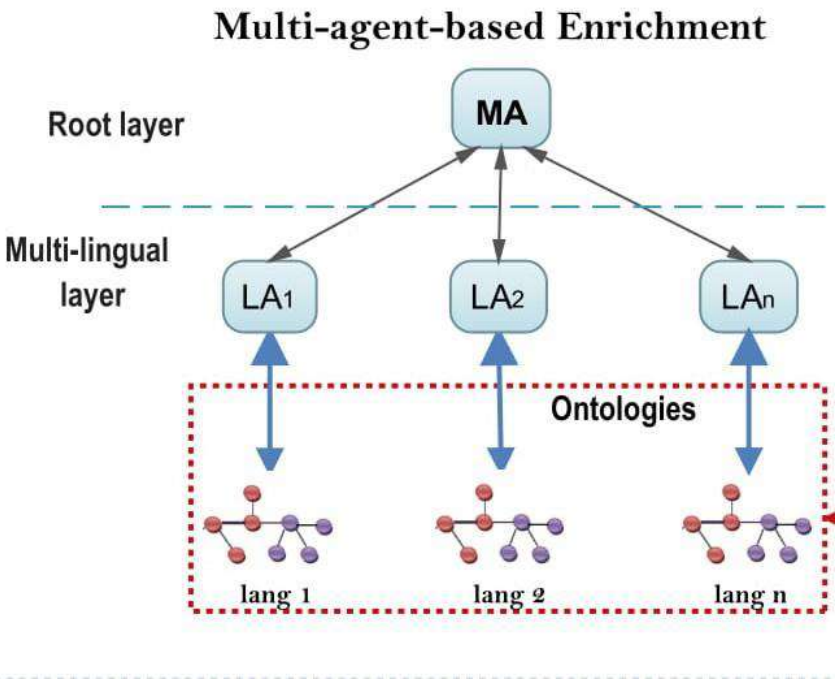
- Agents Registration:
  - For each incoming text, MA takes: candidate sentences S, text language L, and look-up table LT as input from the previous phase and then selects the appropriate LA based on L.
  - Each LA maintains a comprehensive translation table (TT ).
  - Each translation table lists the concepts that the agent knows and maps them to the corresponding concepts in the associated ontology.



# Multi-agent Based Enrichment

- Agents Registration:
  - English is the common language between all translation tables of all agents.

    Example:
    - Translation table of $LA_{Gr}$ translates concepts from German to English and vice-versa.
    - Translation table of $LA_{Ar}$ translates concepts from Arabic to English and vice-versa.
  - Only translations that are credible will be recorded in the translation table.



# Multi-agent Based Enrichment

- Agents Communication:
  - Using the Foundation for Intelligent Physical Agents (FIPA) Agent Communication Language (ACL).
  - ACL is based on speech act theory: messages are communicative acts sent by the sender agent to the receiver agents in order to perform some actions.

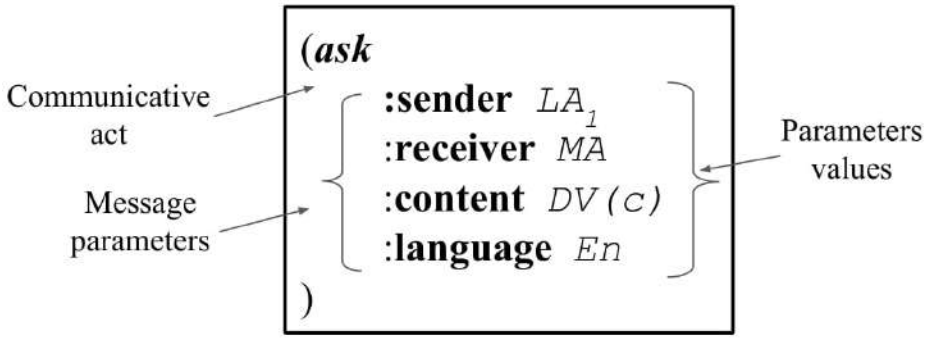



Figure 2: Components of the ACL message

# Multi-agent Based Enrichment

- Agents Communication:
  - Language agents can learn using :
    - Users can teach them by supplying a pre-built ontology and its translation table.
    - An agent can learn a new experience through its communications with its neighbors by querying another agent for a certain concept.
    - From input text.
  - There are two possible dialogues of communication between agents:
    1. Language agent asks the master agent about a certain concept. Then the master agent replies with the **DV** of this concept.
    2. Master agent asks other language agents, except the asking one, about the **DV** of the concept. If a positive reply is received, then the master agent asks the language agent to stream-out the available **n DV** of this concept.
  - With the completion of these actions, we have succeeded in achieving the goal, i.e. **simultaneous ontology enrichment**.

# Querying Another Agent for a Certain Concept

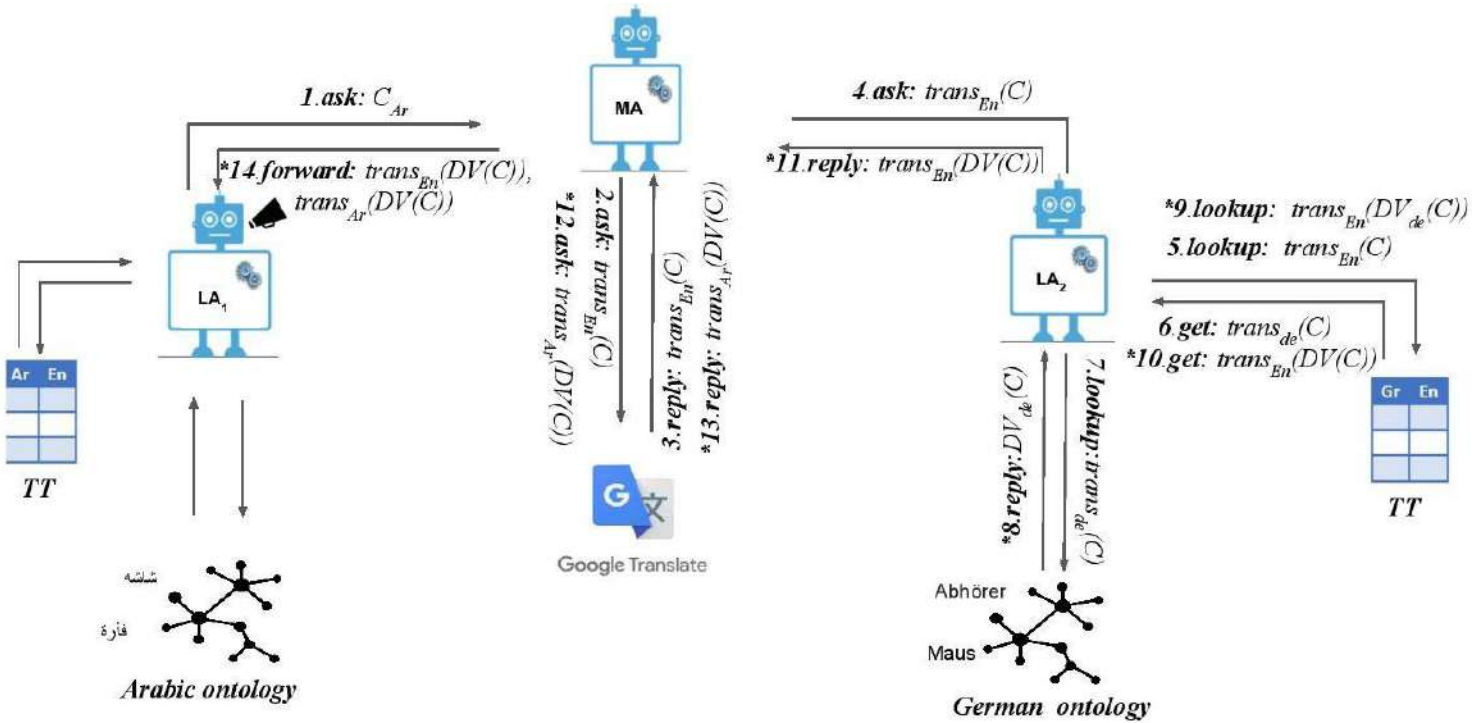



Figure 3: Agent communication actions when $LA_1$ wants to learn about a new incoming Arabic concept **C**. The asterisk "*" means that this step might repeat several times if there are more than one **DV** for a concept.

# A Case Study

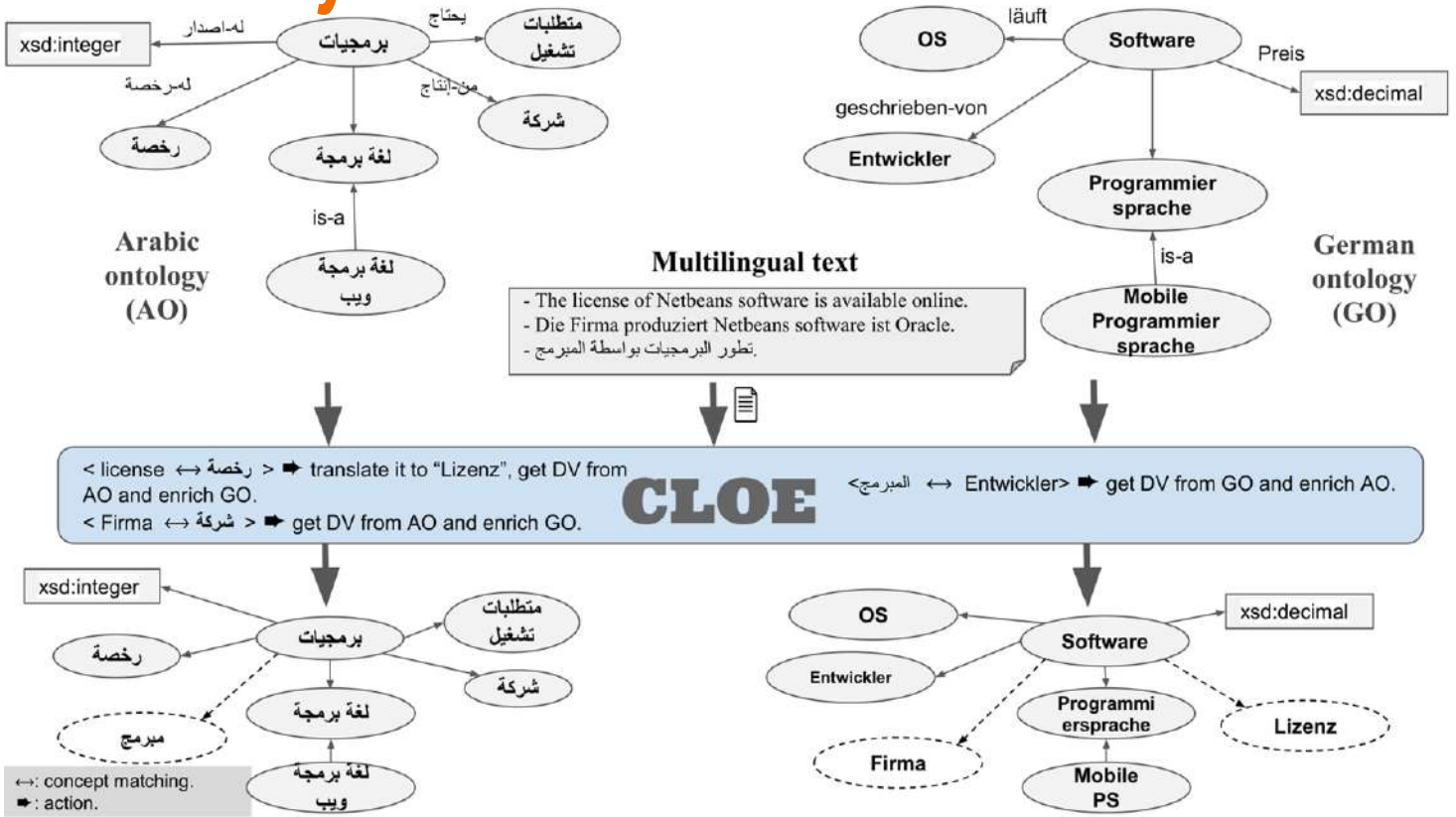



Figure 4: Small fragment from the enriched ontologies before and after submitting a multilingual text chunk containing English, German and Arabic text

# Evaluation

1. Satisfaction Questionnaire Evaluation
    - A total of 24 ontology engineers were recruited for this questionnaire which has 12 questions to assess the distinct phases of the proposed approach.
    - Observations from the satisfaction questionnaire:
        - 70.8% strongly agree that our approach would help ontology engineers.
        - 87.5% very hard to manually develop ontologies from a multilingual text, while only 4.2% responds with Neutral.
        - 62.5% believe that this approach is considered as a step towards a language-independent approach for ontology enrichment.
        - 79% satisfied with the completeness of the candidate relations and concepts, involving 33% of them are strongly satisfied.
        - the overall satisfaction is : 60.9% strongly satisfied, 21.7% satisfied and 17.4% Neutral.

# Evaluation

## 2. Gold standard-based evaluation

The objective of the gold standard-based evaluation is to determine whether the concepts/relations used in the learning process are correct after comparing them with the gold standard results.

- DataSet: 20 Newsgroups - We consider only the computers category (2936 documents)
- We have divided this category into two subsets. The first one is used for manually, by ontology experts, creating a reference standard ontology. Then, we have fed the system with this subset and saved the output. The second subset is used for evaluating CLOE.
- We randomly selected 100 documents from the second subset. Fifty documents out of the 100 were translated using Google API to Arabic and German to be able to submit multilingual text to the system.

# Evaluation

2. **Gold standard-based evaluation** - Comparison with the state-of-the-art systems.

| Metrics | CLOE | Benabdallah et al. [6] | Albukhitan and Helmy 2014 [2] | Albukhitan and Helmy 2016 [1] | Beseiso et al. [7] |
|---|---|---|---|---|---|
| LP | 0.90 | 0.90 | 0.95 | 0.76 | 0.89 |
| LR | **0.99** | 0.71 | 0.77 | 0.45 | 0.73 |
| F | **0.94** | 0.79 | 0.85 | 0.56 | 0.80 |


[1] Saeed Albukhitan and Tarek Helmy. 'Arabic Ontology Learning from Un-structured Text'. In: Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on. IEEE. 2016, pp. 492–496.
[2] Saeed Albukhitan and Tarek Helmy. 'Multi-agent Based System for Multilingual Ontologies Maintenance'. In: Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on. Vol. 1. IEEE. 2014, pp. 419–423.
[6] Ali Benabdallah, Mohammed AlaEddine Abderrahim and Mohammed El-Amine Abderrahim. 'Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology'. In: International Journal of Speech Technology (2017), pp. 1–8.
[7] Majdi Beseiso et al. 'Multilingual Ontology Learning Algorithm for Emails'. In: Soft Computing Applications and Intelligent Systems. Springer, 2013, pp. 229–244.

# Conclusion

- We present a novel multi-agent-based approach (CLOE) in order to enrich several ontologies (Simultaneous Ontology Enrichment) from multilingual data sources.
- The most prominent feature of the proposed approach is that agents could learn from each other, using a predefined communication scheme, to get the benefit of already-learned concepts found in the input ontologies.
- The main contribution of this work is the usage of a text segment to enrich several ontologies from different languages than the language of the input text.
- The results are satisfying compared to four state-of-the-art approaches.

# Future Work

- Adapting the proposed approach to be domain independent approach.
- Translate concepts based on language entries in DBpedia instead of Google Translation.
- Consider more techniques to generate multilingual ontologies from monolingual ones.

UNIVERSITÄT BONN

Smart Data Analytics

Arab Academy
For Science, Technology and Maritime Transport

ALEXANDRIA UNIVERSITY

MINISTRY OF HIGHER EDUCATION EGYPT

# Cross-Lingual Ontology Enrichment Based on Multi-Agent Architecture

Thank You!

Questions?!!!