$Multifarm_{11}$ - Extending the Multifarm Benchmark for Hindi Language

Abhisek Sharma[1,*], Sarika Jain[1] and Cassia Trojahn[2]

[1]National Institute of Technology Kurukshetra, India
[2]Toulouse University (IRIT), France

Abstract
Multifarm is a well-known comprehensive dataset for multilingual ontology matching evaluation.We extend the Multifarm dataset in the eleventh language, i.e., Hindi ($Multifarm_{11}$). This Hindi component of $Multifarm_{11}$[1] has been created by translating the entities using the Google translation service, validating manually, and then creating the reference alignments for the matching task. Work is in progress to determine the impact of $Multifarm_{11}$ on different multilingual ontology alignment systems. The complexities introduced by the Hindi language will introduce novel challenges to the behavior of cross-lingual ontology matching systems.

Keywords
Multifarm Benchmark, Hindi Dataset, Ontology Matching, Reference Alignment, Multilingual.

1. Introduction and Motivation

Multifarm [1] has been the commonly accepted benchmark dataset for Multilingual Ontology Matching since 2011 created on the basis of the Ontofarm dataset from the OAEI campaigns. It consists of seven ontologies of the conference domain. During its first inception, the Multifarm track was available in eight different languages other that English – Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish; later it was extended to include Arabic language in 2015 [2]. Introducing more languages that introduces more challenges and increases the scope of improvement of multilingual ontology matching systems is always beneficial.
Hindi is the fourth most spoken language in the world and is an indispensable part of the Indian identity and culture. It has its root in the mother of all languages, Sanskrit and Prakrit and is written using the Devanagari script. Hindi is known for its free order, multiple variants, and ambiguity in senses. The Hindi language offers different representation in both lexical and contextual sense and suffers from resource unavailability. All these mentioned aspects makes Hindi a worthy addition to the list of languages Multifarm is available in.
We present a proof of concept to extend the Multifarm dataset in the eleventh language, i.e.,

[1]Dataset archive can be found here - https://bit.ly/3QxkHRu

*Corresponding author.
abhisek_61900048@nitkkr.ac.in (A. Sharma); jasarika@nitkkr.ac.in (S. Jain); cassia.trojahn@irit.fr (C. Trojahn)
0000-0003-1568-2625 (A. Sharma); 0000-0002-7432-8506 (S. Jain); 0000-0003-2840-005X (C. Trojahn)

Hindi with an aim to include the same during the OAEI ontology matching workshops. After Chinese, Hindi has more characters than any other available language in the Multifarm track. For any Computer system to take contextually correct decisions and serve the vast population that speaks Hindi, we need more datasets in it that are helping one or more type of computer operations (in this case, Ontology matching). Many letters (and even words) in English has various mappings in Hindi. For example, 'S' can be written as श, ष or स. Hindi has no capitalization, though have short vowels. Version and variation of word in Hindi is context dependent. Same word in English (or any other languages) can have different Hindi associated word, depending on the context.

2. The $Multifarm_{11}$ Benchmark

The Hindi Language Component is developed referring to the multifarm dataset of the OAEI campaign. The structure of the ontologies and reference alignments has been reused while enriching them with the contextually verified entities in Hindi language. After including Hindi, a total of 55 language pairs will be there for the evaluation of the matching systems.

1. **Translation of Ontology Entities** - A total of around 2500 terms were fetched from the seven ontologies of the dataset and enlisted, out of which around 1000 are unique. Translations of the entities were done with the Google translation service.
2. **Validation** - Contextual verification was performed in line with conference domain. Errors like 'paper', which was translated to 'कागज़', which was contextually corrected to 'शोध पत्र. The task requires validators to have knowledge of conference domain and Hindi Language. The authors are well suited for the task, they all are aware of conference domain as they all are researchers and for Hindi, two of them are native Hindi speakers.
3. **Generation of Reference Alignments** - The reference alignments were created by reproducing the alignments based on the reference alignments available in the multifarm track. For example, dokument (of cmt ontology) in German is aligned to सम्मेलन दस्तावेज़ in Hindi (of conference ontology).

Acknowledgements

This work is supported by the IHUB-ANUBHUTI-IIITD FOUNDATION set up under the NM-ICPS scheme of the Department of Science and Technology, India

References

[1] Meilicke, C., Garcia-Castro, R., Freitas, F., Van Hage, W.R., Montiel-Ponsoda, E., De Azevedo, R.R., Stuckenschmidt, H., Šváb-Zamazal, O., Svátek, V., Tamilin, A. and Trojahn, C., 2012. MultiFarm: A benchmark for multilingual ontology matching. Journal of web semantics, 15, pp.62-68.

[2] Khiat, A., Benaissa, M. and Jiménez-Ruiz, E., 2015. ADOM: arabic dataset for evaluating arabic and cross-lingual ontology alignment systems. OM, 1545, pp.50-54.