# Bootstrapping Ontology Alignment Methods with APFEL

Marc Ehrig
Institute AIFB, University of Karlsruhe
Karlsruhe, Germany
ehrig@aifb.uni-karlsruhe.de

Steffen Staab
ISWeb, University of Koblenz-Landau
Koblenz, Germany
staab@uni-koblenz.de

York Sure
Institute AIFB, University of Karlsruhe
Karlsruhe, Germany
sure@aifb.uni-karlsruhe.de

## 1. INTRODUCTION

Semantic alignment between ontologies is a necessary precondition to establish interoperability between agents or services using different ontologies. Thus, in recent years different methods for automatic ontology alignment have been proposed to deal with this challenge. Thereby, the proposed methods were constricted to one of two different paradigms: Either, *(i)*, proposals would include a manually predefined automatic method for proposing alignments, which would be used in the actual alignment process [2, 4]. They typically consist of a number of substrategies such as finding similar labels. Or, *(ii)*, proposals would learn an automatic alignment method based on instance representations [1]. The first paradigm suffers from the problem that it is impossible, even for an expert knowledge engineer, to predict what strategy of aligning entities is most successful for a given pair of ontologies. Also, knowledge encoded in the intensional descriptions of concepts and relations is only marginally exploited by this way.

Hence, there remains the need to automatically combine multiple diverse and complementary alignment strategies of *all* indicators, i.e. extensional *and* intensional descriptions, in order to produce comprehensive, effective and efficient alignment methods. Such methods need to be flexible to cope with different strategies for various application scenarios. We call them "Parameterizable Alignment Methods" (PAM). We have developed a bootstrapping approach for acquiring the parameters that drive such a PAM through machine learning techniques. We call our approach APFEL for "Alignment Process Feature Estimation and Learning".

Our **main benefits**:

- a comprehensive process for ontology alignment with distinct steps
- an exemplary manual allocation of the parameters for each step
- APFEL assists the ontology engineer in optimizing this process by applying machine learning techniques to assign the parameters
- supporting the user in creating the training examples

## 2. ALIGNMENT PROCESS

Given two arbitrary ontologies $O_1$ and $O_2$, we try to find corresponding entities $E_1$ and $E_2$ with the same intended meanings in both ontologies, we try to align the two ontologies [3]. To achieve this we follow a well-defined process as shown in the upper part of Figure 1.

APFEL is based on the general observation that alignment methods like QOM [2] or PROMPT [4] may be mapped onto a generic alignment process:


*WWW2005*, May 10–14, 2005, Chiba, Japan.

1. Feature Engineering, i.e. select small excerpts of the overall ontology definition to describe a specific entity (e.g., the `label` to describe the concept `o1:Daimler`).
2. Search Step Selection, i.e. choose two entities from the two ontologies to compare (e.g., `o1:Daimler` and `o2:Mercedes`).
3. Similarity Assessment, i.e. indicate a similarity for a given description of two entities (e.g., $\text{simil}_{\text{superConcept}}$(`o1:Daimler`,`o2:Mercedes`)=1.0).
4. Similarity Aggregation, i.e. aggregate multiple similarity assessment for one pair of entities into a single measure (e.g., simil(`o1:Daimler`,`o2:Mercedes`)=0.5).
5. Interpretation, i.e. use all aggregated numbers, a threshold and interpretation strategy to propose the alignment (align(`o1:Daimler`)=$\{\emptyset\}$).
6. Iteration, i.e. as the similarity of one entity pair influences the similarity of neighboring entity pairs, the equality is propagated through the ontologies (e.g., it may lead to a new simil(`o1:Daimler`,`o2:Mercedes`)=0.85, subsequently resulting in align(`o1:Daimler`)=`o2:Mercedes`).

Each step requires specific parameters, which are normally engineered manually.

## 3. FEATURE ESTIMATION AND LEARNING

With APFEL (German for 'apple')however we present an approach tailored to assign these parameters automatically through machine learning techniques (see lower part of Figure 1).

**Generation of Feature/Similarity Hypotheses:**

The basis of the feature/similarity combinations is given by an arbitrary successful alignment method e.g. PAM(QOM).

Further, from two given ontologies APFEL extracts additional features by examining the ontologies for overlapping features, including domain-specific features. All features are combined in a combinatorial way with a generic set of predefined similarity assessments including similarity measures for, e.g., equality, string similarity, or set inclusion. Thus, APFEL derives similarity assessments for features. Some feature/similarity combinations will not be very useful, e.g. comparing whether one ID-number is a substring of another one. However, in the subsequent training step machine learning will be used to pick out those which improve alignment results.

From the feature/similarity combinations of PAM(QOM) and of the extracted hypotheses we derive an extended collection of feature/similarity combinations.

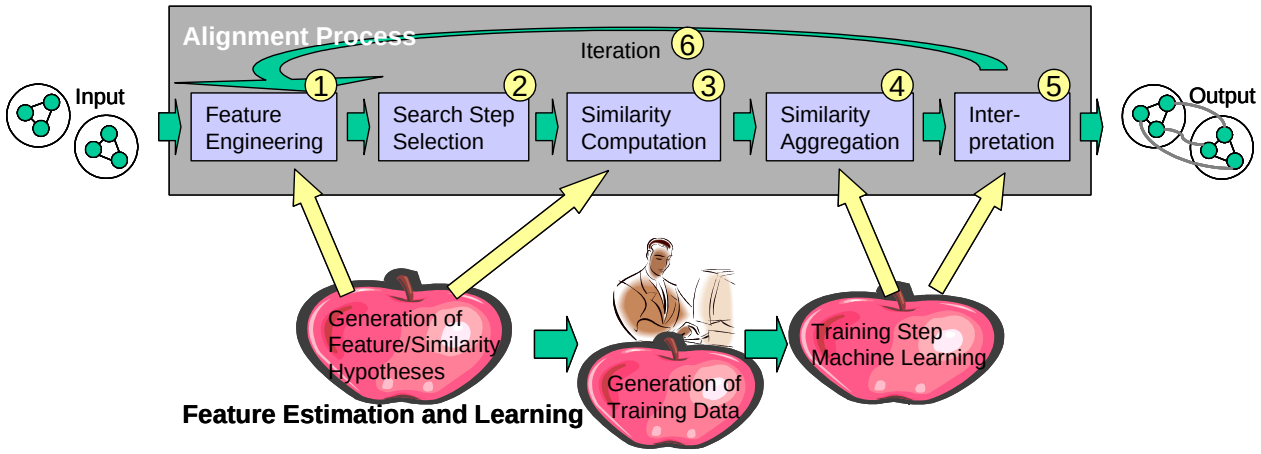


**Figure 1: Alignment Process Feature Estimation and Learning (APFEL)**

**Generation of Training Data:**

Machine learning as used in this paper requires training examples. The assistance in their creation is necessary as in a typical ontology alignment setting there are only a small number of really plausible alignments available compared to the large number of candidates, which might be possible a priori.

Therefore, we use an existing parametrization as input to the Parameterizable Alignment Method, e.g. PAM(QOM) to create the initial alignments for the given pair of ontologies. As these results are only preliminary, PAM does not have to use very sophisticated processes at this stage.

This allows the user to easily validate the initial alignments and thus generate correct training data. If the user further knows additional alignments he can add these alignments to the validated list.

Obviously the quality of the later machine learning step depends on the quality and quantity of these validated alignments.

**Training Step / Machine Learning:**

All validated alignment pairs are processed with the previously automatically generated collection of features and similarities. From each feature/similarity combination a numerical value is returned.

Based on these example training alignments we can now learn a classifier which distinguishes between those entities which align and those which are disjunct. Different machine learning techniques for classification (e.g. decision tree learner, neural networks, or support vector machines) assign an optimal internal weighting and threshold scheme for each of the different feature/similarity combinations. The machine learning methods like C4.5 capture relevance values for feature/similarity combinations. If the latter do not have any (or only marginal) relevance for the alignment, they are given a weight of zero.

From this we finally receive the most important feature/similarity combinations and the weighting and threshold thereof. With this we can set up the final ontology alignment method which we call PAM(APFEL). Depending on the complexity of the alignment problem it might be necessary to repeat the step of test data generation (based on the improved alignment method) and training.

## 4. CONCLUDING REMARKS

To investigate the effectiveness of APFEL, we have tested different strategies against each other. The decision tree learner returns results better than the other machine learning approaches, i.e. neural nets and support vector machines. The margin on improvement as compared to our baseline QOM is both times very good with around 7 percentage points. To sum up, APFEL generates an alignment method which is competitive with the latest existing ontology alignment methods. However, it is important to apply the correct machine learner and a sufficient amount of training data.

From all ontology alignment approaches GLUE[1] is closest to APFEL, but their learning is restricted on concept classifiers for instances based on instance descriptions, i.e. the content of web pages. From the learned classifiers they derive whether concepts in two schemas correspond to each other. Additional relaxation labeling is based solely on manually encoded predefined rules.

To conclude, with the complexity of the alignment task rising it becomes important to use automated solutions to optimize alignment approaches like PAM without losing the advantages of the general human understanding of ontologies. We contributed to this challenge with our approach APFEL. Effectively we received a process outperforming other state-of-the-art manually tailored alignment processes.

## 5. REFERENCES

[1] A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *VLDB Journal*, 50:279–301, 2003.

[2] M. Ehrig and S. Staab. QOM - quick ontology mapping. In F. van Harmelen, S. McIlraith, and D. Plexousakis, editors, *Proc. of the Third International Semantic Web Conference (ISWC2004)*, LNCS, pages 683–696, Hiroshima, Japan, 2004. Springer.

[3] M. Klein. Combining and relating ontologies: an analysis of problems and solutions. In A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, and M. Uschold, editors, *Workshop on Ontologies and Information Sharing, IJCAI01*, Seattle, USA, August 4-5 2001.

[4] N. F. Noy and M. A. Musen. The PROMPT suite: interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.