A Machine Learning Approach to Multilingual and Cross-lingual Ontology Matching

Dennis Spohr, Philipp Cimiano and Laura Hollink

CITEC, University of Bielefeld

Delft University of Technology

ISWC 2011, Bonn

Monnet is supported by the European Union under Grant No. 248458

Outline

Motivation

Background and definitions

Machine learning approach

Evaluation and findings

Conclusion

Open research questions in ontology matching

- Impact of machine translation in cross-lingual scenarios
- Impact of structural information
- Multilingual vs. cross-lingual ontology matching
- "Real" multilingual ontology matching understudied

Open research questions in ontology matching

- **Impact of machine translation in cross-lingual scenarios**
  - *"High-quality MT is prerequisite for good matching"* (Fu et al., 2009)
  - Can this be mitigated by translating to several languages?
- **Impact of structural information**
- **Multilingual vs. cross-lingual ontology matching**
- **"Real" multilingual ontology matching understudied**

Open research questions in ontology matching

- **Impact of machine translation in cross-lingual scenarios**
  - *"High-quality MT is prerequisite for good matching"* (Fu et al., 2009)
  - Can this be mitigated by translating to several languages?
- **Impact of structural information**
  - Importance already attested (Euzenat/Shvaiko, 2007)
  - Can a learning algorithm still learn a good matching function without structural information?
- **Multilingual vs. cross-lingual ontology matching**
- **"Real" multilingual ontology matching understudied**

Open research questions in ontology matching

- **Impact of machine translation in cross-lingual scenarios**
  - *"High-quality MT is prerequisite for good matching"* (Fu et al., 2009)
  - Can this be mitigated by translating to several languages?
- **Impact of structural information**
  - Importance already attested (Euzenat/Shvaiko, 2007)
  - Can a learning algorithm still learn a good matching function without structural information?
- **Multilingual vs. cross-lingual ontology matching**
  - Precise definitions do not exist yet
  - Inconsistent use in literature
- **"Real" multilingual ontology matching understudied**

Open research questions in ontology matching

- **Impact of machine translation in cross-lingual scenarios**
  - *"High-quality MT is prerequisite for good matching"* (Fu et al., 2009)
  - Can this be mitigated by translating to several languages?
- **Impact of structural information**
  - Importance already attested (Euzenat/Shvaiko, 2007)
  - Can a learning algorithm still learn a good matching function without structural information?
- **Multilingual vs. cross-lingual ontology matching**
  - Precise definitions do not exist yet
  - Inconsistent use in literature
- **"Real" multilingual ontology matching understudied**
  - Aggregation of similarity scores within a single language and across multiple languages?

Outline

Why accounting standards?

- Financial accounting standards define financial concepts and their relations

Why accounting standards?

- Financial accounting standards define financial concepts and their relations

- **Formal conceptualisation ≈ ontology**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards

- **Formal conceptualisation ≈ ontology**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages
- Different standards are often defined in *different* languages

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages
- Different standards are often defined in *different* languages

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**
- **Cross-linguality**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages
- Different standards are often defined in *different* languages
- Some manual mappings exist

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**
- **Cross-linguality**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages
- Different standards are often defined in *different* languages
- Some manual mappings exist

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**
- **Cross-linguality**
- **Machine learning possible**

Why accounting standards?

- Financial accounting standards define financial concepts and their relations
- Different jurisdictions typically follow different accounting standards
- Standards are often defined in *several* languages
- Different standards are often defined in *different* languages
- Some manual mappings exist

- **Formal conceptualisation ≈ ontology**
- **Conceptual mismatches**
- **Multilinguality**
- **Cross-linguality**
- **Machine learning possible**

⇒ **Ideal use case for multilingual and cross-lingual ontology matching**

XBRL Europe Business Registers Working Group

- **XBRL Europe Business Registers Working Group (XEBR)**
- Aims at achieving **interoperability of financial information**
- Creation of a **taxonomy of core financial concepts**
- **Manual matching** from national financial concepts to core concepts

XBRL Europe Business Registers Working Group

- **XBRL Europe Business Registers Working Group (XEBR)**
- Aims at achieving **interoperability of financial information**
- Creation of a **taxonomy of core financial concepts**
- **Manual matching** from national financial concepts to core concepts

- **XEBR uses the *eXtensible Business Reporting language (XBRL)***
- **The work presented here is based on an RDF conversion of XBRL**

Outline

Multi-lingual and cross-lingual ontology matching

Monolingual matching

…is the process of matching entities in S and T by comparing the labels in $S(l)$ and $T(l)$ in **a single language** $l \in L_S \cap L_T$.

Multi-lingual and cross-lingual ontology matching

Monolingual matching

…is the process of matching entities in S and T by comparing the labels in $S(l)$ and $T(l)$ in **a single language** $l \in L_S \cap L_T$.

Multilingual matching

…is the process of matching entities in S and T by comparing the labels in $S(l_i)$ and $T(l_i)$ in **at least two languages** $l_i \in L_S \cap L_T$**, with** $|L_S \cap L_T| \geq 2$.

Multi-lingual and cross-lingual ontology matching

Monolingual matching

…is the process of matching entities in S and T by comparing the labels in $S(l)$ and $T(l)$ in **a single language** $l \in L_S \cap L_T$.

Multilingual matching

…is the process of matching entities in S and T by comparing the labels in $S(l_i)$ and $T(l_i)$ in **at least two languages** $l_i \in L_S \cap L_T$**, with** $|L_S \cap L_T| \geq 2$.

Cross-lingual matching

…is the process of matching entities in S and T either

a. by **translating the labels in** $S(l)$ **to at least one language** $l' \in L_T$ and comparing the labels in $S(l')$ with those in $T(l')$, or

b. by **translating the labels in** $T(l)$ **to at least one language** $l' \in L_S$ and comparing the labels in $S(l')$ with those in $T(l')$, or

c. by **translating the labels** $S(l)$ **and the labels** $T(l')$ **to at least one language** $l'' \notin L_S \cup L_T$ comparing the labels in $S(l'')$ with those in $T(l'')$.

Intralingual and interlingual aggregation

Intralingual aggregation

…concerns the **aggregation of similarity scores** for all labels $a_i \in S(l) \cup T(l)$ **in a language** *l*

Intralingual and interlingual aggregation

Intralingual aggregation

…concerns the **aggregation of similarity scores** for all labels $a_i \in S(l) \cup T(l)$ **in a language** l

Interlingual aggregation

…concerns the **aggregation of intralingual scores across all languages** l_i **shared by** S **and** T

Outline

Motivation

Background and definitions

Machine learning approach

- Ranking SVMs
- Feature set definition
- Learning a ranking function with SVM^{rank}

Evaluation and findings

Conclusion

Ranking SVMs

- **SVMs for learning ranking functions**
- **Originally developed to improve search engine quality**
  by analysing clickthrough logs (Joachims, 2002)
- **Input not classes (e.g. *1 or -1*), but pairwise preference constraints**

Ranking SVMs

- **SVMs for learning ranking functions**
- **Originally developed to improve search engine quality**
  by analysing clickthrough logs (Joachims, 2002)
- **Input not classes (e.g. *1 or -1*), but pairwise preference constraints**
- Given 10 query results and knowing the user clicked on results 1, 3 and 5, the **input to the ranking SVM** is:

$$result_3 <_{r*} result_2 \qquad result_5 <_{r*} result_4$$
$$result_5 <_{r*} result_2$$

where $r*$ is the ranking preferred by the user

Applying ranking SVMs to ontology matching

Applied to ontology matching, this means:

We want to **learn a ranking function** which

- ranks ***matching concepts*** **higher than** ***non-matching concepts***
- ranks ***exact matches*** **higher than** ***close matches***

Feature set definition

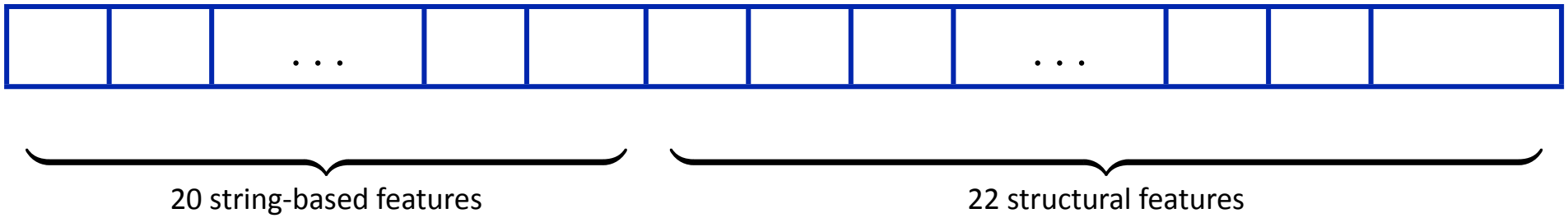
- **Definition of 20 different string-based features**
  - Five different measures (e.g. Levenshtein, Cosine)
  - Four different intralingual and interlingual aggregations (see next slide)

Feature set definition

- **Definition of 20 different string-based features**
  - Five different measures (e.g. Levenshtein, Cosine)
  - Four different intralingual and interlingual aggregations (see next slide)
- **Definition of 22 different (tailormade) structure-based features**
  - Concept type (e.g. monetary vs. abstract)
  - Similarity of calculations of monetary concepts
  - Similarity of position of a concept in a financial statement

Feature set definition

- **Definition of 20 different string-based features**
  - Five different measures (e.g. Levenshtein, Cosine)
  - Four different intralingual and interlingual aggregations (see next slide)
- **Definition of 22 different (tailormade) structure-based features**
  - Concept type (e.g. monetary vs. abstract)
  - Similarity of calculations of monetary concepts
  - Similarity of position of a concept in a financial statement
- **Construction of similarity vector for each pair $C_T \times C_S$**



Intra- and interlingual score aggregation

Intralingual aggregation

For example ***average intralingual similarity***

- $sim^{l}_{intra\sim}(C_S(l), C_T(l)) = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} sim_{lev}(lab^{i}_{C_S}, lab^{j}_{C_T})$

Intra- and interlingual score aggregation

Intralingual aggregation

For example ***average intralingual similarity***

- $sim^{l}_{intra\sim}(C_S(l), C_T(l)) = \frac{1}{n * m} \sum_{i=1}^{n} \sum_{j=1}^{m} sim_{lev}(lab^{i}_{C_S}, lab^{j}_{C_T})$

Interlingual aggregation

For example ***average of average intralingual similarity***

- $sim_{inter\sim/\sim}(C_S, C_T) = \frac{1}{n} \sum_{i=1}^{n} sim^{i}_{intra\sim}(C_S(i), C_T(i))$

Learning the matching function

- **Calculate similarity vectors** for all combinations of source and target concepts
- Use **exact and close matches** from XEBR to **determine rank** of a vector
- **Generate training file for SVM**^{rank}

Learning the matching function

- **Calculate similarity vectors** for all combinations of source and target concepts
- Use **exact and close matches** from XEBR to **determine rank** of a vector
- **Generate training file for SVM**^{rank}

```
3 qid:1 1:0.5155 2:0.6491 3:0.5757 ...  # C_Si × C_Tm
2 qid:1 1:0.4012 2:0.5660 3:0.4401 ...  # C_Si × C_T1
1 qid:1 1:0.2613 2:0.3333 3:0.2884 ...  # C_Si × C_T2
1 qid:1 1:0.2972 2:0.3492 3:0.0913 ...  # C_Si × C_T3
  ...
3 qid:2 1:0.2903 2:0.3548 3:0.5303 ...  # C_Sj × C_Tn
  ...
```

Outline

Motivation

Background and definitions

Machine learning approach

Evaluation and findings

Conclusion

Evaluation settings

- **Construction of training set:**

  Use XEBR mappings to derive mappings between national standards

C_{XEBR} *exactMatch* C_S	C_{XEBR} *exactMatch* C_S
C_{XEBR} *exactMatch* C_T	C_{XEBR} *closeMatch* C_T
———	———
C_S *exactMatch* C_T	C_S *closeMatch* C_T

  $\Rightarrow$ **Mappings between XEBR, German "Handelsgesetzbuch" (HGB) and Italian "Codice Civile" (ITCC)**

Evaluation settings

- **Construction of training set:**

  Use XEBR mappings to derive mappings between national standards

C_{XEBR} *exactMatch* C_S	C_{XEBR} *exactMatch* C_S
C_{XEBR} *exactMatch* C_T	C_{XEBR} *closeMatch* C_T
———	———
C_S *exactMatch* C_T	C_S *closeMatch* C_T

  $\Rightarrow$ **Mappings between XEBR, German "Handelsgesetzbuch" (HGB) and Italian "Codice Civile" (ITCC)**

- **Languages:**
  - XEBR: English
  - HGB: English, German
  - ITCC: English, French, German, Italian

Evaluation settings

- **Different scenarios**
  - **Monolingual:** English
  - **Multilingual:** English, German (for HGB × ITCC)
  - **Cross-lingual:** English × Italian, English × German, Italian × English
    - First translate ontology, then apply monolingual matching (cf. Fu et al., 2009)
    - Additional: *translate to several languages and apply multilingual matching*
  - **Transfer:** train on 2 pairs and test on third pair (*also multilingual*)

Evaluation settings

- **Different scenarios**
  - **Monolingual:** English
  - **Multilingual:** English, German (for HGB × ITCC)
  - **Cross-lingual:** English × Italian, English × German, Italian × English
    - First translate ontology, then apply monolingual matching (cf. Fu et al., 2009)
    - Additional: *translate to several languages and apply multilingual matching*
  - **Transfer:** train on 2 pairs and test on third pair (*also multilingual*)
- **Different settings**
  - With and without structural information (*Struct* vs. *NoStruct*)

Evaluation settings

- **Different scenarios**
  - **Monolingual:** English
  - **Multilingual:** English, German (for HGB × ITCC)
  - **Cross-lingual:** English × Italian, English × German, Italian × English
    - First translate ontology, then apply monolingual matching (cf. Fu et al., 2009)
    - Additional: *translate to several languages and apply multilingual matching*
  - **Transfer:** train on 2 pairs and test on third pair (*also multilingual*)
- **Different settings**
  - With and without structural information (*Struct* vs. *NoStruct*)
- **SVM^{rank} configuration**
  - Linear kernel with default configuration
  - No parameter optimisation

Evaluation settings

- **Different scenarios**
  - **Monolingual:** English
  - **Multilingual:** English, German (for HGB × ITCC)
  - **Cross-lingual:** English × Italian, English × German, Italian × English
    - First translate ontology, then apply monolingual matching (cf. Fu et al., 2009)
    - Additional: *translate to several languages and apply multilingual matching*
  - **Transfer:** train on 2 pairs and test on third pair (*also multilingual*)
- **Different settings**
  - With and without structural information (*Struct* vs. *NoStruct*)
- **SVM^{rank} configuration**
  - Linear kernel with default configuration
  - No parameter optimisation
- **Evaluation**
  - Precision: how often is exact match at rank 1
  - Further measure how often it is among top 5 and top 10

Evaluation results

Scenario	Setting	XEBR / ITCC			ITCC / HGB		
		1	5	10	1	5	10
Monolingual	$Struct_1$	**51.67**	**76.67**	**81.67**	**44.83**	**65.52**	**68.97**
	$NoStruct_1$	46.67	66.67	76.67	41.38	55.17	58.62
Multilingual	$Struct_n$	--	--	--	**51.72**	**68.97**	**68.97**
	$NoStruct_n$	--	--	--	51.72	55.17	65.52
Cross-lingual, *S* translated to one language	$Struct_1^S$	**35.00**	**56.67**	**63.33**	34.48	**65.52**	**68.97**
	$NoStruct_1^S$	28.33	53.33	53.33	**41.38**	51.72	55.17
Cross-lingual, *S* translated to several langs	$Struct_n^S$	**56.67**	**78.33**	**86.67**	44.83	**65.52**	**72.41**
	$NoStruct_n^S$	46.67	70.00	81.67	**48.28**	58.62	65.52
Cross-lingual, transfer	$Struct_n^{S_{tr}}$	**45.67**	**76.67**	**85.00**	**41.38**	**62.07**	**75.86**
	$NoStruct_n^{S_{tr}}$	23.33	60.00	73.33	31.03	58.62	65.52

Findings

- ***"With structure"* clearly outperforms *"without structure"***
  - 25 out of 27 values are significantly better

Findings

- ***"With structure"* clearly outperforms *"without structure"***
  - 25 out of 27 values are significantly better
- **Multilingual outperforms monolingual**
  - **Even in cross-lingual scenario!**
  - For XEBR × ITCC, cross-lingual produced best results

Findings

- ***"With structure"* clearly outperforms *"without structure"***
  - 25 out of 27 values are significantly better
- **Multilingual outperforms monolingual**
  - **Even in cross-lingual scenario!**
  - For XEBR × ITCC, cross-lingual produced best results
- **Transfer learning produces good results as well**

Conclusion

- Impact of machine translation in cross-lingual scenarios
- Impact of structural information
- Multilingual vs. cross-lingual ontology matching
- "Real" multilingual ontology matching understudied

Conclusion

- **Impact of machine translation in cross-lingual scenarios**
  - Translating to multiple languages can mitigate dependence on *"High-quality MT"*
- **Impact of structural information**
- **Multilingual vs. cross-lingual ontology matching**
- **"Real" multilingual ontology matching understudied**

Conclusion

- **Impact of machine translation in cross-lingual scenarios**
  - Translating to multiple languages can mitigate dependence on *"High-quality MT"*
- **Impact of structural information**
  - Further evidence that structural information is *very* important
- **Multilingual vs. cross-lingual ontology matching**
- **"Real" multilingual ontology matching understudied**

Conclusion

- **Impact of machine translation in cross-lingual scenarios**
  - Translating to multiple languages can mitigate dependence on *"High-quality MT"*
- **Impact of structural information**
  - Further evidence that structural information is *very* important
- **Multilingual vs. cross-lingual ontology matching**
  - We have given precise definitions of these notions
- **"Real" multilingual ontology matching understudied**

Conclusion

- **Impact of machine translation in cross-lingual scenarios**
  - Translating to multiple languages can mitigate dependence on *"High-quality MT"*
- **Impact of structural information**
  - Further evidence that structural information is *very* important
- **Multilingual vs. cross-lingual ontology matching**
  - We have given precise definitions of these notions
- **"Real" multilingual ontology matching understudied**
  - We have given precise definitions of *intralingual* and *interlingual* aggregation

monnet

Thank you very much for your attention!

Monnet is supported by
the European Union
under Grant No. 248458

COOPERATION