“Language and Domain Aware Lightweight Ontology Matching”

Talk

hasTitle

hasVenue

ISWC 2017

presents

Journal paper

hasTitle

hasPublisher

Journal of Web Semantics

hasAuthor

Gábor Bella

Fausto Giunchiglia

Fiona McNeill

hasSpeaker

hasWorkplace

University of Trento

Heriot-Watt University
Lightweight Ontology

• Formal representation of a hierarchy:
  – classifications,
  – subject headings,
  – catalogues, etc.;
• more common than formal ontologies;
• backbones of domain knowledge:
  – ICD (medicine),
  – NAICS (commerce),
  – UDC (library science), etc.

Use Case: Medical Research

High-precision medical research requires the integration of worldwide patient data:

– codes for diseases, interventions, etc.;
– need to map country-specific standards;
– some mappings exist but often partial or underspecified: ≡ / ≅ / ⊏

Language and Domain Aware Matching

- Need to match across languages;
- frequent use of domain terminology;
- some applications need high-precision mappings and precise mapping semantics (equivalence, subsumption, similarity, etc.).
State of the Art

Advantages:
- trivial to implement, no “awareness”;
- lots of languages supported by state-of-the-art MT;
- good-quality translations for major languages.

Disadvantages:
- good-quality MT (Google Translate) is not for free;
- MT is a black box, no adaptability or extensibility;
- much lower quality on “smaller” languages.
Goal

Compare the translation-based approach to a knowledge-based one where language and domain knowledge are built into the matcher and are extensible.
High-Level Architecture

Multilingual NLP

- labels
- label meanings
- lexical concepts

Formalisation → Matching

- axioms
- mappings

Background Knowledge
Background Knowledge

Multilingual NLP

Language Support

Domain Knowledge

Matcher

labels
label meanings

Language-independent concepts, relations

domain ontology

domain terms

Lexicon

axioms

Lexicon
1. Lightweight preprocessing: tokenisation, lemmatisation, optional lightweight parsing;
2. language-independent, domain-based word sense filtering.
   [Bella et al. Domain-Based Sense Disambiguation over Multilingual Structured Data, DIVERSITY Workshop at ECAI 2016.]

We currently have NLP support for EN, IT, ES, DE, AR, CN, MN.
Evaluation Setup

• **Datasets:**
  - Eurovoc (1,000 labels);
  - Universal Decimal Classification (600 labels);
  - English, Italian, Spanish in all combinations.

• **Testbed:** two versions of the SMATCH semantic matcher:
  - old: English-only SMATCH,
  - new: LDA SMATCH.
Evaluation Method

- Google or Apertium
- SMATCH
- Language and Domain Aware SMATCH
- Mappings
Evaluation Results

- **F-measure**: 70-80% for Eurovoc (short labels), 50-60% for UDC (longer labels).
- **Google Translate** is better by 0-12%, **Apertium** is worse by 13-15%.
- **Spanish-Italian** is the strongest pair, we benefit from the closeness of languages while MT is penalised by having to use English as pivot;
- With **complete lexical coverage** we approach the scores of Google and outperform it on Spanish-Italian.
Idea: take advantage of the strengths and weaknesses of each approach.
Combined Matching Results

- An extremely simple “OR” method that boosts recall+F with a slight detriment to precision.
- A consistent 5-10% increase in F-measure with respect to the best individual score (Google MT).
Conclusions

- We compared **two approaches** that reflect the current tensions in the AI community between formal and statistical methods;
- results of the formal method of language-aware matching are **slightly below but comparable** to those obtained using the best statistical MT;
- a **combined approach** can give an additional 10% win;
- **advantage** on less widespread languages and when none of the inputs are in English;
- **initial investment** needed for lexical and NLP support but lots of existing resources are available;
- in return, we get **full control & extensibility**.
Language and Domain Aware Lightweight Ontology Matching

Gábor Bella, Fausto Giunchiglia, and Fiona McNeill

Journal of Web Semantics, 2017
ISWC 2017
Uses of Matching: Emergency Response

Cross-border emergency response:
• a multi-domain scenario (geography, medicine, police, hydrology, etc.);
• may require cross-lingual matching;
• requires fast & accurate data exchange.

Analysis of Matching Mistakes

Typical limitations of our approach:

- lexical incompleteness;
- NLP mistakes;
- less robust w.r.t. approximative translations: *manufacturing vs attività manifatturiere*.

Typical limitations of machine translation:

- committing on wrong meanings: translation always resolves ambiguity, but WSD is hard;
- training anomalies: *psicotecnica* -> *psycho*
- cumulative mistakes in non-English language pairs, as English is used as pivot.
Label Formalisation

**Label Formalisation**

(Language Aware, Domain Independent)

- Language Detection
- Label Structure Formalisation
- Label Concept Formalisation

**Label Sense Filtering**

(Domain Aware, Language Independent)

---

Growing of plants and animals

Growing \( \cap (\text{Plant} \sqcup \text{Animal}) \)

Growing
of
plant and animal

“grow”
“cultivate”
“maturation”
“growth”

“industrial plant”
“plant as organism”
“floral”

“animal”
“beastly”

Scalable and extensible multilingual NLP (SCROLL)

Language-independent formula building

Language-independent concepts

Concepts of derived words
Label Sense Filtering

**Label Formalisation**

(Language Aware, Domain Independent)

- Language Detection
- Label Structure Formalisation
- Label Concept Formalisation

**Label Sense Filtering**

(Domain Aware, Language Independent)

- Domain Detection
- Domain-Based Filtering

---

**Growing \( \cap (\text{Plant} \cup \text{Animal}) \)**

- "grow"
- "cultivate"
- "maturation"
- "growth"
- "industrial plant"
- "plant as organism"
- "floral"
- "animal"
- "beastly"

1. agriculture
2. botany
3. zoology
4. industry
5. ...

Filtering of concepts that belong to less relevant domains

Ordering of domains according to ontology-level relevance
Language and Domain Aware Background Knowledge

DOMAINS
- industry
- sports
- science
- arts
- botany
- agriculture

CONCEPTS
- #12 industrial plant
- #45 plant as organism
- #890 book
- #567 cycling
- #34 bicycle
- #321 vegetable

SYNSETS
- plant (industrial)
- plant (flora)
- book (written work)
- cycling
- bicycle
- bicicletta
- libro
- pianta (organismo)
- ortaggio

ENSES
- related form

WORDS
- plant
- book
- cycling
- bicycle
- bicicletta
- libro
- pianta
- ortaggio

English lexical DB

Italian lexical DB
Language and Domain Aware Matching

Growing of *plants* – Allevamento di *piante*
Language and Domain Aware Matching

Cycling tourism – Turismo a bicicletta
Language and Domain Aware Matching

Cycling tourism – Turismo a bicicletta
Cross-Lingual String Similarity

• String matching is used for out-of-vocabulary words:
  – lexical incompleteness (frequent!);
  – name matching;
  – monolingual threshold: 80-90%.

• Can we match strings across languages? **Yes**!
  – up to 20% gain in recall while 0-5% loss of precision;
  – the lexically/orthographically closer the languages, the better it works:
    • Spanish-Italian: +18%,
    • English-Spanish: +7%,
    • English-Italian: +3%.
  – cross-lingual threshold: 60-70% (e.g., *tourism*—*turismo*: 71%), should be dependent on the language pair.