Boosting collaborative ontology building with key-concept extraction

Marco Rospocher

Fondazione Bruno Kessler (FBK) - Trento, Italy

DKM DATA & KNOWLEDGE MANAGEMENT

http://dkm.fbk.eu/rospocher

rospocher@fbk.eu

Joint work with:

Sara Tonelli, Emanuele Pianta, Luciano Serafini

IEEE ICSC 2011

Stanford, USA – September 19-21, 2011

Automatic Concept Extraction

- Support ontology modeling by **extracting concepts** characterizing a domain from a **reference text corpus.**

- Automatic concepts extraction plays an important role in ontology modeling:
  - To boost the ontology **construction/extension** phase;
  - To "**validate**" an ontology against a domain corpus.

FONDAZIONE BRUNO KESSLER

PESCaDO

Our Contribution

- A framework for supporting ontology **building/ validation** by automatic **concept extraction** from a reference text corpus

- A fully-working and publicly available **implementation** of the proposed framework

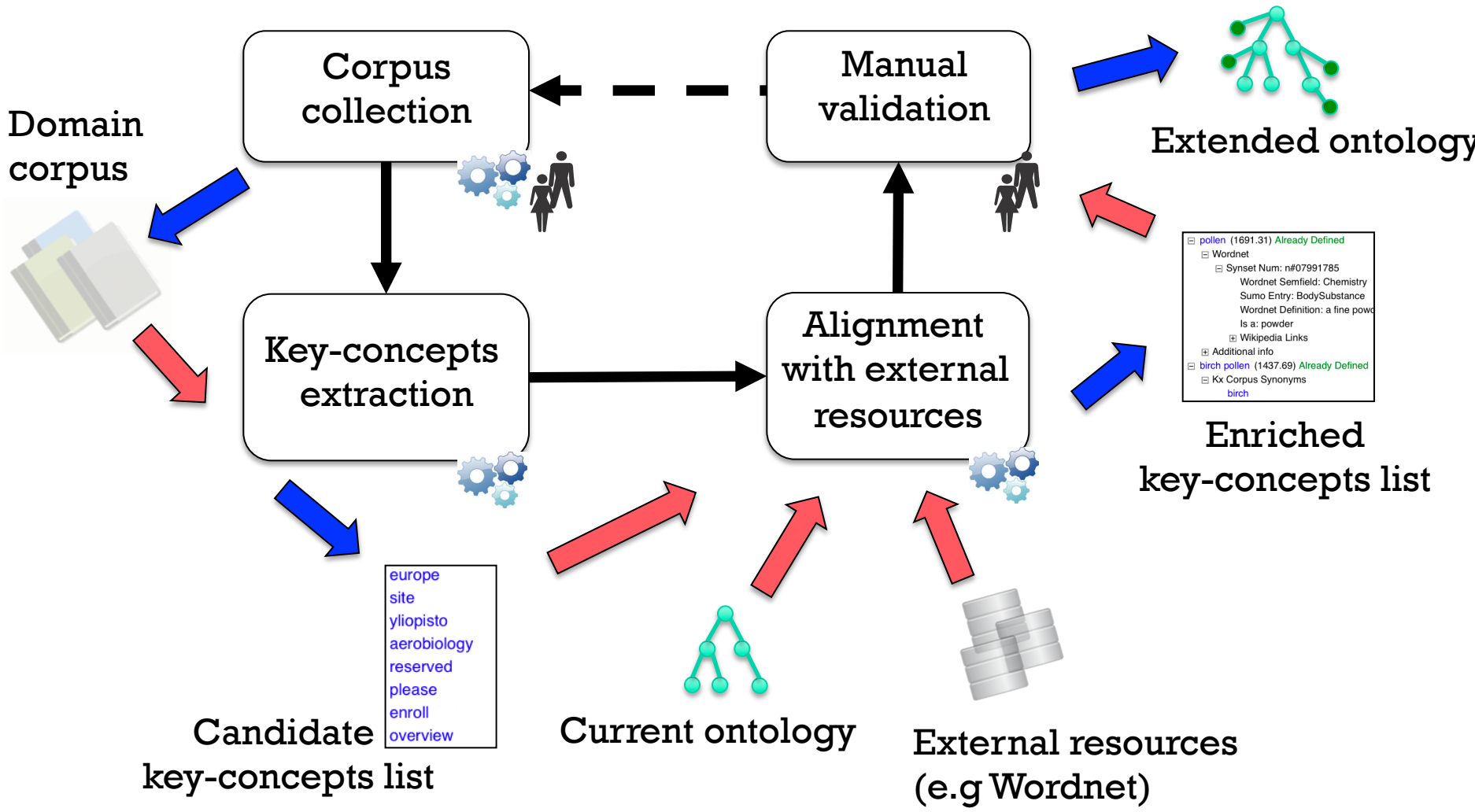
FONDAZIONE BRUNO KESSLER

PESCaDO

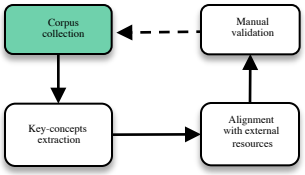
Outline

- The Framework
- Implementation of the Framework
- Evaluation
- Application Scenarios
- Concluding Remarks

The Framework



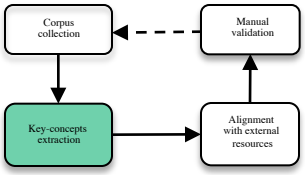
Corpus Selection



- The corpus can be **manually** or **automatically** selected (e.g. crawling web pages).

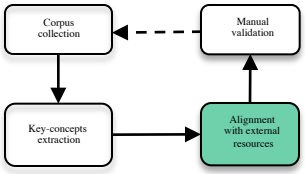
- Corpus could consist of:
  - (large) **collection** of documents
    - e.g. pollen bulletins crawled on-line
  - A **single** big document
    - e.g. the BPMN specification.

Key-concept extraction



- Performed by **KX** (**K**eyphrase e**X**traction) tool.
  - exploits **linguistic** information and **statistical** measures to select a list of **weighted keywords** from documents;
  - handles **multi-words**;
  - flexible **parameters** configuration;
  - easily adaptable to **new languages**;
  - ranked 2nd (out of 20) at SemEval2010, task on "*Automatic Keyphrase Extraction from Scientific Articles*".

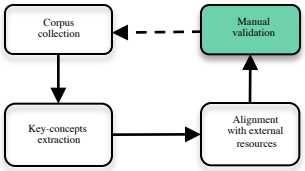
Alignment with external resources



- Extracted key-concepts **aligned** and **enriched** with additional resources:
  - **WordNet** (& WN domains): synonyms, definitions, SUMO labels;
  - **Wikipedia**: link to the Wikipedia page corresponding to the term (exploiting BabelNet);
  - Other external resources (e.g. dictionary).

- Enriched key-concepts list **matched against** the ontology under development (to detect already defined key-concepts).

Manual Validation



- The user **decides** which of the extracted key-concepts to add to the ontology;

- The additional details provided in the enriched list may **guide the formalization**;
  - e.g. is-a related synsets, definitions, …

FONDAZIONE BRUNO KESSLER

MoKi

the Modelling WiKi ---

PESCaDO

- **Collaborative** wiki-based tool for modeling (integrated) **ontologies** and **business processes**;
- Supports an agile collaboration between domain experts and knowledge engineers via **multi-mode knowledge access** modalities;
- Offers several different functionalities:
  - **Import/export** of formal models;
  - **Views** on the is-a hierarchy and processes decomposition;
  - **Graphical editing**.
- **Available @ http://moki.fbk.eu**

FONDAZIONE BRUNO KESSLER

MoKi

the Modelling WiKi ---

PESCaDO

DEMO

Evaluation

- Applied in **PESCaDO** (EU FP7 2010-2012) for building an ontology describing the environmental domain.

- Corpus: plain text corpus composed of **390 pollen bulletins** (541,000 tokens).

- The system outputted **91 key-concepts**:
  - **26 pollen names** (further validated against the Pollen Atlas);
  - **38** key-concepts **enriched** with additional information;
  - Extracted key-concepts having up **to 4 tokens**:
    - e.g. "oil seed rape pollen".

Application Scenarios

- The proposed approach can support several different ontology modeling tasks:
  - **Ontology construction boosting**: building an ontology from scratch;
  - **Ontology extension**: adding new concepts to an existing ontology;
  - **Ontology validation**: terminologically validating an ontology against a domain corpus;
  - **Ontology ranking**: ranking candidate ontologies wrt a given domain corpus;
  - **Ranking of ontology concepts**: determining which are the domain-wise most relevant concepts defined in an ontology.

Concluding Remarks

- We presented a framework for ontology **building**/ **validation** based on automatic concept extraction;
- **Fully-implemented** in a working system;
- Several **application** scenarios;
- Current/Future works:
  - Implementing specific support for ontology validation/ranking (e.g. computation of **ontology metrics**);
  - Extend for extraction of **structural information** (e.g. is-a relations defined in the corpus).

FONDAZIONE
BRUNO KESSLER

PESCaDO

Thank You!

Questions?

MoKi
http://moki.fbk.eu

Marco Rospocher
http://dkm.fbk.eu/rospocher
rospocher@fbk.eu