ADEL@OKE 2017: A Generic Method for Indexing Knowledge Bases for Entity Linking

Julien Plu, Giuseppe Rizzo, Raphaël Troncy

julien.plu@eurecom.fr

@julienplu

ADEL in a nutshell

- **ADaptable Entity Linking Framework**
  http://adel.eurecom.fr/api/

- **ADEL (in 2015): OKE Challenge winner**

- **ADEL (in 2016):**
  - Combine the results from multiple NER models
  - NIL clustering
  - New overlap resolution heuristic

What's new in ADEL (2017)?

- **Generic indexing process compliant with any indexing software (main focus of this talk)**
- **Type harmonization across data source (DBpedia, CoNLL, NEEL, Musicbrainz, etc.)**
- **Linking entities from video subtitles**
- **New co-reference extractor based on a deep neural network**
- **New linking method for French based on the JeuxDeMots lexical network**

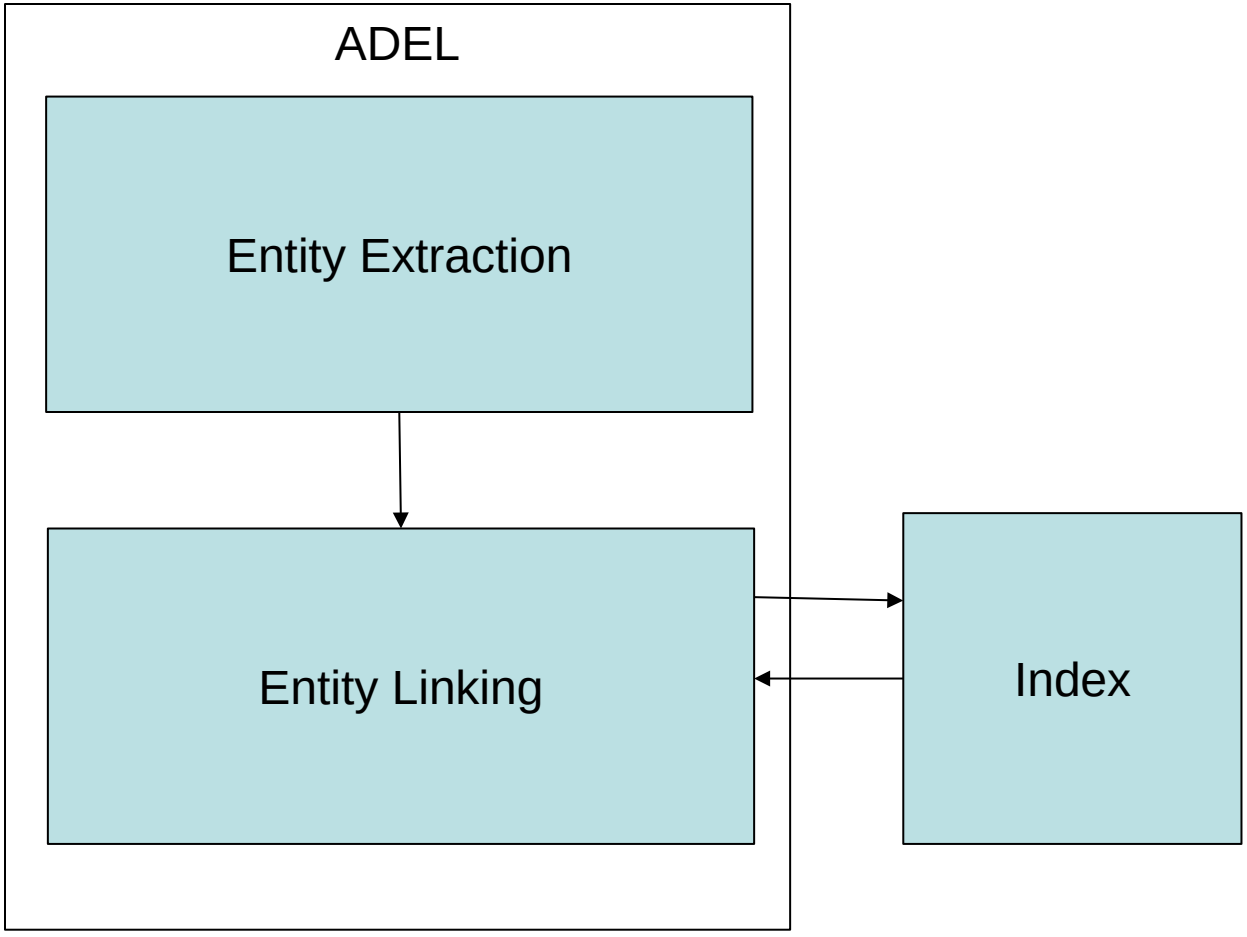
Different Approaches

E2E approaches:
A dictionary of mentions and links is built from a referent KB. A text is split in n-grams that are used to look up candidate links from the dictionary. A selection function is used to pick up the best match

Linguistic-based approaches:
A text is parsed by a NER classifier. Entity mentions are used to look up resources in a referent KB. A ranking function is used to select the best match

ADEL is a combination of both to make a hybrid approach

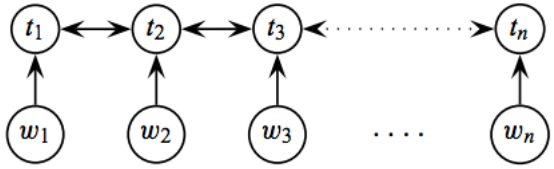
ADEL 30,000 foot view

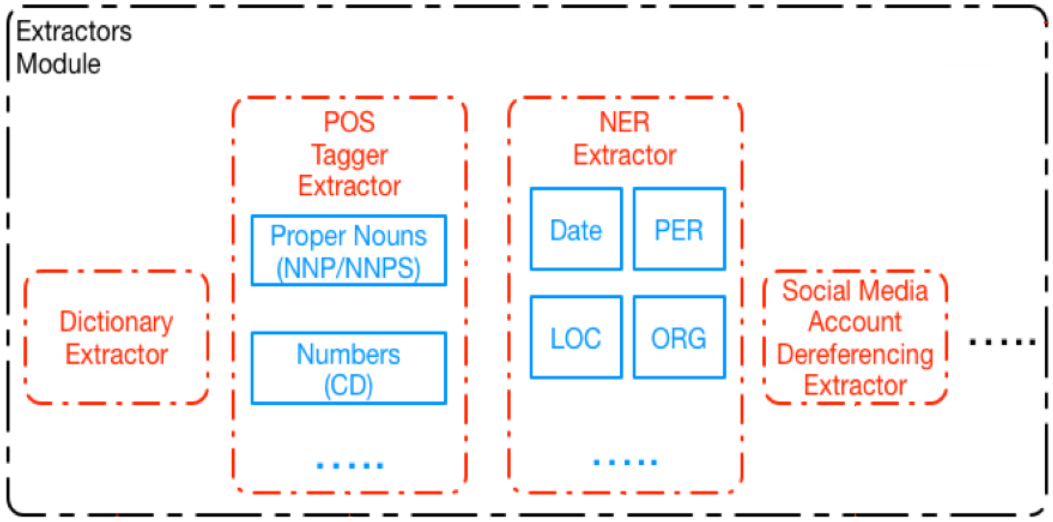


EURECOM

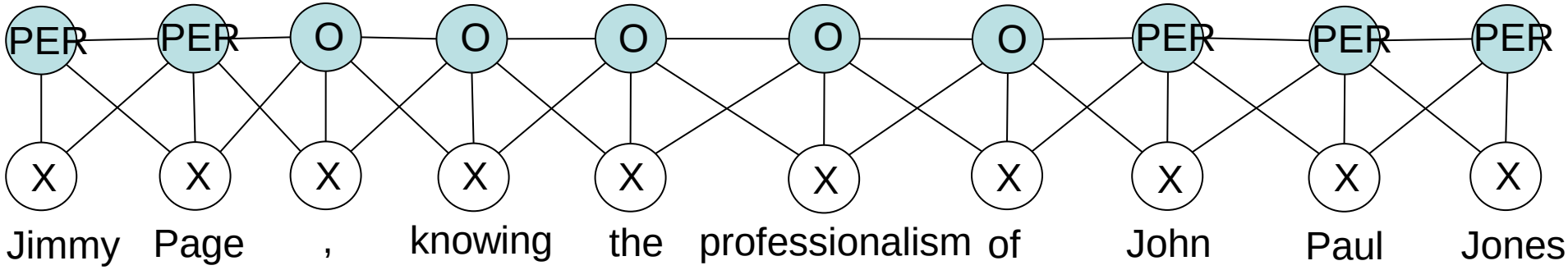
Entity Extraction: Extractors Module

- **POS Tagger:**
  - bidirectional CMM (left to right and right to left)





- **NER Combiner:**
  - Use a combination of CRF with Gibbs sampling (Monte Carlo as graph inference method) models. A simple CRF model could be:

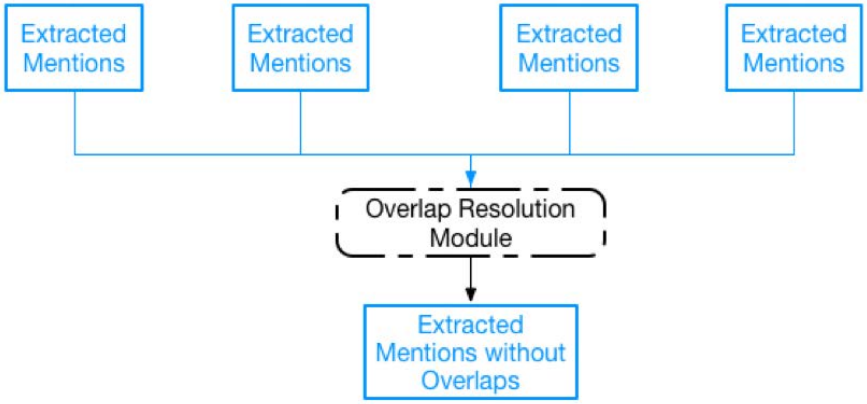


X set of features for the current word: word capitalized, previous word is “de”, next word is a NNP, … Suppose P(PER | X, PER, O, LOC) = P(PER | X, neighbors(PER)) then X with PER is a CRF

EURECOM

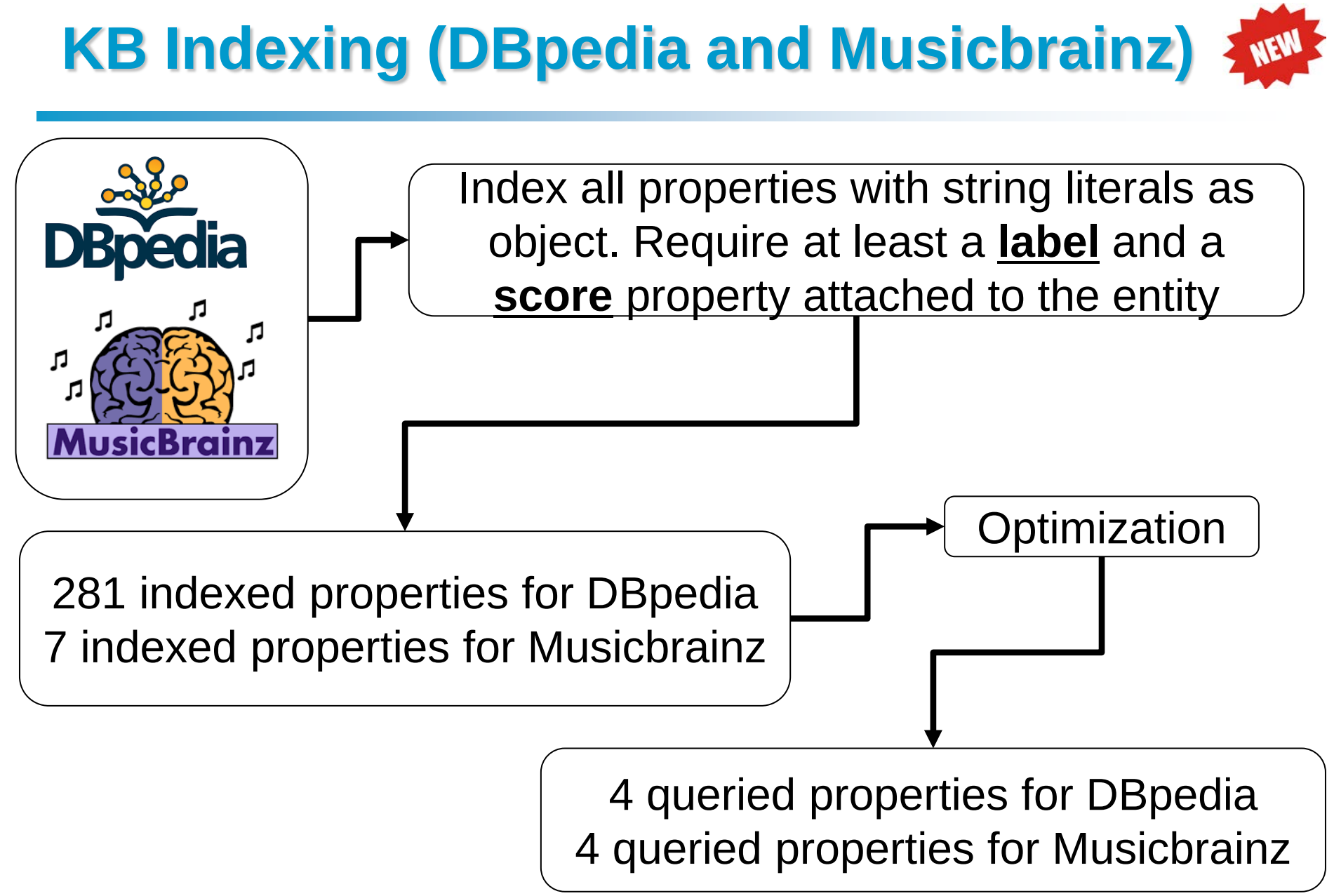
Entity Extraction: Overlap Resolution

- **Detect overlaps among boundaries of entities coming from the extractors**



- **Different heuristics can be applied:**
  - Merge: (“United States” and “States of America” => “United States of America”) **default behavior**
  - Simple Substring: (“Florence” and “Florence May Harding” => ”Florence” and “May Harding”)
  - Smart Substring: (”Giants of New York” and “New York” => “Giants” and “New York”)

KB Indexing (DBpedia and Musicbrainz)

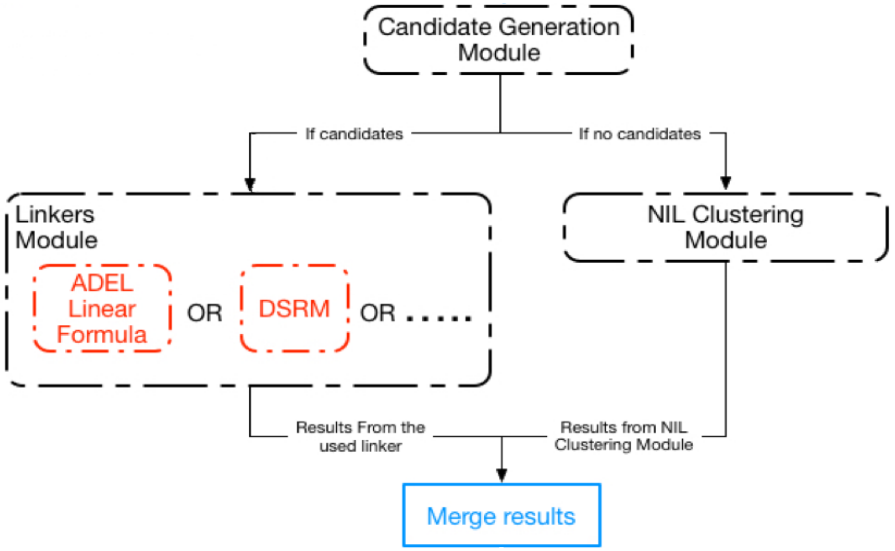


KB Indexing: Optimization

- **The optimization is done for a specific index**
- **3 steps process:**
  1. Create a dataset based on gold standard annotations to get a list of pairs: « mention, link »
  2. Query every mention against each property in the index:
     - ☞ If the proper link appears among the retrieved links, add the property in the list attached to the pair, otherwise, ignore the property
     - ☞ At the end, each pair has a list of successful properties attached
  3. Get an optimized list by following three merging heuristics:
     - ☞ a) A set of one property cannot be removed
     - ☞ b) If a set of size n has at least one intersection with another set of size m, knowing n > m, the set is removed
     - ☞ c) If n = m (in b) then one of the two sets is randomly removed

Entity Linking: Linking tasks

- **Generate candidate links for all extracted mentions:**
  - If any, they go to the linking method
  - If not, they are linked to NIL via NIL Clustering module
- **Linking method:**
  - Filter out candidates that have different types than the one given by NER
  - ADEL linear formula:



$$r(l) = \left(a.L(m, title) + b.\max\left(L(m, R)\right) + c.\max\left(L(m, D)\right)\right).PR(l)$$

r(l): the score of the candidate l
L: the Levenshtein distance
m: the extracted mention
title: the title of the candidate l
R: the set of redirect pages associated to the candidate l
D: the set of disambiguation pages associated to the candidate l
PR: Pagerank associated to the candidate l

a, b and c are weights following the properties:
$a > b > c$ and $a + b + c = 1$

Preliminary Results

- **Results over the OKE 2017 training set (4-fold cross-validation) … via nel-eval scorer**
- **Task 1: Focused NE identification & linking, 3 classes**

	Precision	Recall	F-measure
strong mention match	74.9	75.4	75.1
strong link match	47	54.5	**50.3**

- **Task 2: Broader NE identification & linking, 12 top classes**

	Precision	Recall	F-measure
strong mention match	70.7	67.9	69.2
strong link match	40.8	45.8	**42.9**

Lessons Learned

- **NER: combination of Stanford NER 3-classes, 4-classes, 7-classes trained on CoNLL 2003 + NER model trained on each OKE 2017 fold**
  - Dictionary extractor: +2% recall …BUT … -13% precision (e.g. *against* in sentence 22, famous bands!)
  - POS extractor: +5% recall … BUT … -36% precision (e.g. *World War II* in sentence 22)
  - Adding NER models from previous OKE years
  - Similar observations made on Task 2
- **NEL: drop in performance due to the current disambiguation formula (privileging entity popularity)**
- **KB indexing:**
  - Task 1: recall@1729 = 94.65% / Task 2: recall@1943 = 90.22%

Questions?

http://multimediasemantics.github.io/adel

http://jplu.github.io

julien.plu@eurecom.fr

@julienplu

http://www.slideshare.net/julienplu

EURECOM