THEY BUY FOR YOU

Multilingual Document Analysis

Carlos Badenes-Olmedo, Francisco Yedro, Oscar Corcho

Ontology Engineering Group at Universidad Politecnica de Madrid (OEG-UPM)

Spain

Motivation

Explore **large-scale multilingual** corpora to discover **similar** documents

- ***M1*** *(large-scale)**:***
  - *all pair-wise comparisons shouldn't be required (avoid O(N2))*
- ***M2*** *(multilingual):*
  - *translations shouldn't be required*
- ***M3*** *(similarity):*
  - *On-demand content-based ranking*

Plan

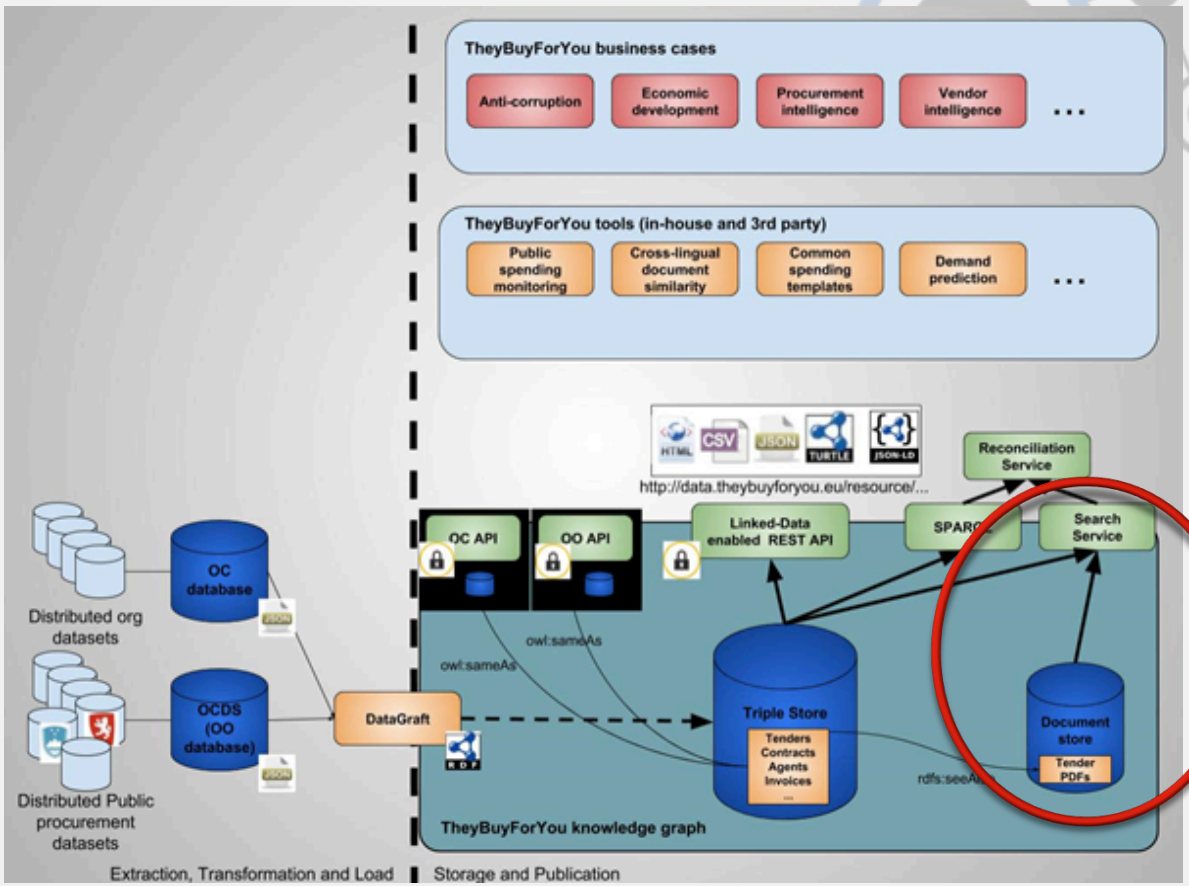
Design and Implement an ***efficient cross-lingual*** *textual* ***similarity*** *algorithm*

- ***P1*** *(efficient)**:***
  - *Approximate Nearest Neighbour (ANN) approach*
  - *low-dimensional representational space based on Probabilistic Topic Models (PTM)*
- ***P2*** *(cross-lingual):*
  - *unique simplex space for all documents*
  - *topic alignment between different language models*
- ***P3*** *(similarity):*
  - *Distance based on topic distributions*

Integration

On top of Document Store:

- ***As Annotator***
  - *Insert/Update Document Meta-Data*
- ***As Explorer***
  - ***Query Documents by Filter***



Build Corpora

- Parallel or Comparable Data based on multi-language labels
- Public Procurement or Legal Domain in European Union
- High data volume (more is better)

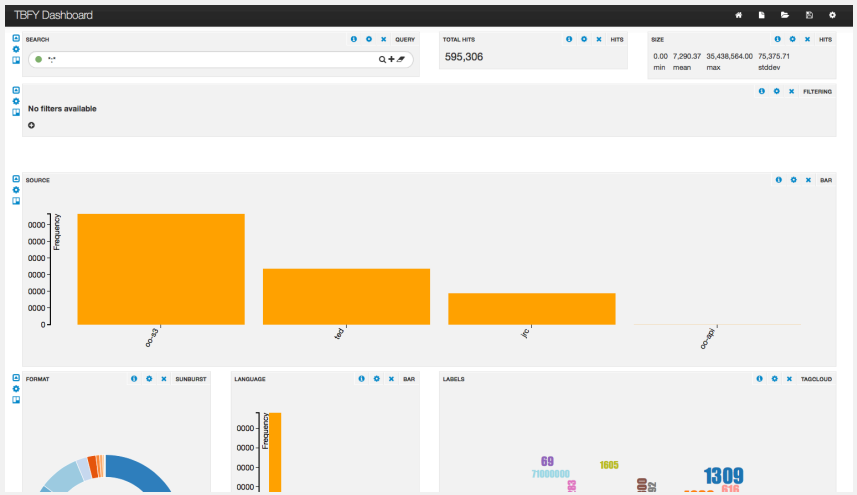
Build Corpora

Features	OO-API	OO-S3	TED	JRC-Acquis
Domain	Procurement Data	Procurement Data	Procurement Notices	Legislative Texts
Languages	EN	EN, ES, NL, FR..	EN,ES,DE,FR..	EN,ES,DE,FR..
Labels	Type	-	CPV	EuroVOC
Format	json	ZIP,Images,PDF, Docx,XSLX..	XML	XML
Size *(num docs)*	< 400	> 332,000	> 1M	> 90,000
Length *(chars mean)*	527	8,301	916	15,111

Build Corpora > Results

- **Corpus Repository / Dashboard**

  http://librairy.linkeddata.es/solr/banana



- **Harvester**

  http://github.com/TBFY/harvester

  Features:
  - *Amazon-S3 Collector*
  - *OO-API Collector (credential required)*
  - *TED Parser*
  - *JRC-Acquis Parser*
  - *EuroVoc Thesaurus Analyzer*

Train Models

- Automatic Identification of **stop-words** based on topic distribution
- Automatic identification of **stop-labels** based on word distribution
- Filter **words** by length (>2 chars), content (only alpha) and PoS (noun, proper_noun, verb and adjectives)
- Filter **texts** by length (>100 chars)
- Constraint LDA by defining a one-to-one correspondence between topics and labels (***Labeled LDA***)
- **Topic Independence Assumption**: only labels with no broader terms

Train Models > JRC > EuroVoc Categories

- A *multilingual* and *multidisciplinary* **thesaurus [1]** covering the activities of the EU, the European Parliament in particular. It contains terms in 23 EU languages.
- Vocabulary [2]:
  - 21 domain areas (*politics, international relations, european union, law, economics, trade, finance, social questions..*)
  - 7,193 concepts/labels
  - 4,904 reciprocal hierarchical relationships (no polyhierarchy)
  - 6,992 reciprocal associative relationships
- Document Labels:
  - Unbalanced Distribution (0~7):
    - e.g. Concept '1309' ("*import*"): 22.3%,
    - e.g. Concept '3330' ("*voluntary work*"): 0.05%
- Required tasks:
  - Re-Annotate documents *only* with ***root-labels***: iterate on broader concepts
  - Discard geo concepts (e.g. "Nordic Council Countries", "Andean Community Countries"..)
    - thesaurus_id=7216 : "America"
    - thesaurus_id=7226: "Asia and Oceania"

[1] - https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[2] - http://publications.europa.eu/resource/cellar/7eecbd11-c00d-11e5-9e54-01aa75ed71a1.0002.01/DOC_1

Train Models > Models-as-a-Service (MaaS)

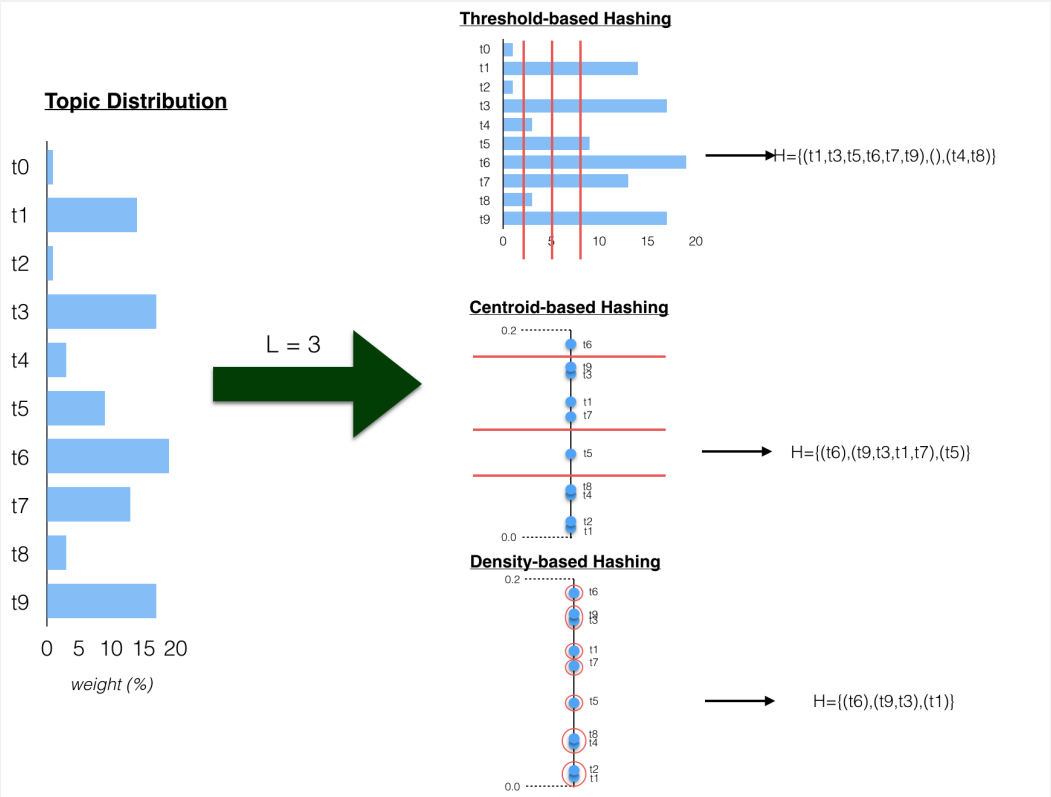
Features	OpenOpps API	TED (EN)	TED (ES)	JRC (EN)	JRC (ES)
Corpus	OpenOpps API	TED	TED	JRC-Acquis	JRC-Acquis
Size *(docs)*	272	9,768	7,144	20,000	19,997
Language	EN	EN	ES	EN	ES
Vocabulary *(words)*	815	10,048	7,510	35,796	54,806
Topics	177	1,612	1,070	453	470
Endpoint *(model-as-a-service)*	http://librairy.linkeddata.es/oo-api-en-model	http://librairy.linkeddata.es/ted-en-model	http://librairy.linkeddata.es/ted-es-model	http://librairy.linkeddata.es/jrc-en-model	http://librairy.linkeddata.es/jrc-es-model

THEY BUY FOR YOU

Annotate Documents

- **Topic-based Hierarchical Hash** expressions
- Approximate Nearest Neighbour (**ANN**) oriented
  - *Similar* documents *have larger probability of collision in feature space*
- Use of **Inverted Index** to avoid all pair-wise calculations

Annotate Documents > Topic-based Hashing Algorithm



*"**Efficient Exploration of Scientific Articles using Topic-based Hashing Algorithms**" – Badenes-Olmedo C., Redondo-Garcia J.L, Corcho O, 2019 (under review)*

Annotate Documents > Hashing Algorithm

- Documents annotated with a **Hierarchical** Expression:
  - ***topics0_t***: *primary topics*
  - ***topics1_t***: *secondary topics*
  - ***topics2_t***: *tertiary topics*
- Example:


```
{
        "id": "jrcC2006#261#70-en",
        "name_s": "Case F-81/06: Action brought on 21 July 2006 — Duyster v Commission",
        "topics0_t": ”5130",
        "topics1_t": "108",
        "topics2_t": ”5371",
        "labels_t": "1048 105 3979 4038 4271",
        ”root-labels_t": "5130 108 5371 1451 2166",
        "_version_": 1628471447846387712
}
```

Evaluate Mono-lingual Similarity

- **Unsupervised Classification** based on topics
  - Automatic generation of Eurovoc categories (labels)
- **Content-based Document Retrieval**

Evaluate Mono-Similarity > Unsupervised Classification

- **Gold-Standard**
  - Eurovoc categories (*i.e. root-concepts*) manually assigned to documents in JRC-Acquis corpus
- **Test-Set:**
  - 1k docs

Precision	JRC (EN)	JRC (ES)
Topics0_t	0.920	0.990
Topics1_t	0.647	0.670
Topics2_t	0.438	0.501

Recall	JRC (EN)	JRC (ES)
Topics0_t	0.189	0.211
Topics1_t	0.086	0.066
Topics2_t	0.028	0.024

Evaluate Mono-Similarity > Content-based Document Retrieval

- **Gold-Standard:**
  - Result documents from *MoreLikeThis* query based on root-concepts
- **Test-Set:**
  - 1k docs

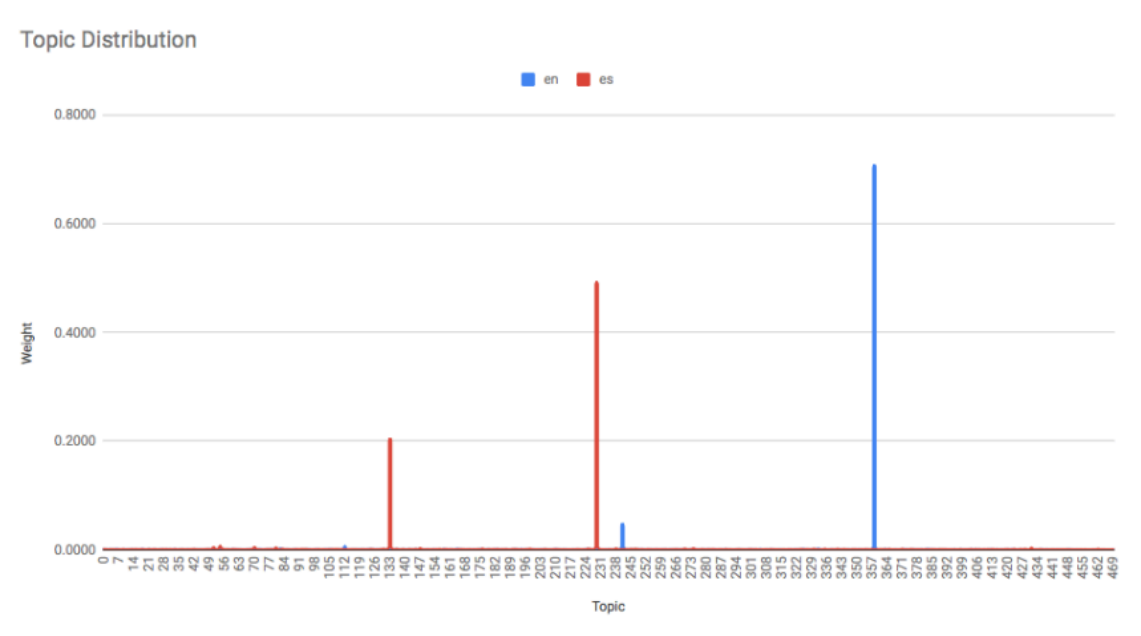
Measure	JRC (EN)	JRC (ES)
Precision	0.933	0.940
Recall	0.677	0.749
fMeasure	0.761	0.813

Align Multi-lingual Topics

- **Mapping** between topics from different language models

Align Multi-lingual Topics

- Based on topic names (~Eurovoc concepts)
  - e.g. **Concept '335'** - "*Educational Institution*"
    - Topic(EN): {id=241, **name=335**}
      - *partnership, stabilisation, western, balkan, priority, association, ..*
    - Topic(ES): {id=229, **name=335**}
      - *asociacion, estabilizacion, balcanes, occidentales, pacto, estabilidad..*
- Re-train models by only using ***shared*** topics:
  - (EN): -2258
  - (ES): -5974 -2117 -3689 -1039 -3232 -5776 -3013 -1191 -3257 -4587 -778 -2180 -317 -565 -2814 -3928 -1652 -3952



Topic Distributions of Documents:
- **jrc32004D0928-en** {t0=3062, t1=335, t2=8278}
- **jrc32004D0928-es** {t0=335, t1=4060, t2=5769}

Align Multi-lingual Topics > Results

Features	JRC (EN)	JRC (ES)
Corpus	JRC-Acquis	JRC-Acquis
Size *(docs)*	19,996	19,999
Language	EN	ES
Vocabulary *(num words)*	35,792	54,809
Topics	452	452
Endpoint	http://librairy.linkeddata.es/jrc-en-model	http://librairy.linkeddata.es/jrc-es-model

Evaluate Multi-lingual Similarity

- **Retrieval** of **cross-lingual** content-based related documents

Evaluate Multi-Similarity > Documents

English JRC	Feature	*Spanish* JRC
jrcC2006#261#66-en	***ID***	jrcC2006#261#66-es
Designation of the judge to replace the President of the Civil Service Tribunal for the purpose of dealing with applications for interim measures	***title***	*Designación del Juez que sustituirá al Presidente del Tribunal de la Función Pública en calidad de Juez de medidas provisionales*
1163, 1451, 5130	***root-labels***	1163, 1451, 5130
695 (“*election*”)	***topics0***	1163 (*“personnel administration”*)
1163 (*“personnel administration”*)	***topics1***	1538 *(“ruling”)*
1538 *(“ruling”)*	***topics2***	c_964c9649 *(“BRICS countries”)*

Evaluate Multi-Similarity > Results

- **Gold-Standard:**
  - Result documents from *MoreLikeThis* query based on root-concepts in **both languages**
- **Test-Set:**
  - 1k docs

Measure	JRC (EN,ES)
Precision	0.938
Recall	0.700
fMeasure	0.778

THEY BUY FOR YOU

Retrieve Similar Documents - DEMO

- *Tender and Contract Data* from **OpenOpps** (EN)
- *Procurement Notices* from **TED** corpora (EN,ES)
- *Legislative Texts* from **JRC-Acquis** corpora (EN,ES)
- *63,900 documents* in Solr Collection:
  http://librairy.linkeddata.es/solr/tbfy/select?q=*:*

Explore Similar Documents > By Text

[POST] http://librairy.linkeddata.es/api/items


```
{
        "dataSource": {
                "dataFields": {
                        "id": "id",
                        "name": "name_s"
                },
                "filter":"source_s:ted",
                "format": "SOLR_CORE",
                "url": "http://librairy.linkeddata.es/solr/tbfy"
        },
        "reference": {
                "text": {
                        "model": "http://librairy.linkeddata.es/jrc-en-model",
                        "content": "The Decarbonisation Project is in support of the Scottish Government's Energy Efficient Scotland Programme 2017-18. The project is to utilise existing Solar PV panels to 52 Council domestic properties to reduce tenant services bills, reduce carbon emissions and support Fife Council in meeting its EESSH 2020 and EESSH2 obligations (refer to Tenderers Submission Part 2, Property Information, Appendix 1). The tender includes for the supply, installation, testing and commissioning of the battery storage systems and associated work in domestic properties, which involves connecting to the existing Solar PV panels. The project is to utilise existing Solar PV panels to 52 Council domestic properties to reduce tenant services bills, reduce carbon emissions and support Fife Council in meeting its EESSH 2020 and EESSH2 obligations (refer to Tenderers Submission Part 2, Property Information, Appendix 1). The tender includes for the supply, installation, testing and commissioning of the battery storage systems and associated work in domestic properties, which involves connecting to the existing Solar PV panels."
                }
        },
    "size": 10
}
```

Explore Similar Documents > By ID

[POST] http://librairy.linkeddata.es/api/items

```
{
        "dataSource": {
                "dataFields": {
                        "id": "id",
                        "name": "name_s"
                },
                "filter":"source_s:ted",
                "format": "SOLR_CORE",
                "url": "http://librairy.linkeddata.es/solr/tbfy"
        },
        "reference": {
                "document": {
                        "id": "19-044187-001-en"
                }
        },
    "size": 10
}
```

Next Steps

- Add *Mean Average Precision* (MAP) to evaluation metrics
- Evaluate similarities based on TED corpora
- Merge annotations in only one meta-data field
- Deploy a Search-API using this annotations

Thanks!

TheyBuyForYou.eu has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780247