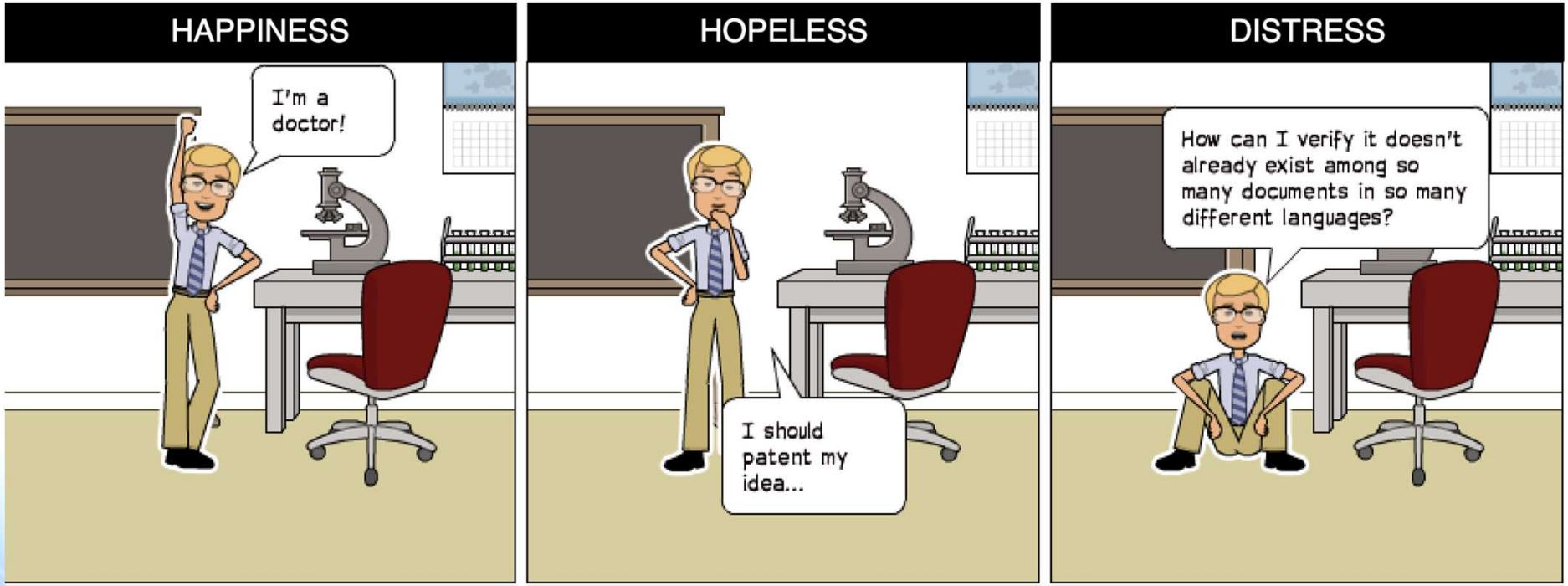
Plan TL
IBERLEGAL

Legal Document Retrieval Across Languages: Topic Hierarchies based on Synsets

Carlos Badenes-Olmedo - *Universidad Politécnica de Madrid (Spain)*
Jose-Luis Redondo-Garcia - *Amazon Research (UK)*
Oscar-Corcho - *Universidad Politécnica de Madrid (Spain)*

11 de Diciembre de 2019

MOTIVATION: multi-language corpora



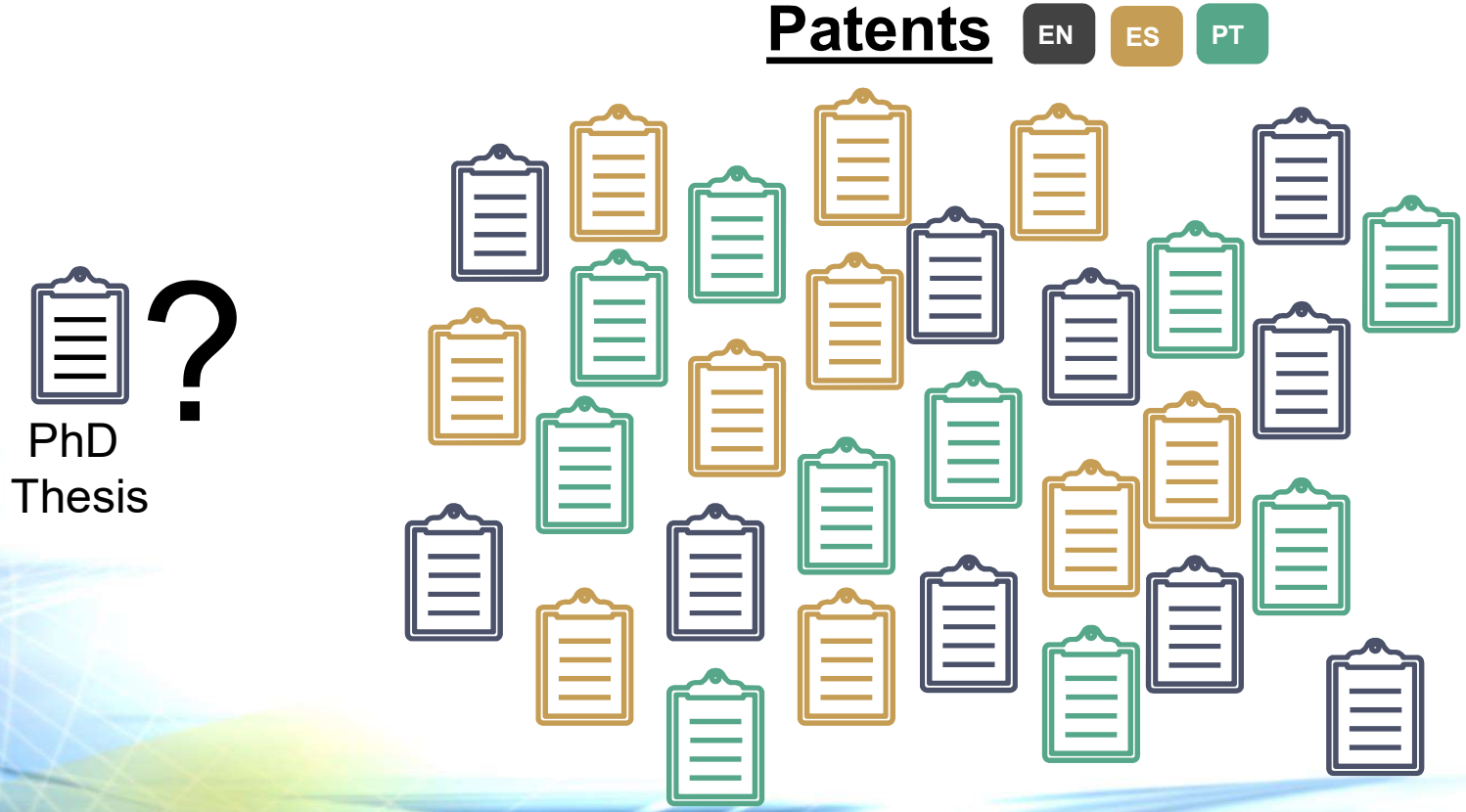
Carlos has just finished his doctorate thesis. It describes a really disruptive algorithm.

He believes his work will have a high impact on the industry and its intellectual property should be protected.

But first he must make sure that it has not been previously registered, not only in his country but somewhere in the world.

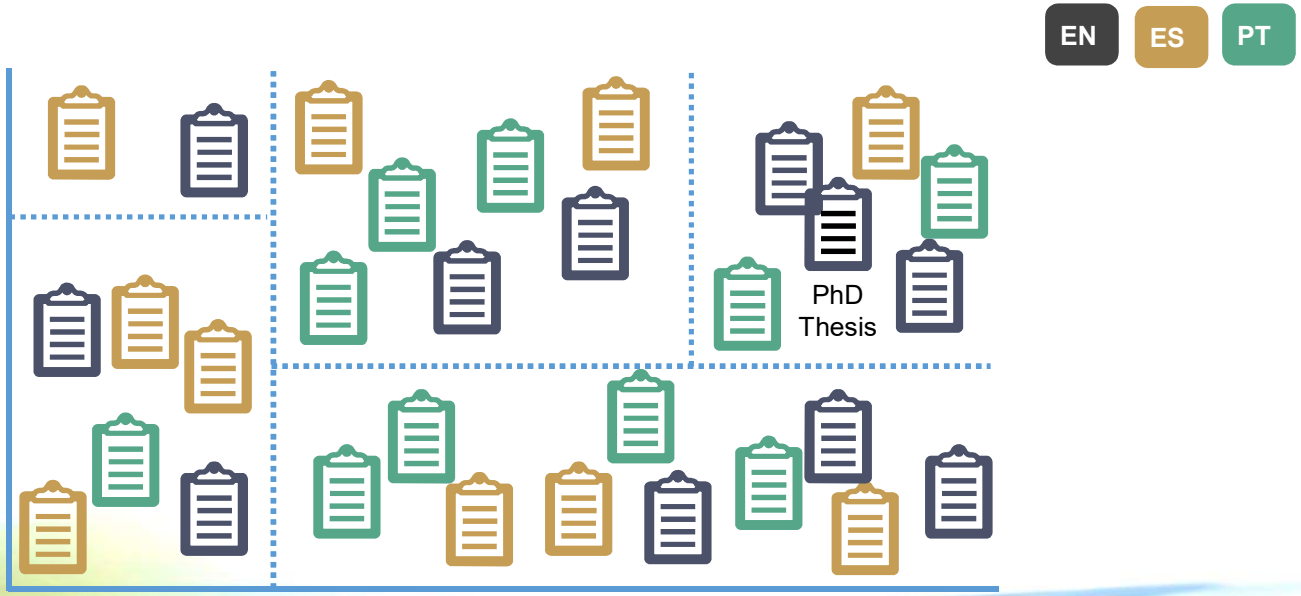
PROBLEM: Cross-language Information Extraction

- **Large-scale** retrieval of documents in multi-lingual corpora requires:
  - **Document representation (P1)**
  - **Comparison across languages (P2)**
  - **High-dimensional correlation matrix (P3)**



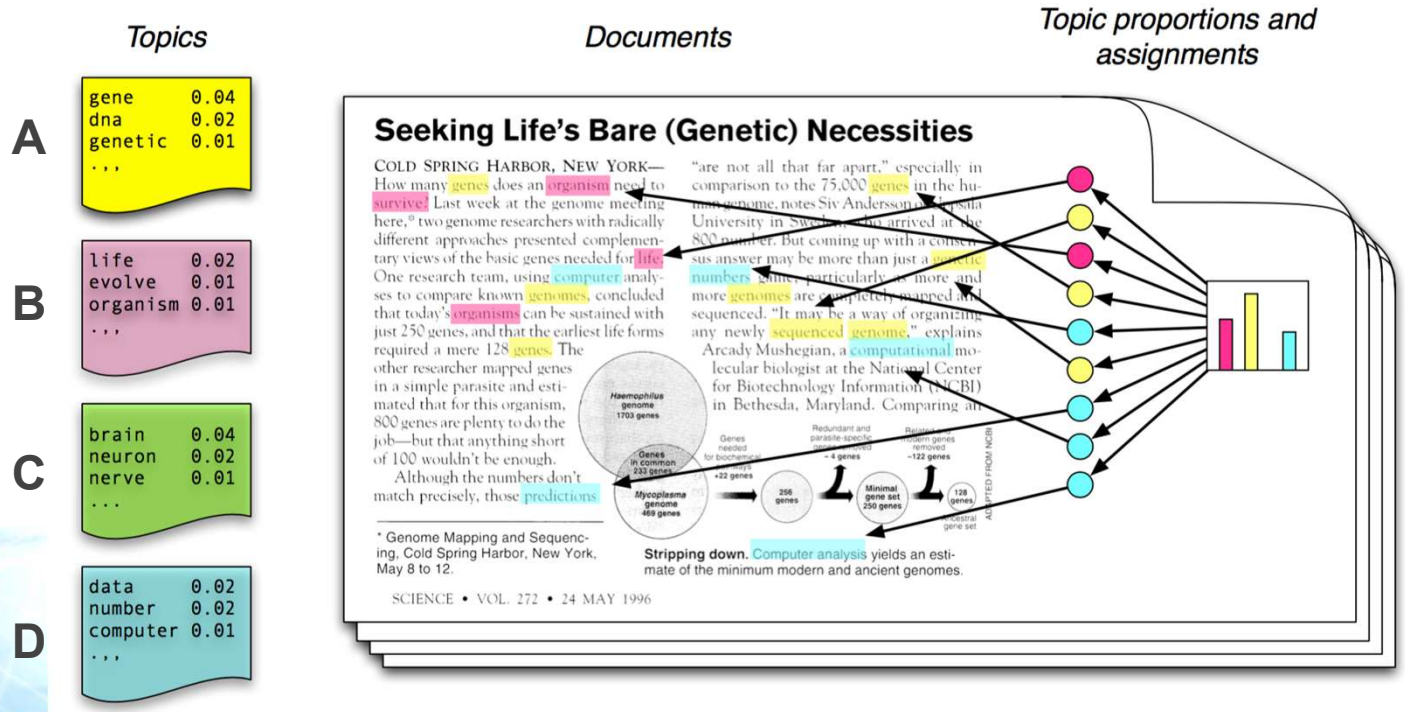
CONTRIBUTION: Cross-lingual Semantic Annotations

- a novel **cross-lingual document similarity algorithm** based on hierarchies of *synsets*
- an open-source **implementation** of the algorithm
- **data-sets** and **pre-trained models** to facilitate other researchers to replicate our experiments and validate and test their own ideas
  - https://github.com/cbadenes/crosslingual-semantic-similarity



RELATED WORK: (P1) Document Representation

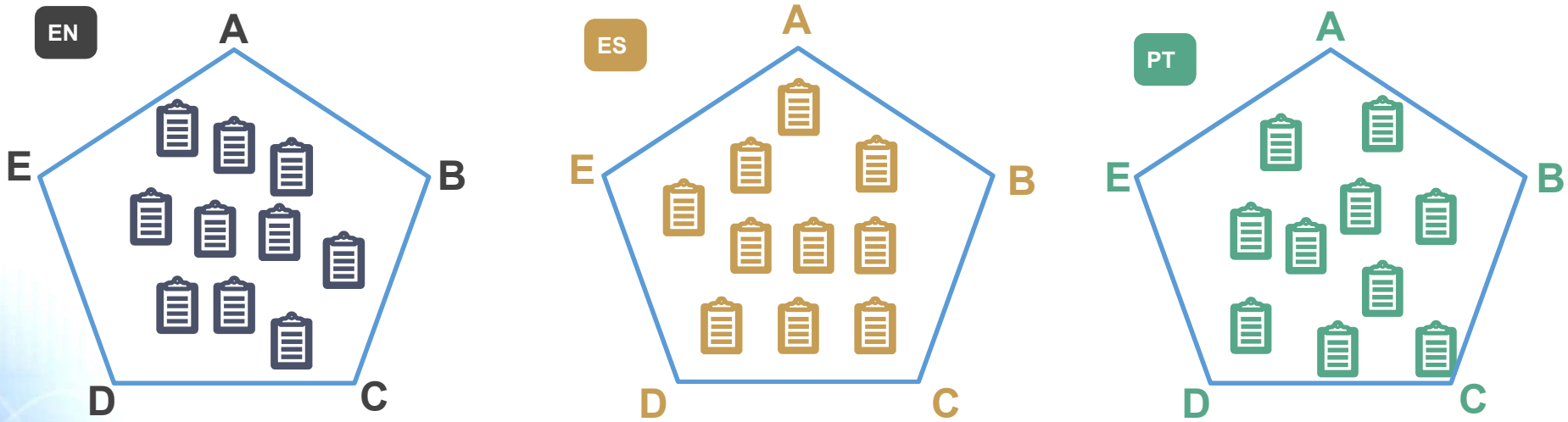
- **Probabilistic Topic Models** [Blei et al, 2003]
  - Each **topic** is a distribution over words
  - Each **word** is drawn from one of those topics
  - Each **document** is a mixture of corpus-wide topics
    - Vector of topic distributions



Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4–5), 993–1022.

RELATED WORK: (P2) Comparison across languages

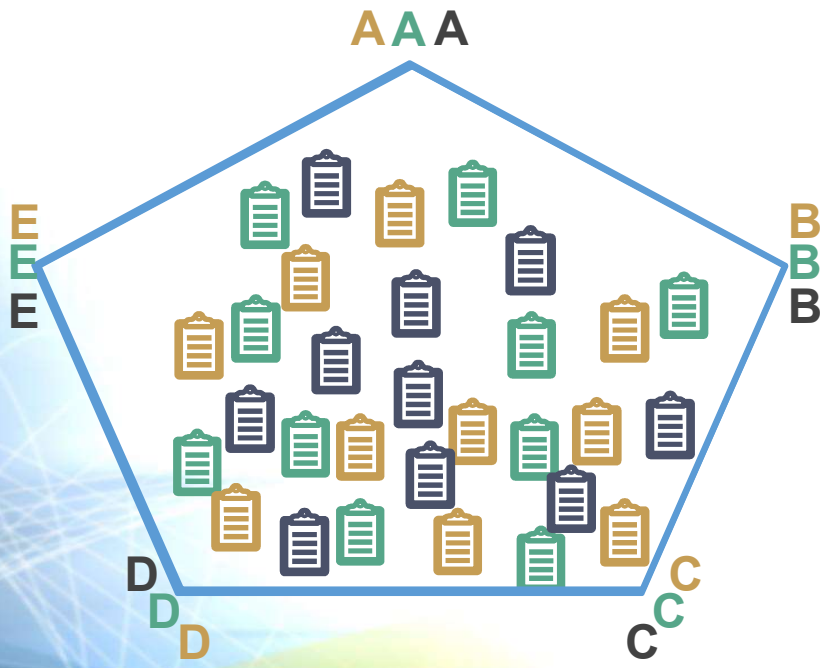
- **Multi-Lingual Topic Models** [Viulic et al. 2015]
  - ***language-specific*** descriptions of each ***topic*** from documents in multi-lingual corpora
  - adding ***supervised*** *association* between languages by using:
    - ***parallel*** corpus (sentence-aligned documents)
    - or ***comparable*** corpus (theme-aligned documents)



A 'communication system'	A 'sistema de comunicación'	A 'sistema de comunicação'
radio	equipo	rede
equipment	red	comunicação
network	comunicación	electrónico
communication	espectro	acesso
regulatory	electromagnético	utilizador

RELATED WORK: (P2) Comparison across languages

- similarity based on **density distributions** derived from the topic distributions
- shared labels as **supervised method** to align topics from different languages
- require ***parallel*** or ***comparable*** **corpora**



Distance Metrics

$$KL(P,Q) = \sum_{i=1}^{K} p(x_i) \log \frac{p(x_i)}{q(x_i)} \quad (1)$$

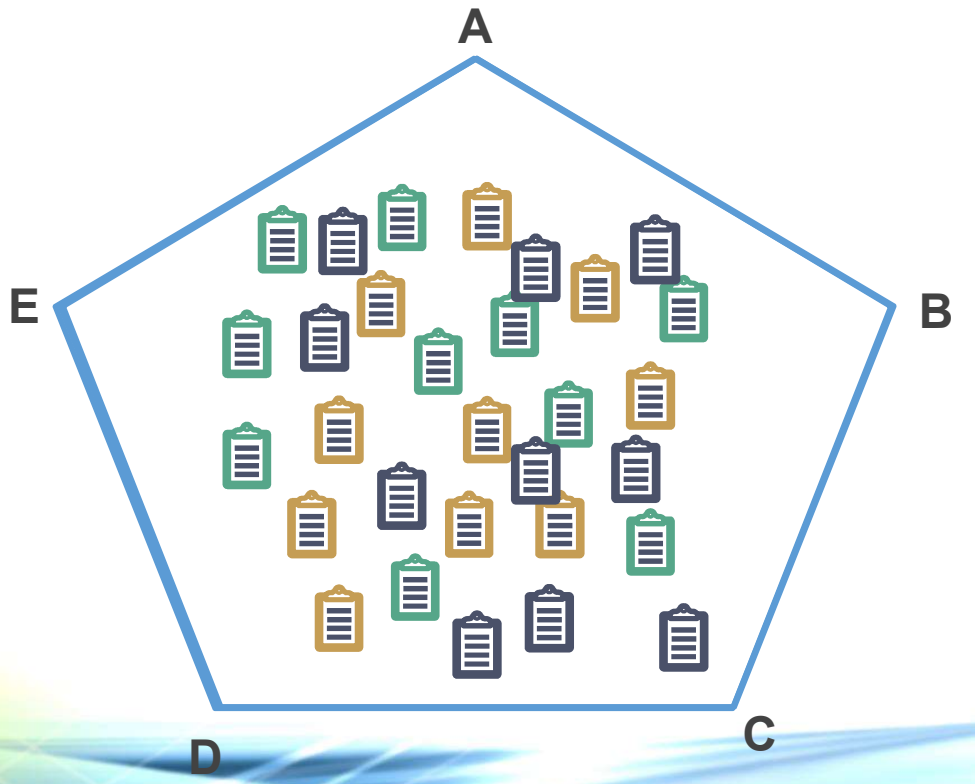
$$JS(P,Q) = \frac{1}{2} KL\left(p, \frac{p+q}{2}\right) + \frac{1}{2} KL\left(q, \frac{p+q}{2}\right) \quad (2)$$

$$He(P,Q) = \sum_{i=1}^{K} \left(\sqrt{p(x_i)} - \sqrt{q(x_i)}\right)^2 \quad (3)$$

$$S2JSD(P,Q) = \sqrt{2 * JS(P,Q)} \quad (4)$$

RELATED WORK: (P2) Comparison across languages

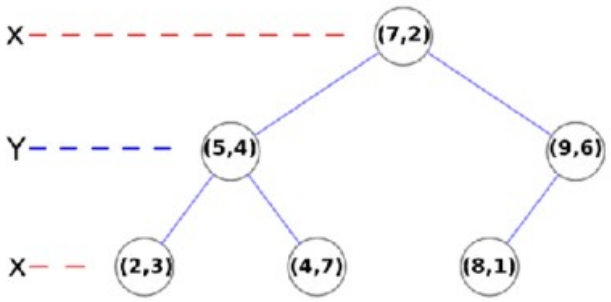
- **Multi-Lingual Dictionaries** [Hao and Paul, 2018]
  - easier to obtain and more **widely available** than parallel corpora (e.g PANLEX or Wiktionary)
  - models are built from words in a **target language**
  - dictionaries as **supervised method** to align topics
    - **topics conditioned** by pre-established language relations

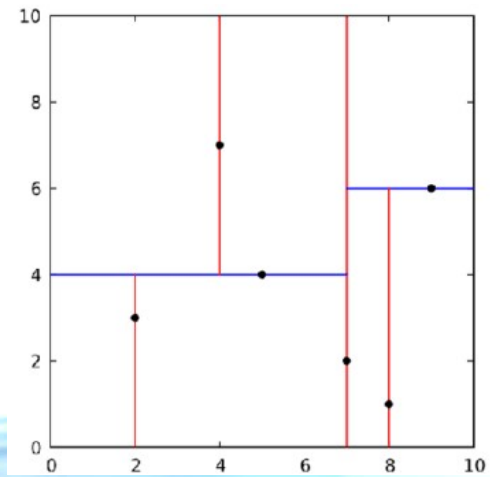


RELATED WORK: (P2) High-Dimensional Matrix

- Exact similarity computations require to have complexity $O(n^2)$ for neighbours detection tasks or $O(k \cdot n)$ computations when k queries are compared against a dataset of n documents
- Computation can be an **approximate nearest neighbour** (ANN) search problem based on topic distributions [Mao et al, 2017]
- It transforms data point (i.e vector of topic distributions) from the original feature space into a binary-code space, so that similar data points have larger probability of collision

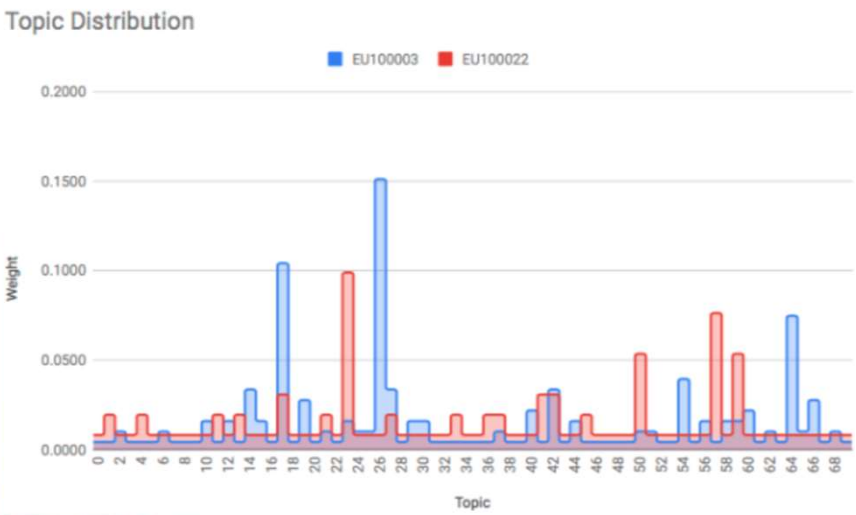
	d1	d2	d3	d4	..
d1	1.0	?			
d2	?	1.0			
d3			1.0		
d4				..	
..					1.0

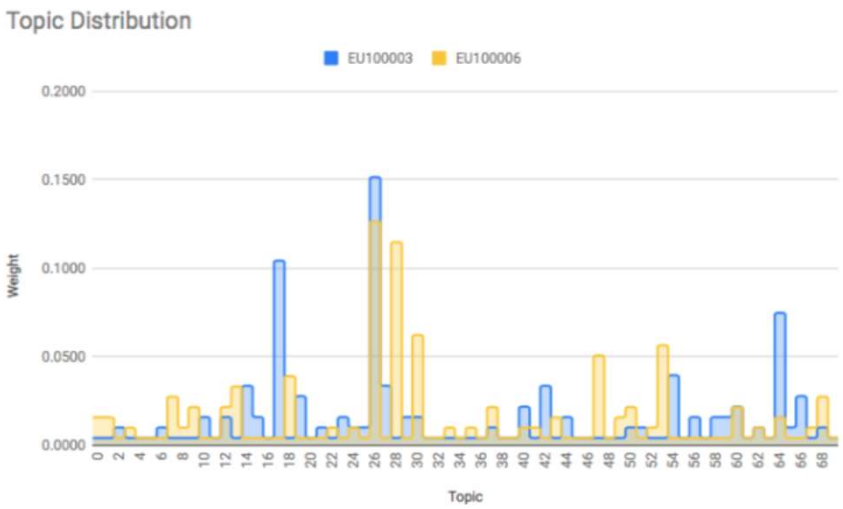




RELATED WORK: (P3) High-Dimensional Matrix

- ***density-based metrics*** consider that similar documents do not necessarily share the most relevant topic for each of them.



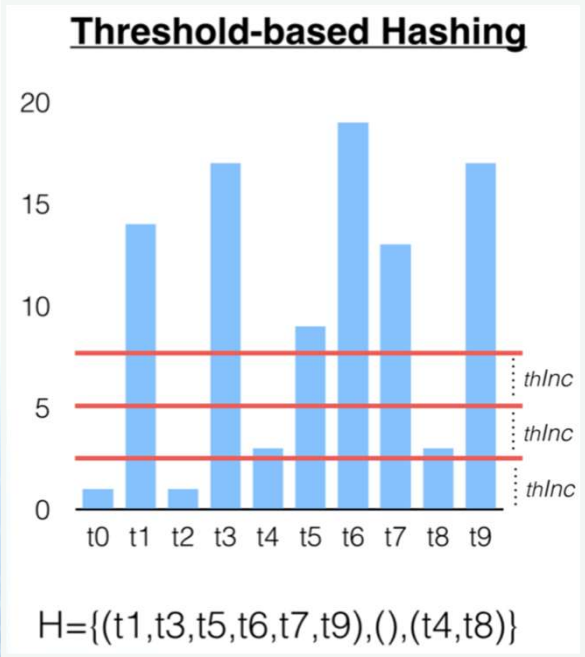


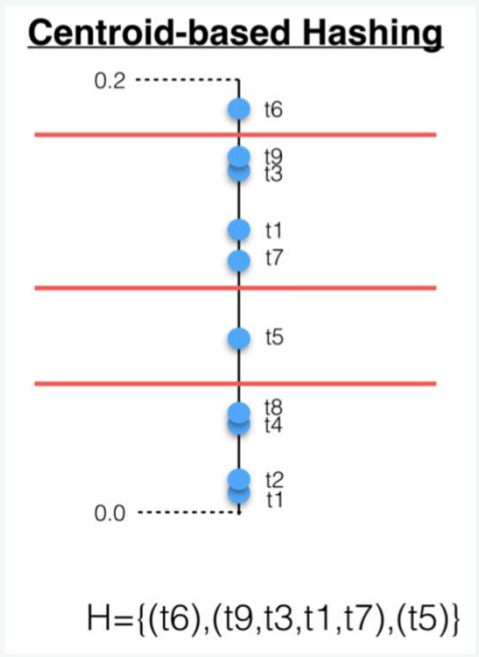
a) $sim_{JSD} = 0.74$

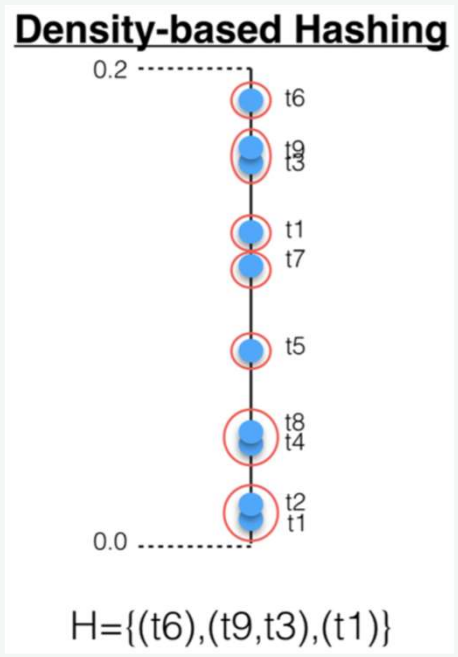
b) $sim_{JSD} = 0.71$

RELATED WORK: (P3) High-Dimensional Matrix

- Hashing Topic Distributions [Badenes-Olmedo et al, 2019]
  - hierarchical set of topics based on their relevance



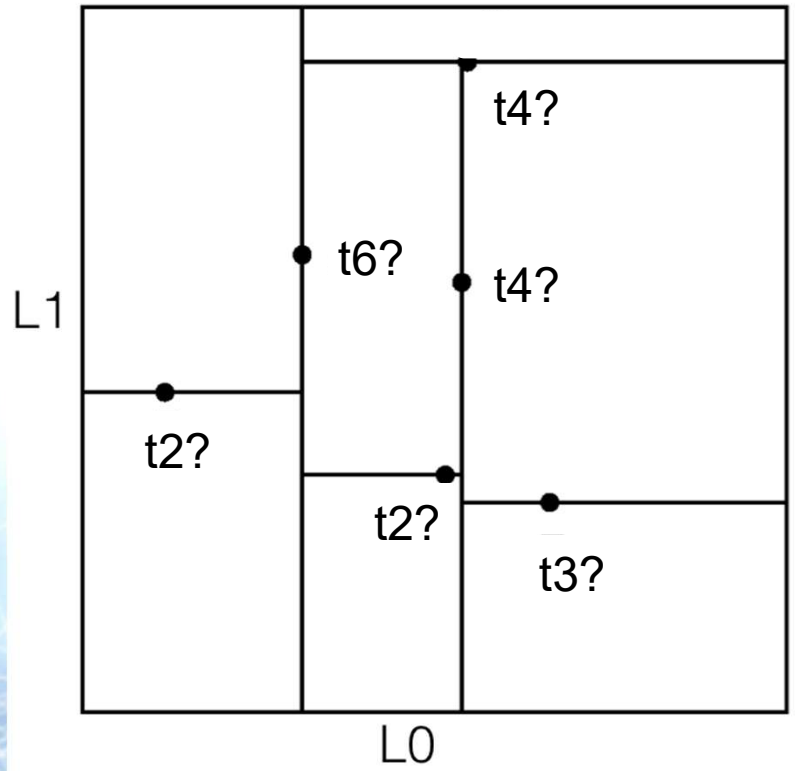


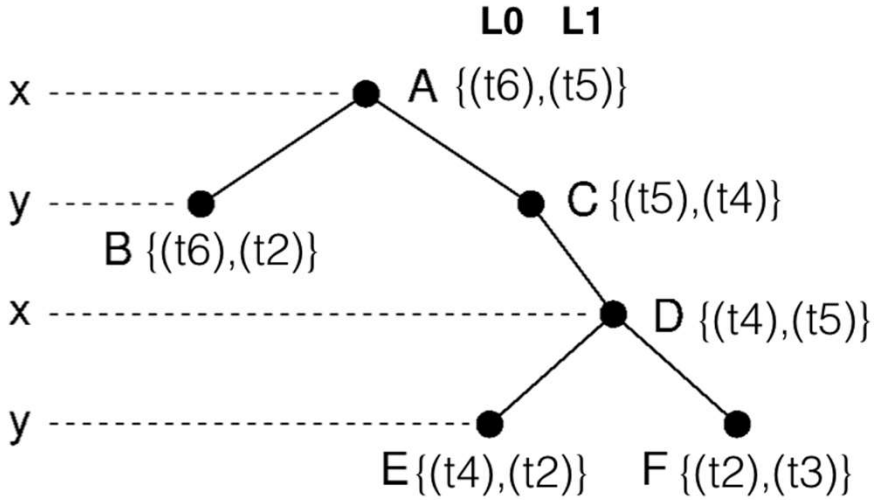


Badenes-Omedo, C., Redondo-García, J. L., & Corcho, O. (2019). ***Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms***. Semantic Web Journal.

RELATED WORK: (P3) High-Dimensional Matrix

- Hashing Topic Distributions [Badenes-Olmedo et al, 2019]
  - hierarchical set of topics based on their relevance

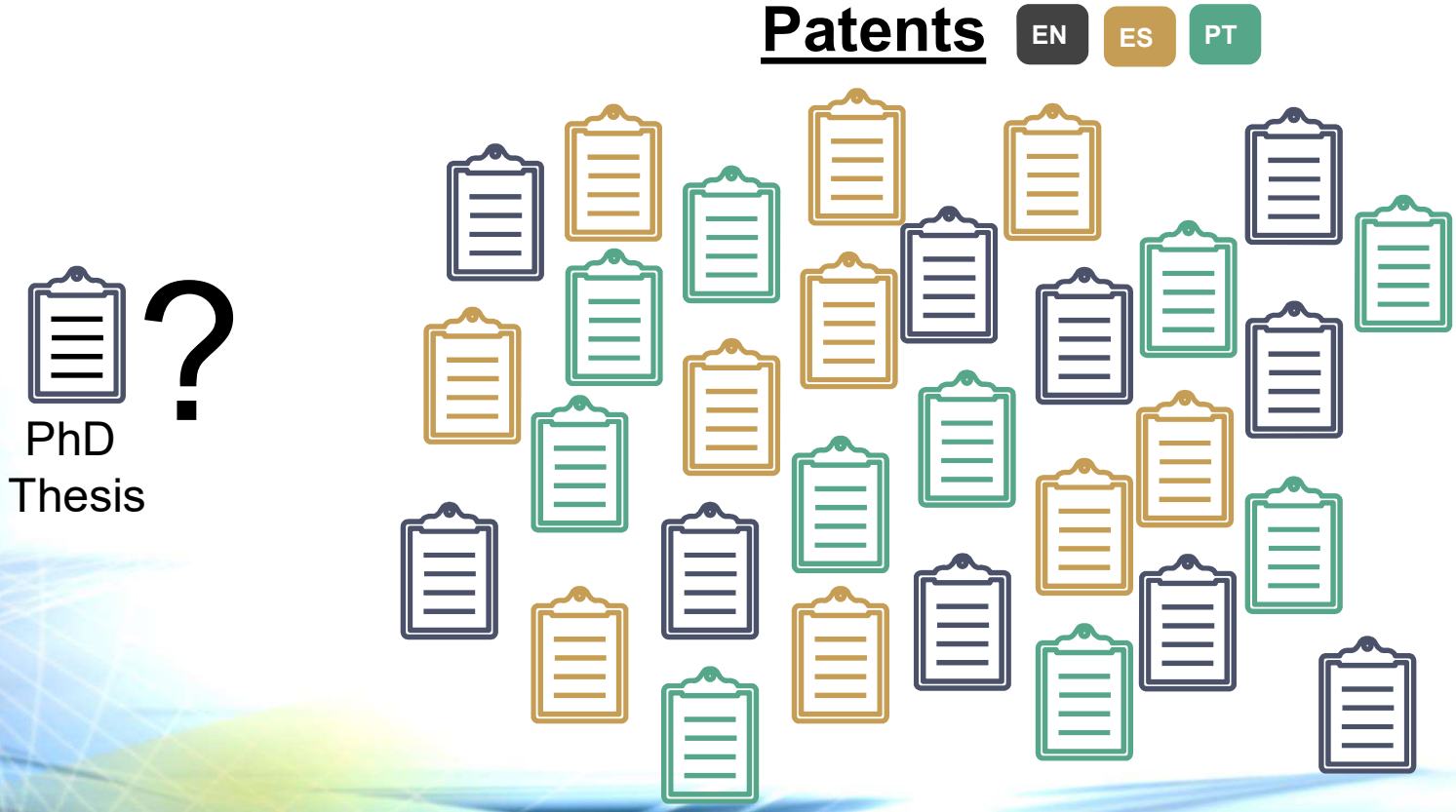




Badenes-Omedo, C., Redondo-García, J. L., & Corcho, O. (2019). ***Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms***. Semantic Web Journal.

PROBLEM: Cross-language Information Extraction

- **Large-scale** retrieval of documents in multi-lingual corpora requires:
  - ✓ **Document representation (P1)**
  - ✗ **Comparison across languages (P2)** ***(supervised solution)***
  - ✓ **High-dimensional correlation matrix (P3)**



PROPOSAL: Cross-lingual Document Similarity

Plan TL IBERLEGAL

Hypothesis:

"similar documents share synsets derived from their main topics that have not been previously aligned"

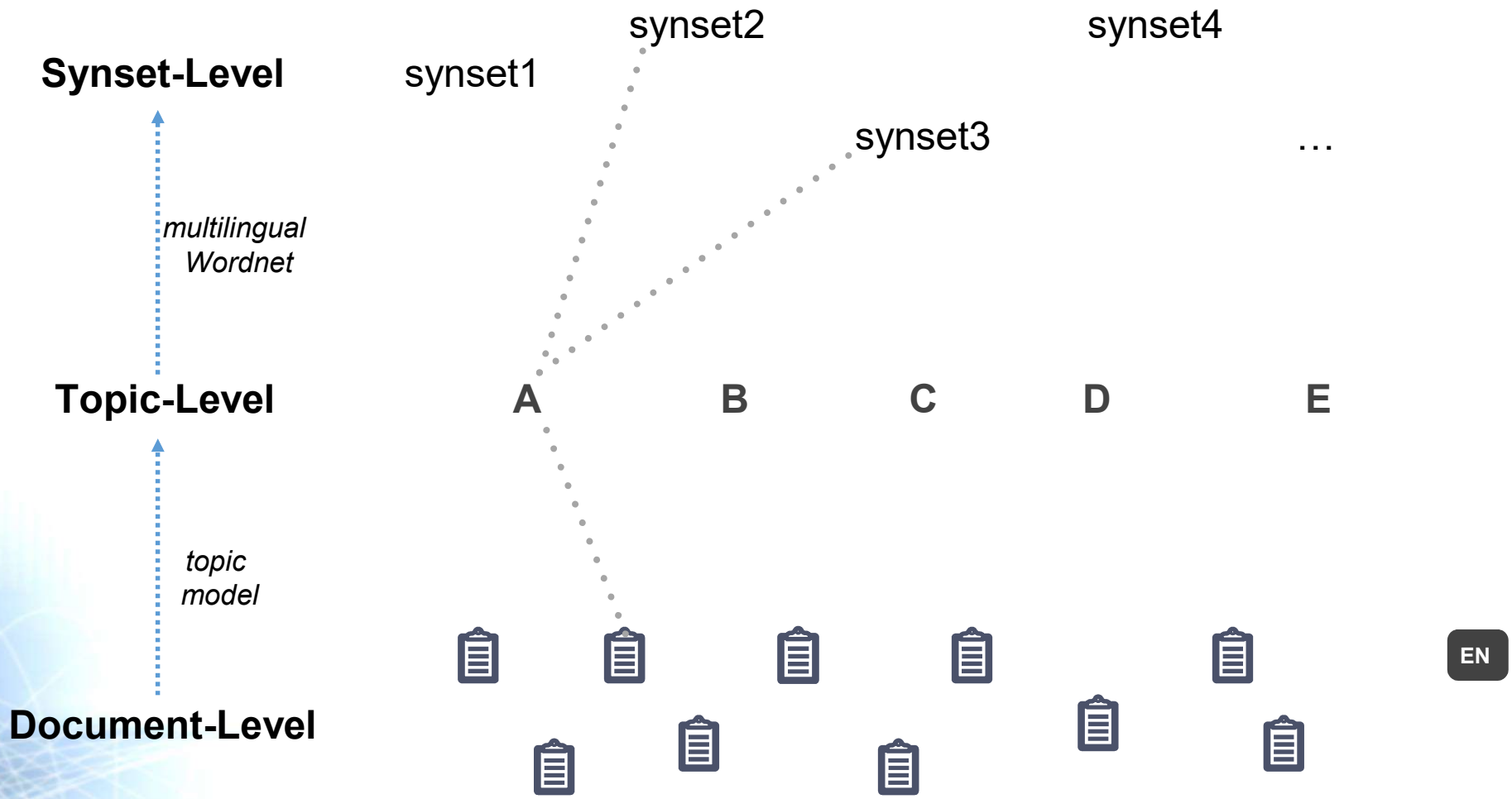
We propose an **unsupervised** algorithm to:

- relate similar documents in multi-lingual corpora
  (no translations required)

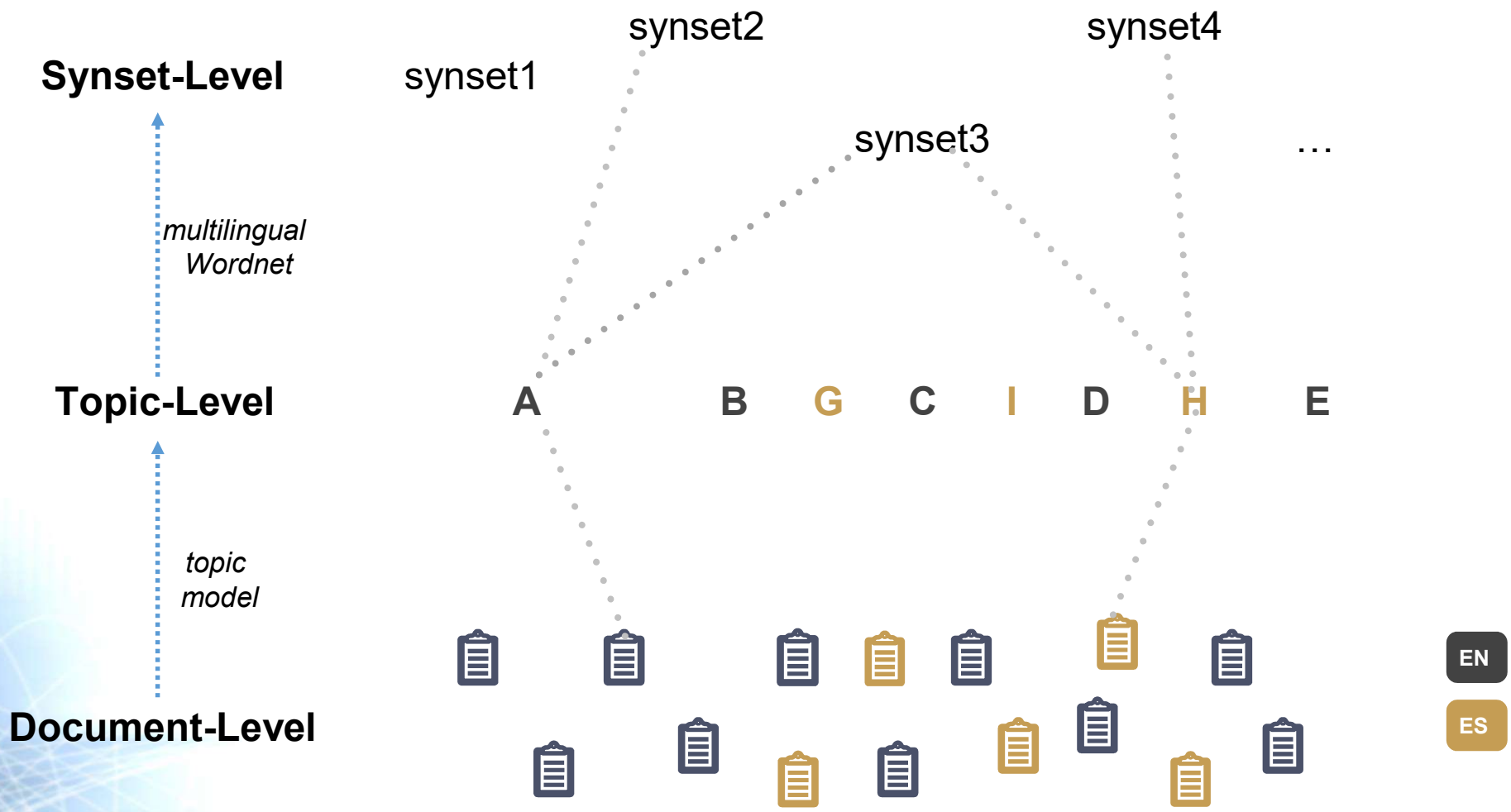
- creating cross-lingual annotations through language-specific concept hierarchies
  (no parallel or comparable corpora required)

- based on the most relevant topics
  (no density-based distance metrics)

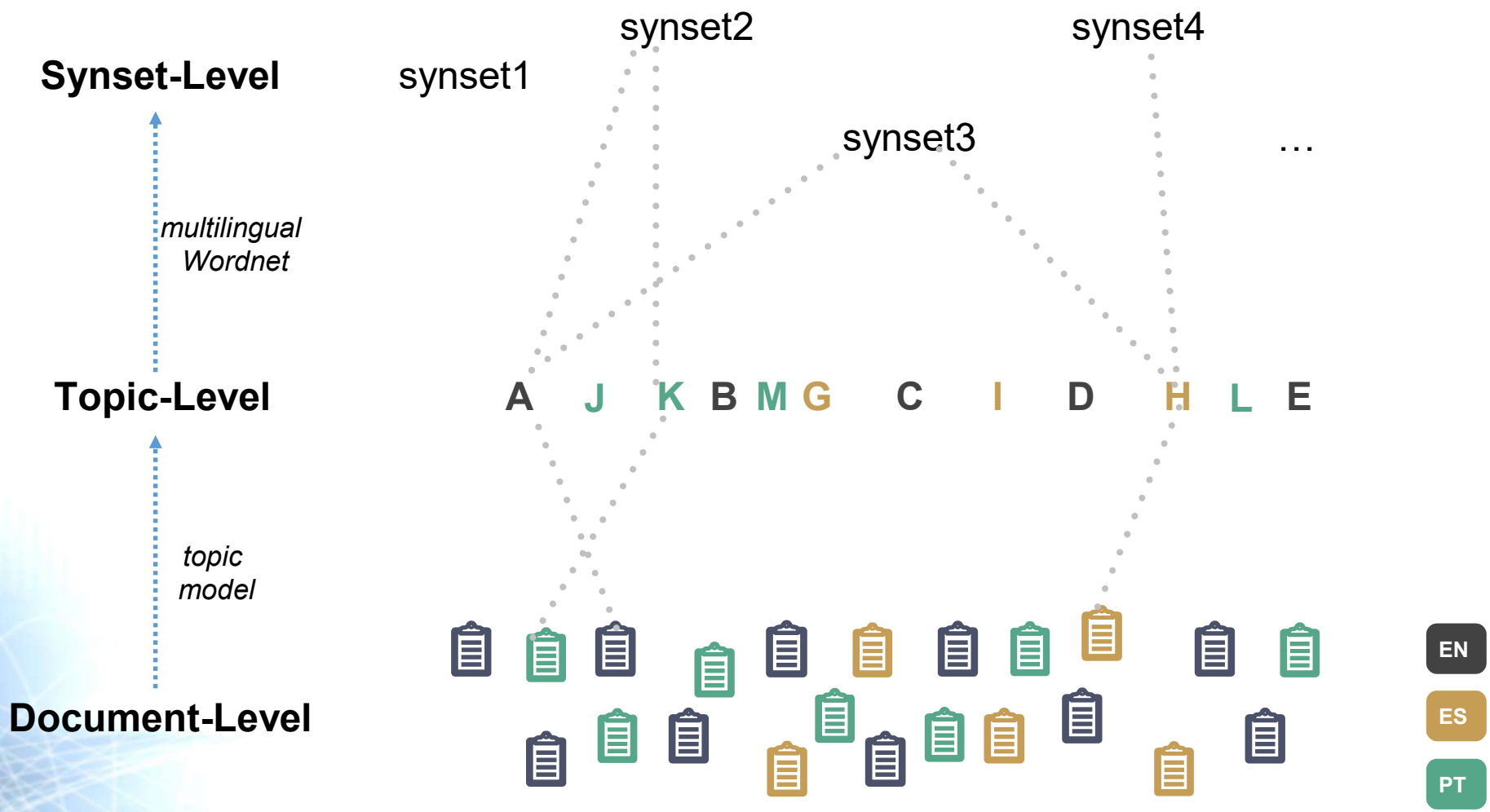
PROPOSAL: Cross-lingual Annotations



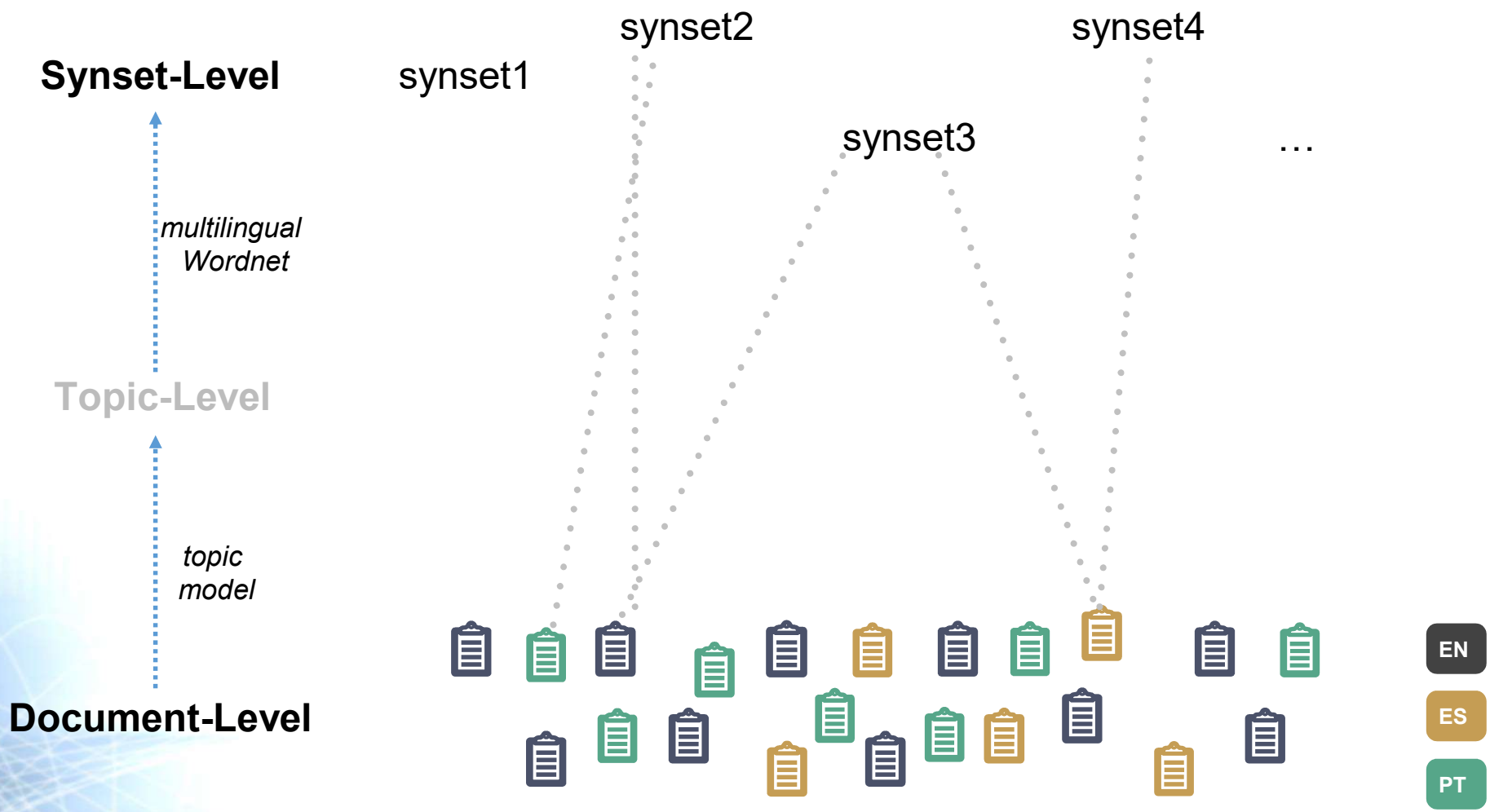
PROPOSAL: Cross-lingual Annotations



PROPOSAL: Cross-lingual Annotations

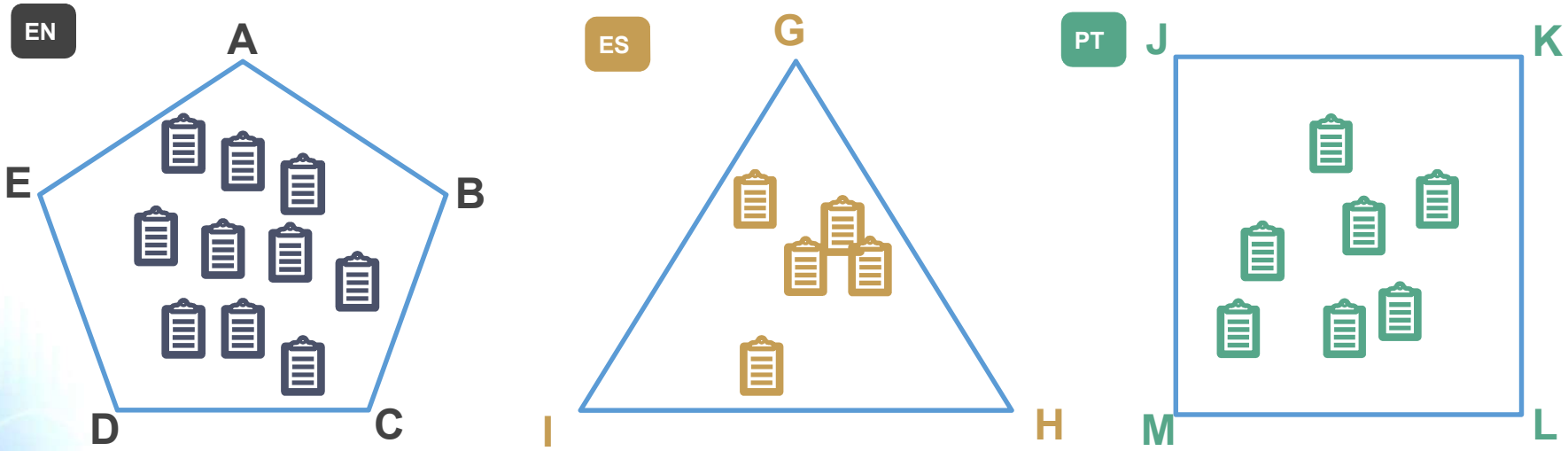


PROPOSAL: Cross-lingual Annotations



METHOD: Cross-lingual Synset-based Topics

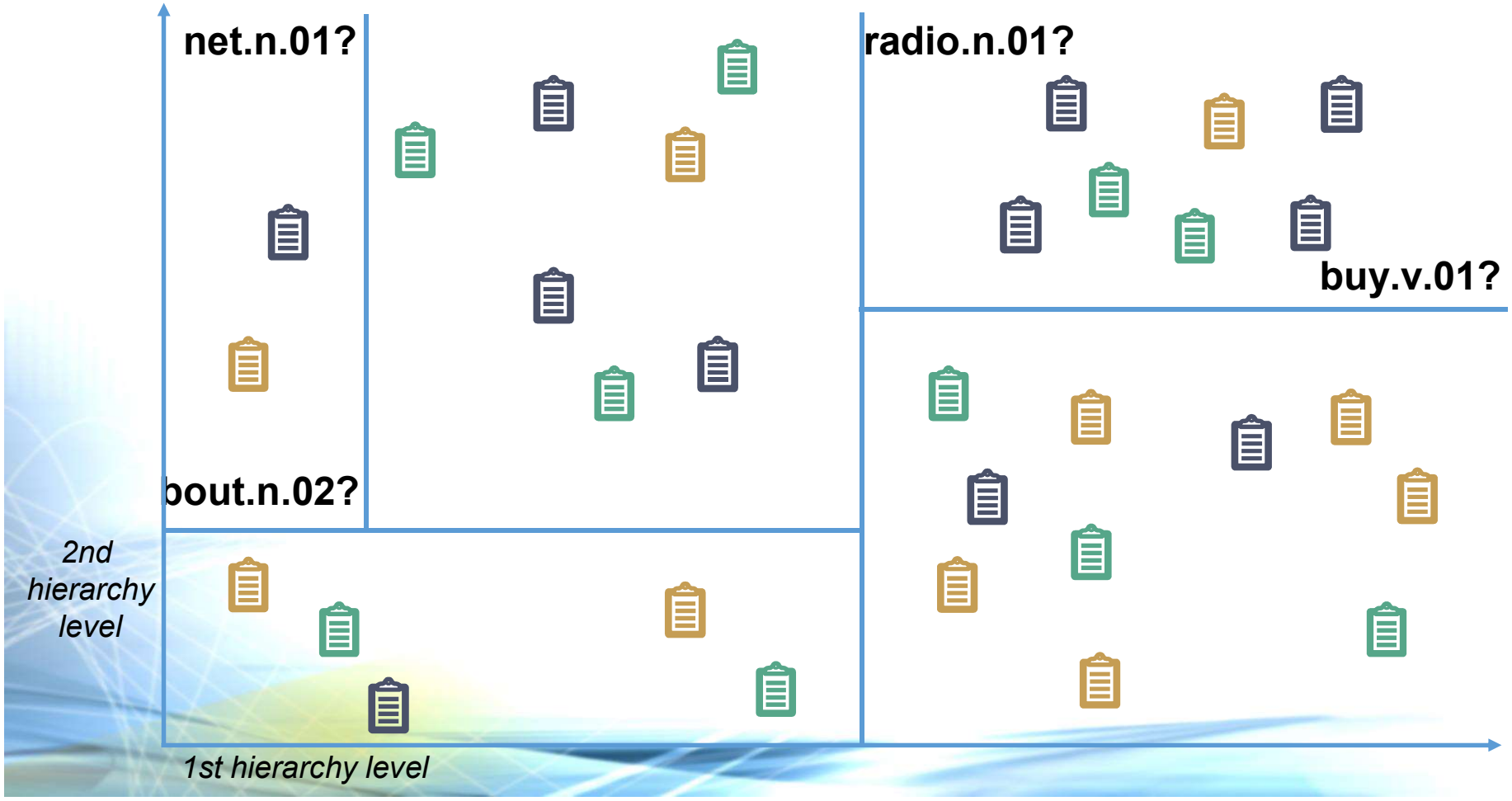
- based on **language-specific** concepts
- no parallel or comparable data required
- **wordnet** synset-based alignment



A	G	K
radio.n.01	kit.n.02	access.n.02
equipment.n.01	**equipment.n.01**	approach.n.07
network.n.02	**net.n.02**	entree.n.02
net.n.06	web.n.06	**communication.n.02**
communication.n.02	**communication.n.02**	bout.n.02

METHOD: Scalable Search Space

- hierarchical-set of topics from ***relevance***
- ***nearest neighbour searches*** (*k-d tree)*
- Boolean Similarity (***Jaccard Index***)

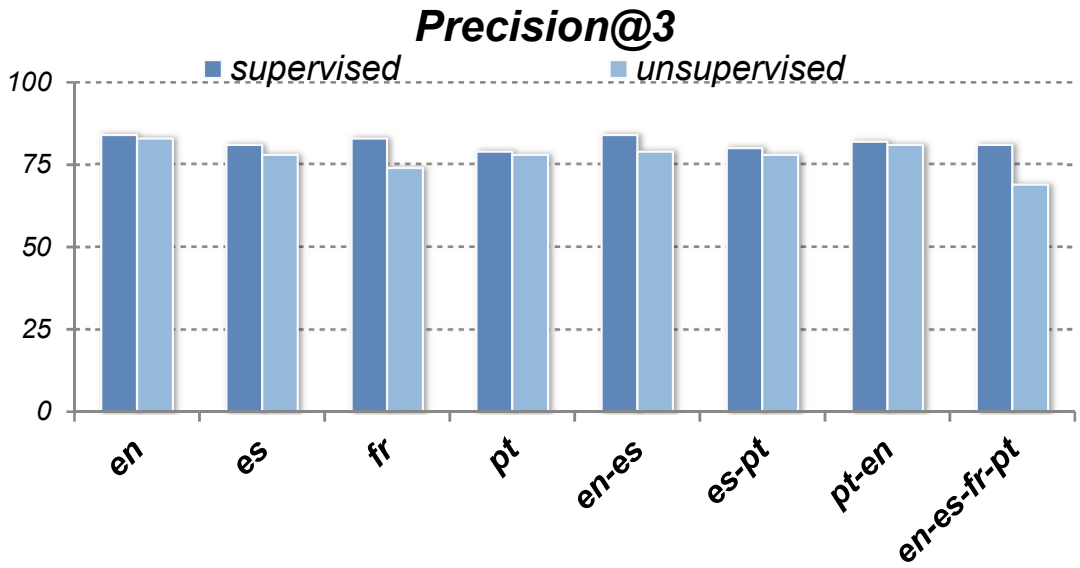


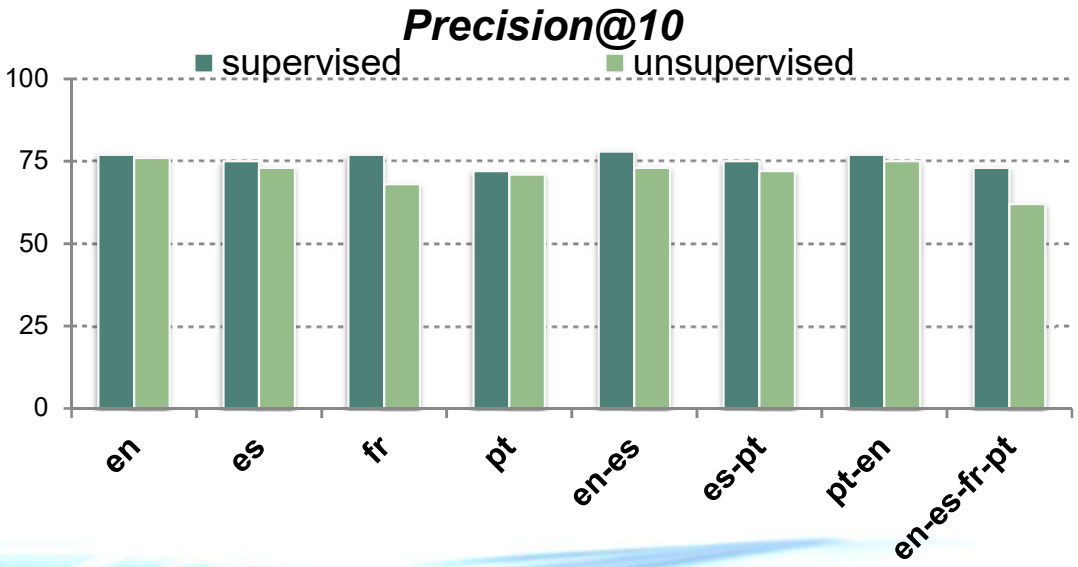
EXPERIMENTAL SETUP

- Comparison of supervised vs unsupervised methods for: **document classification** and **document retrieval** tasks
- **Datasets:** Parallel corpora is required to use supervised method
  - *JRC-Acquis Corpora*: legislative texts in European Union
    - English, Spanish, French and Portuguese editions (~80k docs)
    - Documents manually annotated with **EUROVOC** categories (~6k labels)
- **Models:**
  - categories processed to satisfy the independence assumption of probabilistic topics (~400 topics)
  - Lemmatized expressions of names, verb and adjetives
  - LabeledLDA (supervised) and LDA (unsupervised) models
- **Resources**:
  - Datasets and Models available at: https://github.com/cbadenes/crosslingual-semantic-similarity

EVALUATION

- **Document Retrieval Task**
  - Metrics: *precision@3, precision@5 and precision@10*
  - Test Data: ~1k docs *(monolingual, bi-lingual or multilingual documents)*
  - *Comparison of topN similar documents based on EUROVOC categories and based on annotations created by the model:*
    - *supervised = alignment + labeledLDA*
    - ***unsupervised** = LDA + WordNet Synsets*





WRAPPING UP

- documents written in different languages are aligned in a **single representation space** without the need for translation
- the feature space ***does not lose the semantics*** offered by the topics when is approximated to nearest neighbours.
- **No parallel or comparable corpora** is required to train the models.
- topic annotation by set of synonyms should be improved to **filter** those concepts that are **not sufficiently representative.**
- <u>Next Steps:</u>
  - multi-lingual embeddings to align topics

¡¡ GRACIAS !!

Legal Document Retrieval Across Languages: Topic Hierarchies based on Synsets

Carlos Badenes-Olmedo

Jose-Luis Redondo García

Oscar Corcho

www.upm.es

www.oeg-upm.net

theybuyforyou.eu

www.PlanTL.es

PlanTecnologiasLenguaje@mineco.es