Cross-lingual Word Embeddings for Translation Inference

Sina Ahmadi
Atul Kr. Ojha
Shubhanker Banerjee
John P. McCrae

Data Science Institute, NUI Galway

A World Leading SFI Research Centre

Science Foundation Ireland For what's next

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

elexis european lexicographic infrastructure

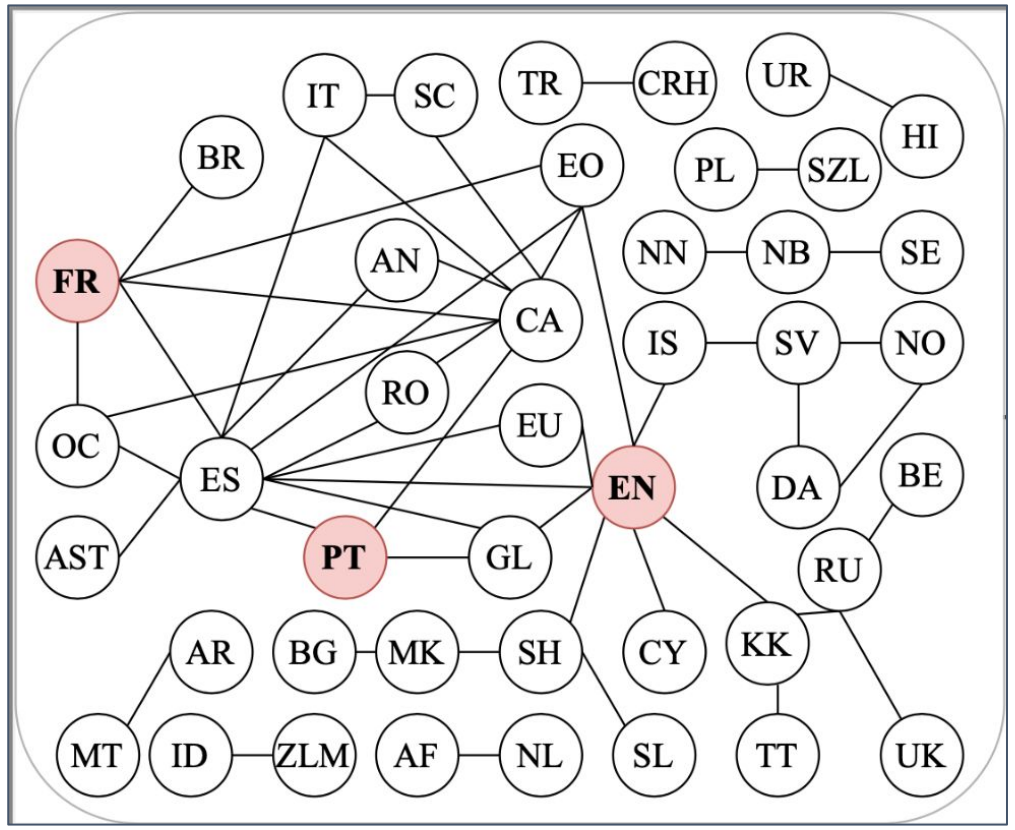
HOST INSTITUTIONS
DCU
NUI Galway OÉ Gaillimh
UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh
UCD DUBLIN

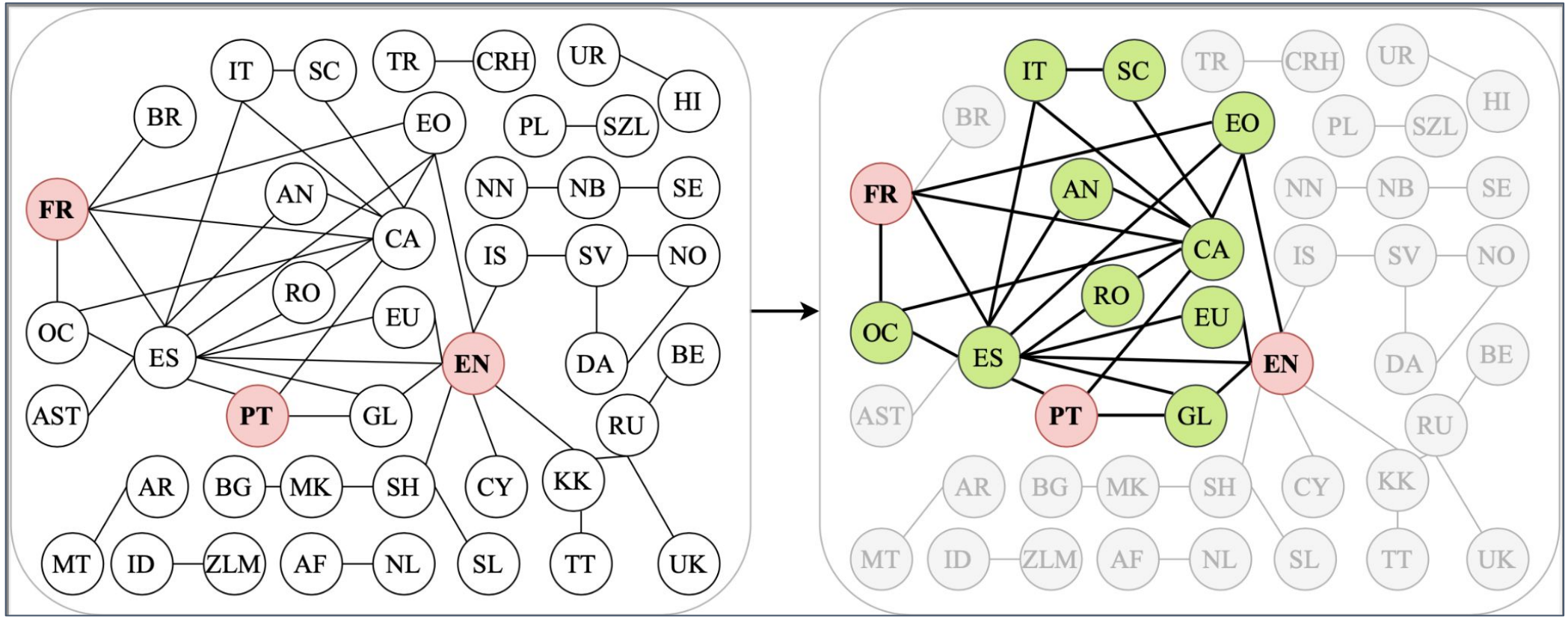
PARTNER INSTITUTIONS
Maynooth University National University of Ireland Maynooth
Tyndall National Institute
Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin
University of Limerick Ollscoil Luimnigh

FUNDED BY:
European Union European Regional Development Fund
Science Foundation Ireland For what's next

NUIG's submission to TIAD 2021

- Submitted five systems based on **graph analysis** and **cross-lingual word embeddings**
- This year's submission is in line with the previous year's submission





Approach 1: Exploring the graph structure

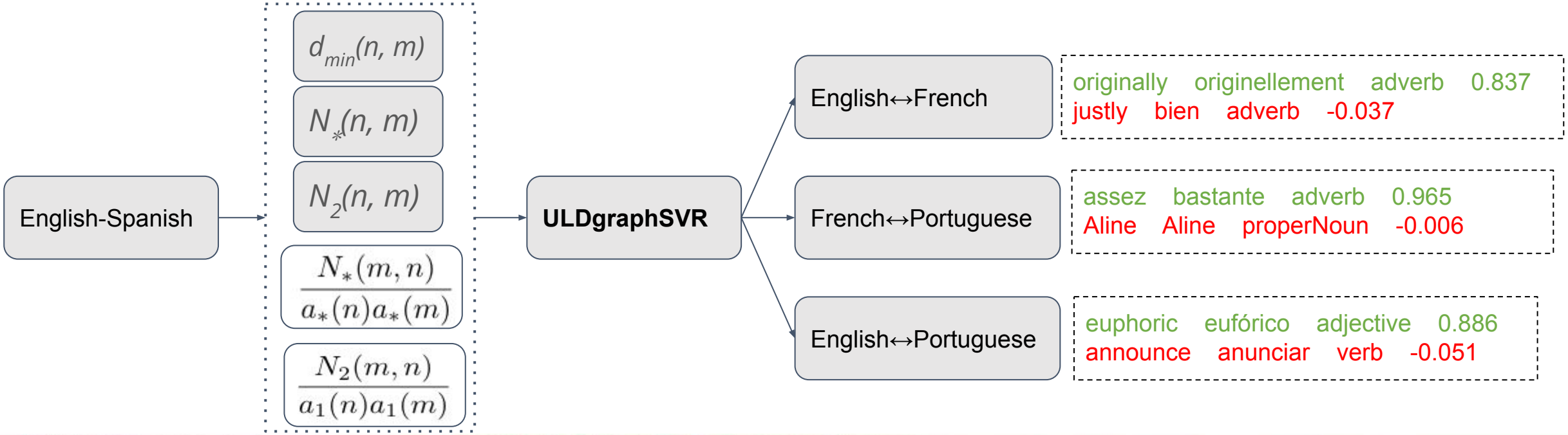
Approach 1: Graph Analysis

1. **ULD_OnetaSVR:** We use McCrae and Arcan's algorithm (submitted to TIAD 2020) as features to train a classifier

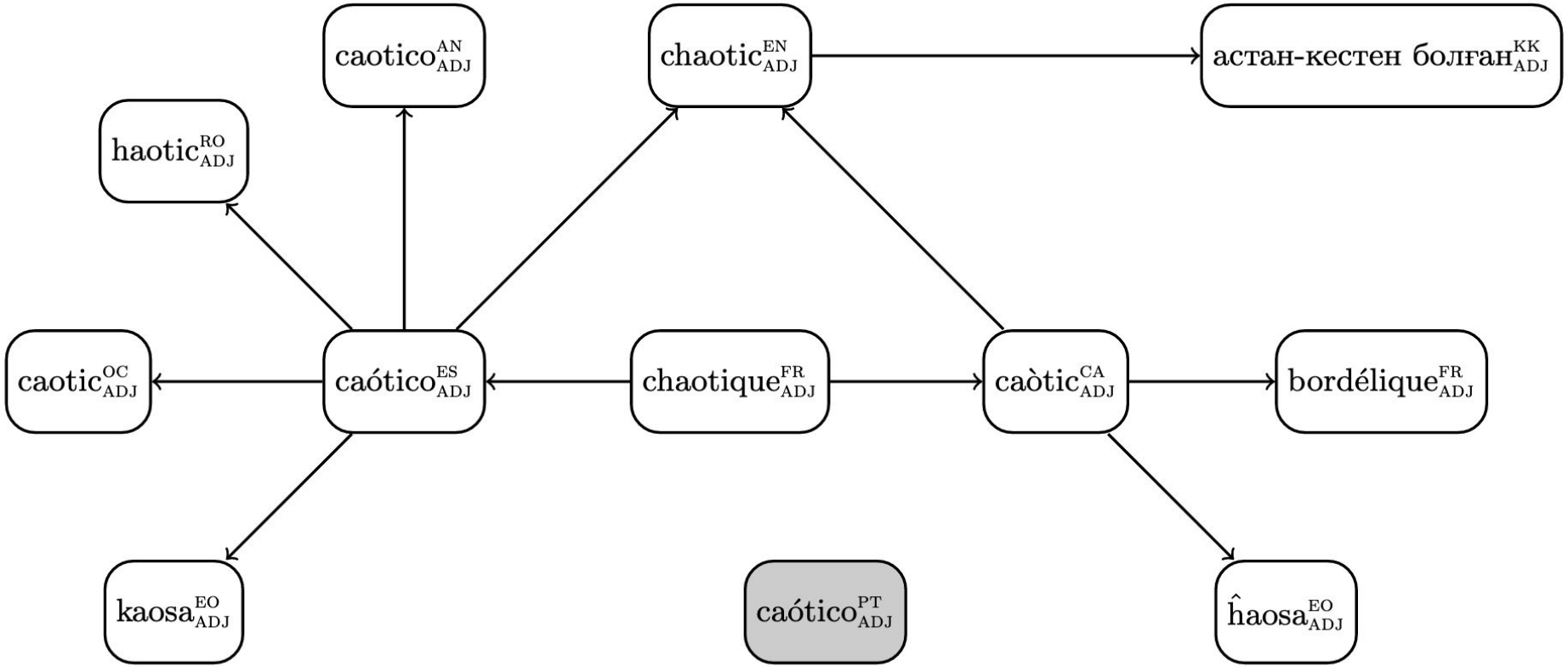
2. **ULD_graphSVR**: We create regression models based on:
   - $d_{min}(n, m)$: The minimum distance in the graph between two nodes.
   - $N_*(n, m)$: The number of paths between the nodes of any length.
   - $N_2(n, m)$: The number of paths between the nodes of length 2.
   - $a_*(n)$: The number of nodes reachable from node *n*.
   - $a_1(n)$: The number of nodes directly connected to node *n*.

Approach 1: Graph Analysis

For ULD_graphSVR, we use the following five features (based on One-Time Inverse Consultation or OTIC) to train a support vector regression models:



Approach 1: Limitations due to coverage



Approach 2: Cross-lingual word embeddings mapping

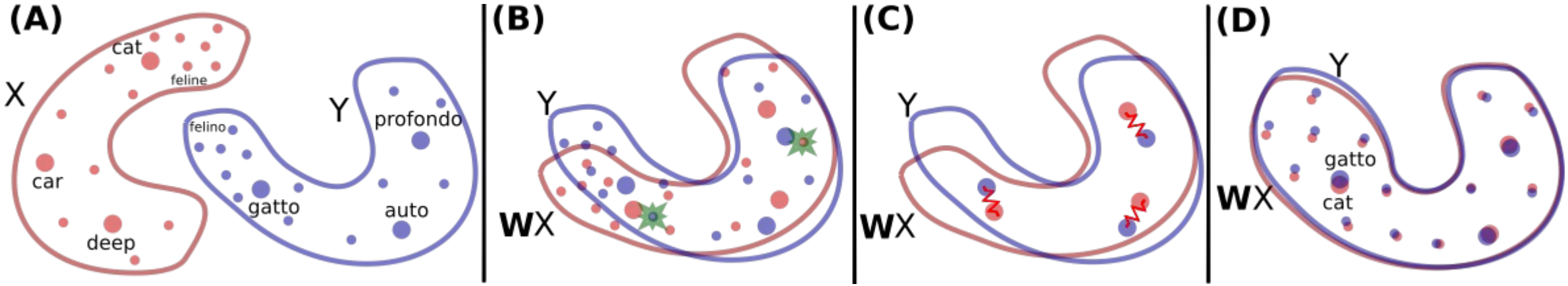
Using unsupervised cross-lingual word embedding mapping techniques, find a **mapping between the monolingual word embedding spaces** of the source and target languages.

- VecMap*
- MUSE**

Both methods, induce a seed lexicon automatically assuming approximate isomorphism between source and target spaces

* Artetxe, M., Labaka, G., & Agirre, E. (2017, July). Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 451-462).
** Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. A. (2017). Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043.

Approach 2: Cross-lingual word embeddings mapping



A translation matrix ***W*** is learned based on the spaces X and Y

Approach 2: Muse

- a mapping is learned using the MUSE* unsupervised method and fastText monolingual embeddings of French, English and Portuguese
- the mappings are used to create new translation pairs between the 10 most nearest translations in the target language using cosine similarity
- The cosine similarity score is considered as the confidence score in the final submission
- the part-of-speech of the source word is used for the target predictions as well

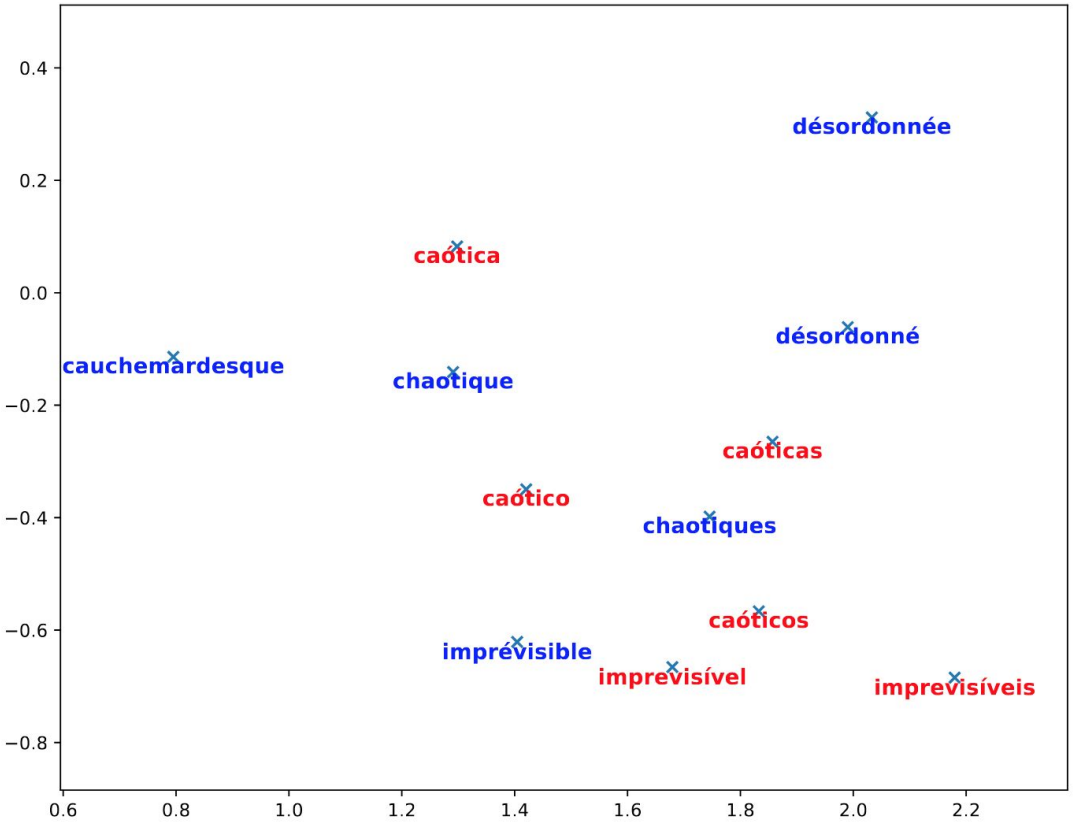
* https://github.com/facebookresearch/MUSE

Approach 2: VecMap*

- an unsupervised method based on cross-lingual embeddings
- only focused on English-French
- using pre-trained French and English fastText monolingual embedding models
- pre-trained UDPipe 2.5 models framework to generate the part-of-speech features (only of the target (French) language)
- the generated parts-of-speech tags were mapped with parts-of-speech tags of the shared task

* https://github.com/artetxem/vecmap

Approach 2: Cross-lingual word embeddings mapping



t-SNE visualization of *chaotique* (adjective in French) in the MUSE multilingual word embeddings of French and Portuguese

Approach 2: Cross-lingual word embeddings mapping

0.73	plus	less	adverb
0.68	plus	more	noun
0.68	plus	more	adverb
0.68	plus	very	adverb
0.66	plus	than	adverb
0.65	plus	quite	adverb
0.63	plus	most	adverb
0.62	plus	comparatively	adverb
0.62	plus	even	adverb
0.61	plus	extremely	adverb
0.60	plus	much	adverb

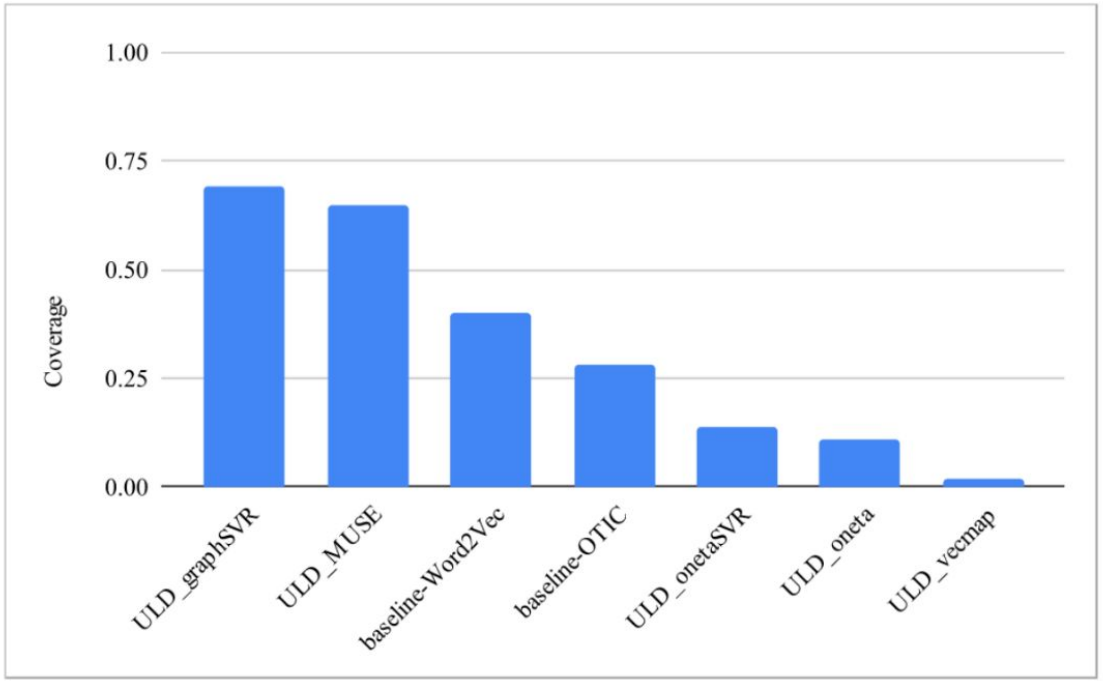
0.73	plutonium	uranium	noun
0.7	plutonium	plutonium	noun
0.68	plutonium	thorium	noun
0.67	plutonium	reactor	noun
0.65	plutonium	deuterium	noun
0.64	plutonium	fission	noun
0.63	plutonium	radioactivity	noun

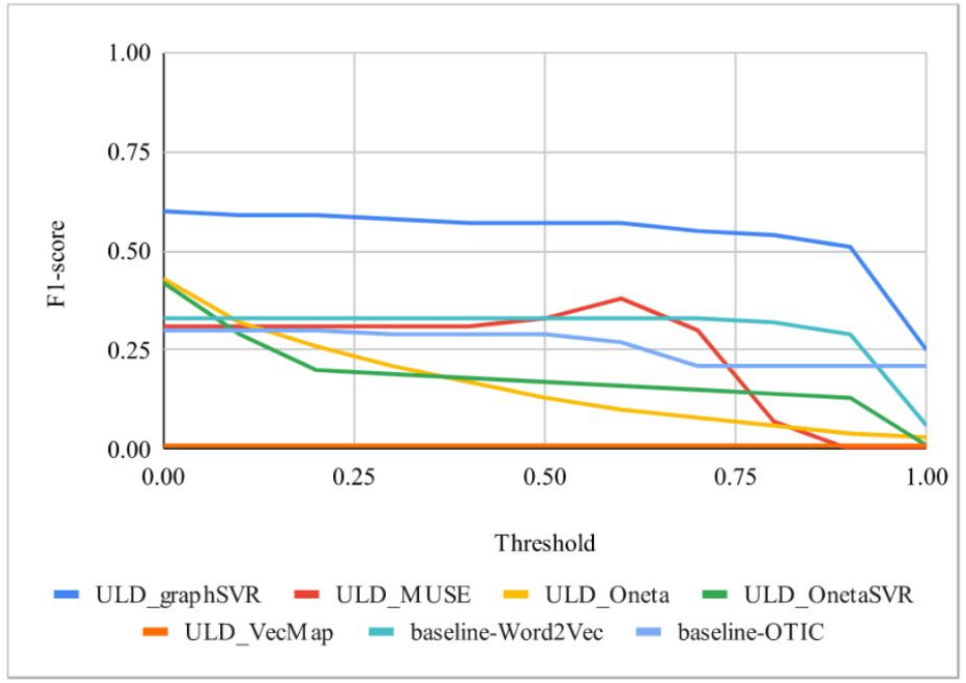
0.5	plurinominal	electoral	adjective
0.55	plurinominal	ballot	adjective
0.54	plurinominal	elect	adjective
0.87	plusieurs	several	determiner
0.77	plusieurs	many	determiner
0.74	plusieurs	various	determiner
0.66	plusieurs	multiple	determiner

Averaged systems results

SYSTEM	PRECISION	RECALL	F1-Measure	COVERAGE
ULD_graphSVR	**0.70**	**0.49**	**0.57**	**0.69**
baseline-Word2Vec	0.69	0.23	0.33	0.40
ULD_MUSE	0.29	0.41	0.33	0.65
baseline-OTIC	0.78	0.18	0.29	0.28
ULD_onetaSVR	0.76	0.10	0.17	0.14
ULD_oneta2	0.64	0.07	0.13	0.11
ULD_vecmap	0.36	0.01	0.01	0.02

Results (English-French)

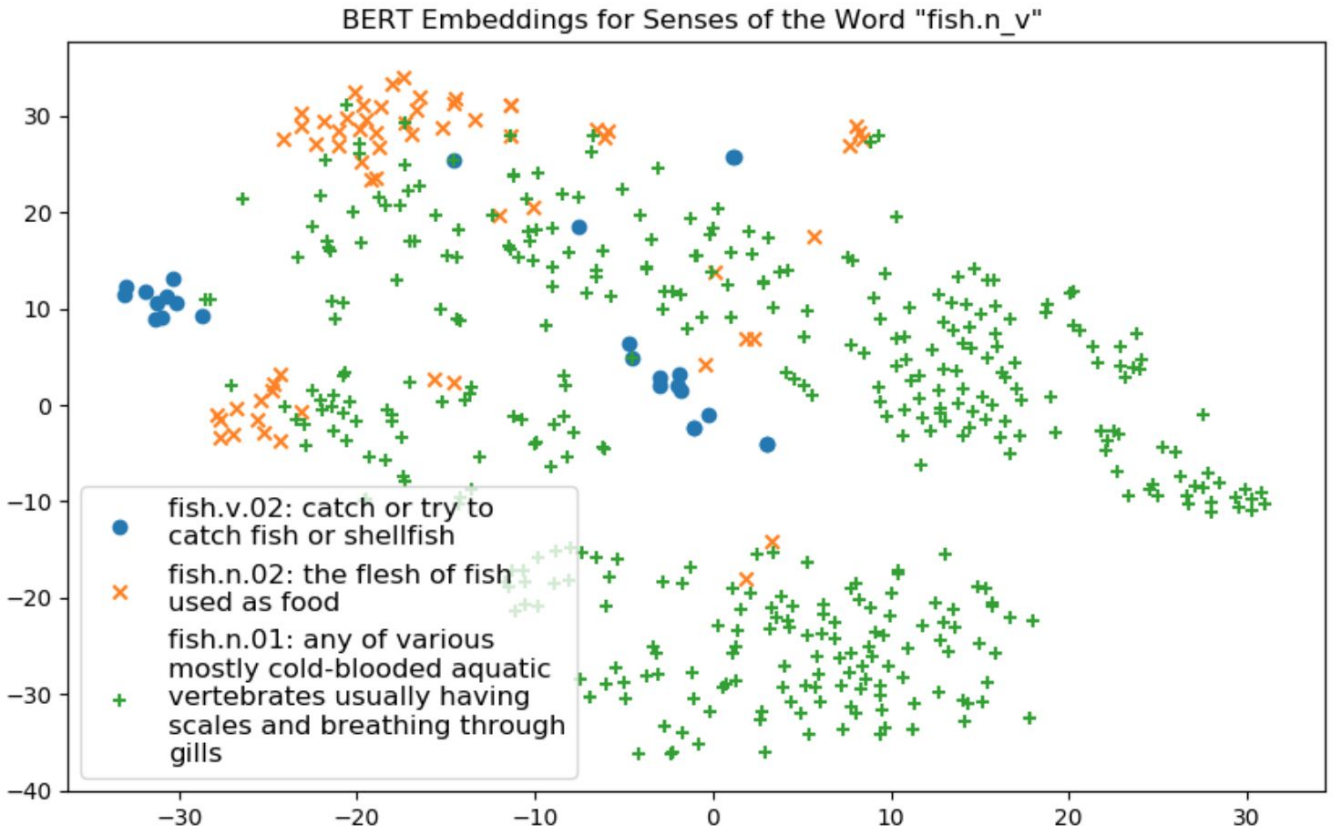




Future work

- word and contextual embeddings such as BERT should be studied for this task
- lemmatization and part-of-speech tagging should also be taken into account when using word and contextual embeddings which lack such information

t-SNE visualization of *fish* (noun, verb) in BERT vector for each occurrence in the SemCor corpus (http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor)



Questions?

elexis european lexicographic infrastructure

Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS