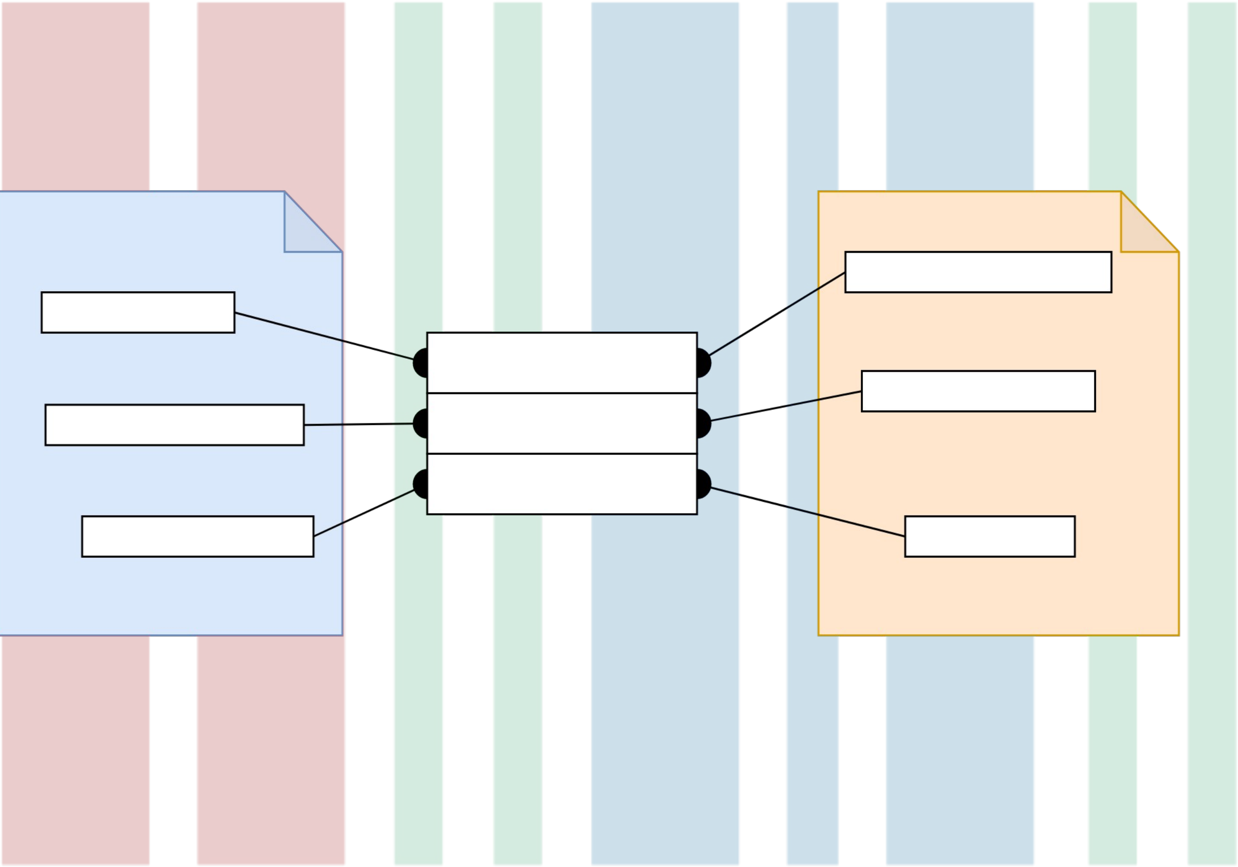


Detecting Cross-Language Plagiarism using Open Knowledge Graphs

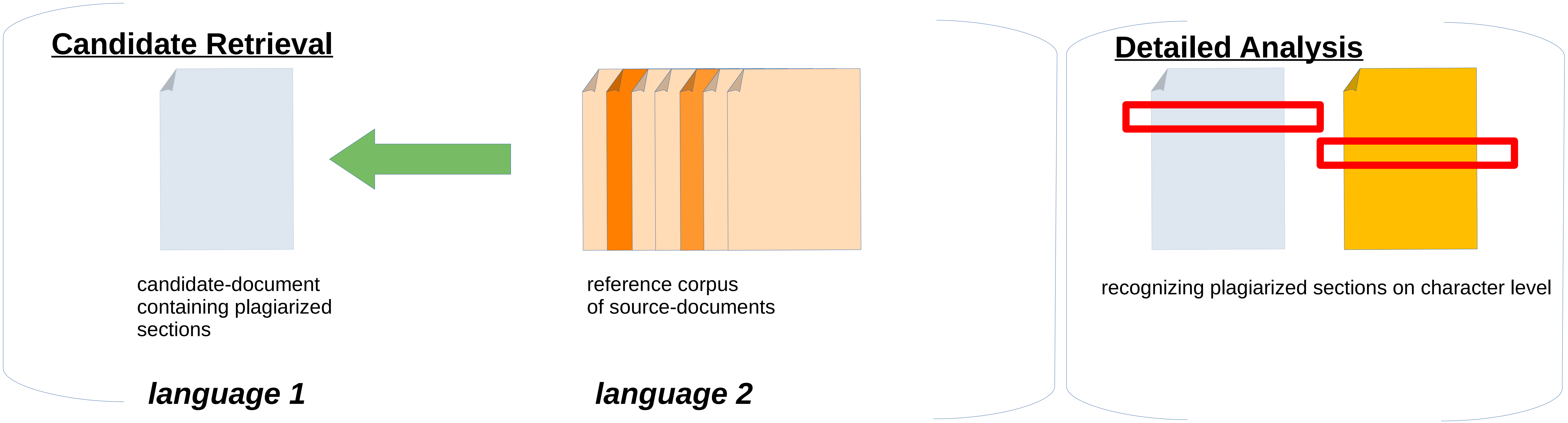
- Presenter: Johannes Stegmüller
- Venue: 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2021) at JCDL 2021 (Online)

Authors: *Johannes Stegmüller*, Fabian Bauer-Marquart*, Norman Meuschke, Moritz Schubotz, Terry Ruas, Bela Gipp*

**contributed equally*

Multilingual Plagiarism Detection

- What is multilingual plagiarism detection?



Candidate Retrieval

candidate-document
containing plagiarized
sections

reference corpus
of source-documents

language 1

language 2

Detailed Analysis

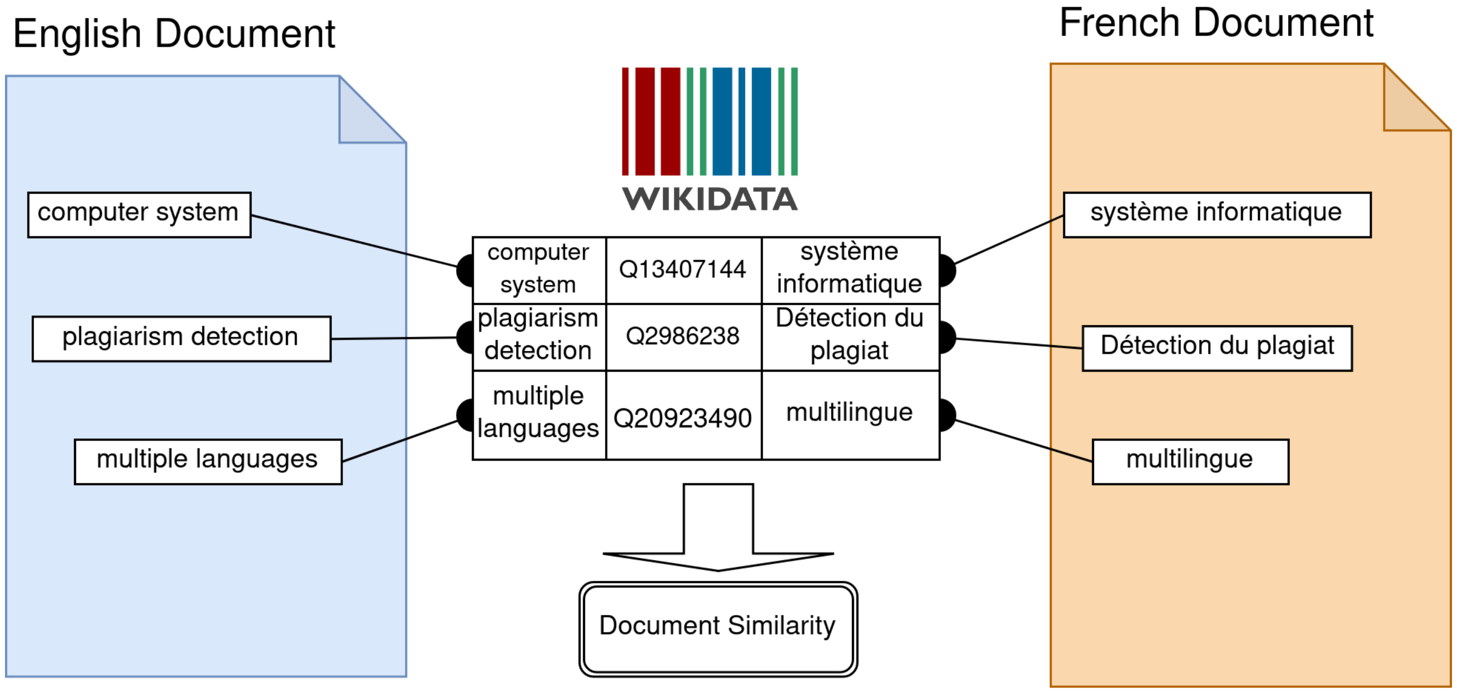
recognizing plagiarized sections on character level

What is CL-OSA ?

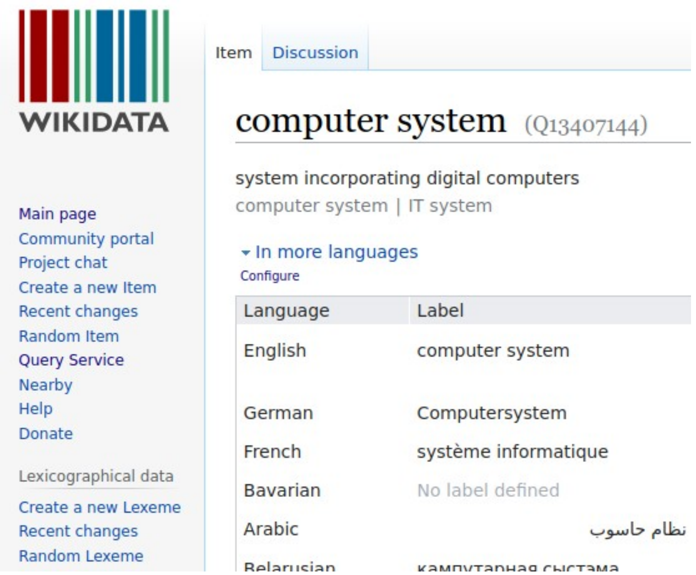
- How does CL-OSA detect multilingual plagiarism ?

English candidate document: “The computer system is used for plagiarism detection in multiple languages.”

French source document: “Le système informatique est utilisé pour la détection de plagiat multilingue.”



computer system	Q13407144	système informatique
plagiarism detection	Q2986238	Détection du plagiat
multiple languages	Q20923490	multilingue



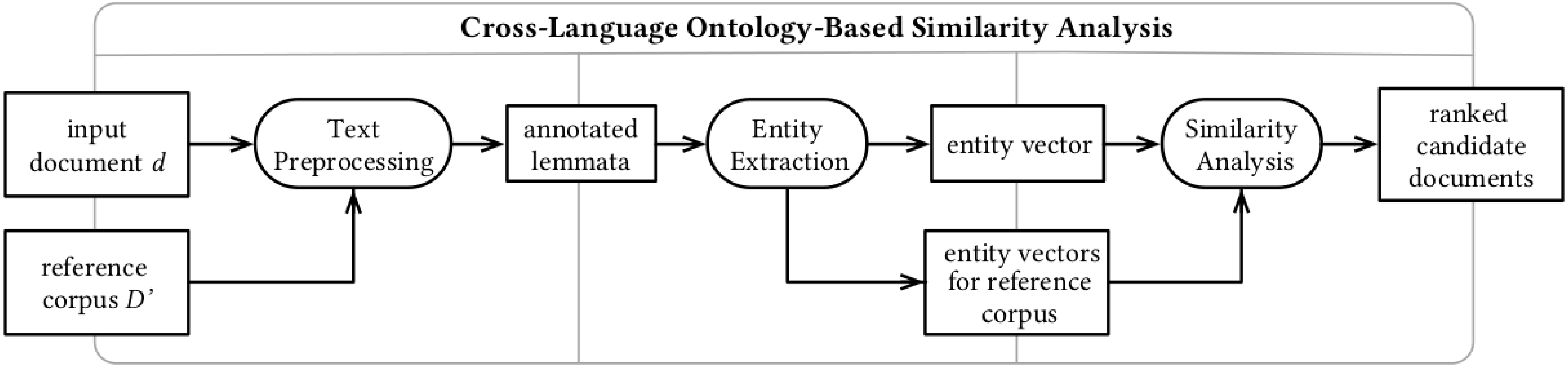
computer system (Q13407144)

system incorporating digital computers

computer system | IT system

Language	Label
English	computer system
German	Computersystem
French	système informatique
Bavarian	No label defined
Arabic	نظام حاسوب
Belarusian	кампутарная сыстэма

Overview of CL-OSA



Text-Preprocessing

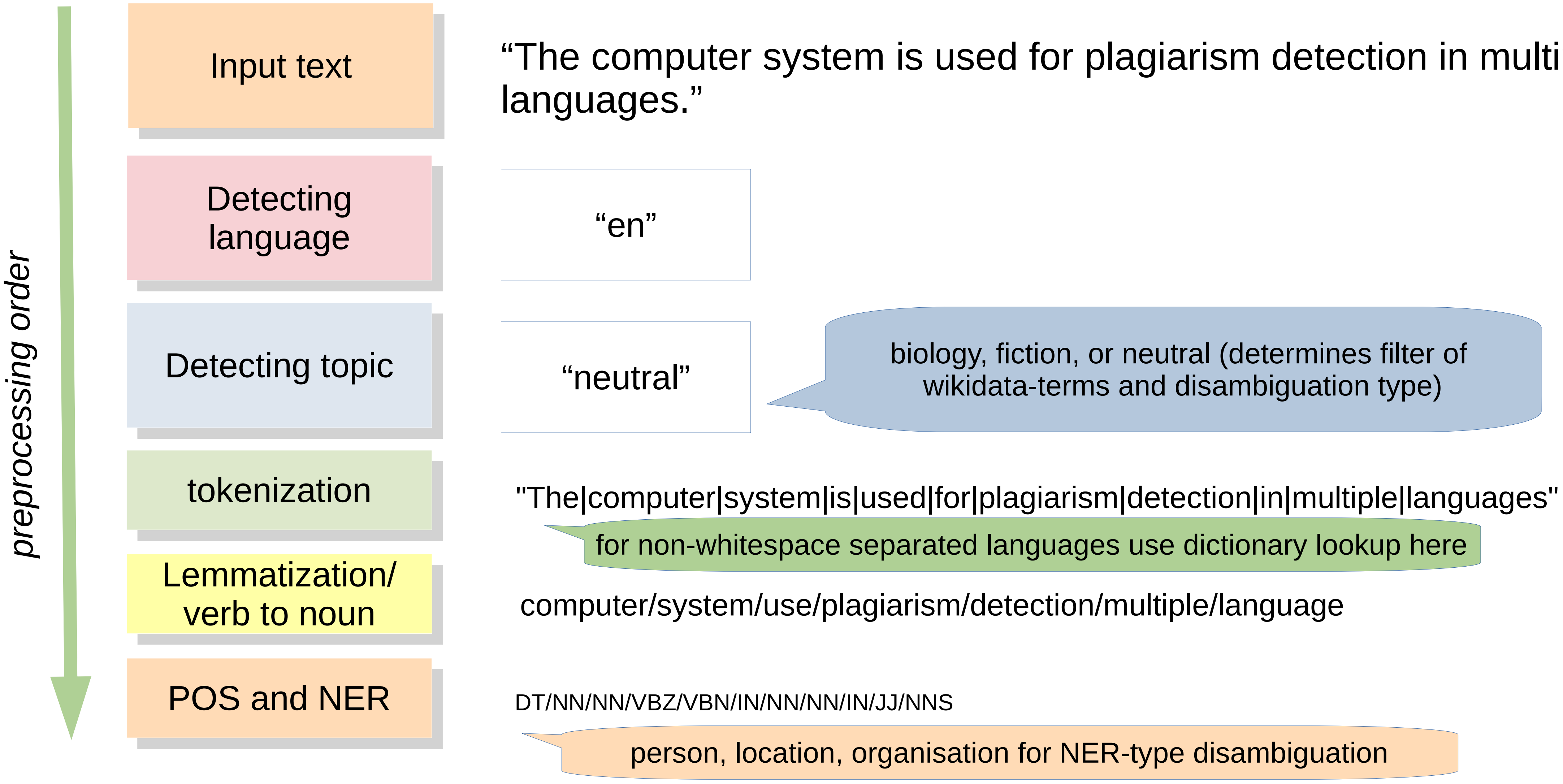
Entity-Extraction

Entity-Disambiguation

Similarity Analysis

Evaluation Method and Results

CL-OSA: Preprocessing



preprocessing order

Input text

“The computer system is used for plagiarism detection in multiple languages.”

Detecting language

“en”

Detecting topic

“neutral”

biology, fiction, or neutral (determines filter of wikidata-terms and disambiguation type)

tokenization

"The|computer|system|is|used|for|plagiarism|detection|in|multiple|languages"

for non-whitespace separated languages use dictionary lookup here

Lemmatization/ verb to noun

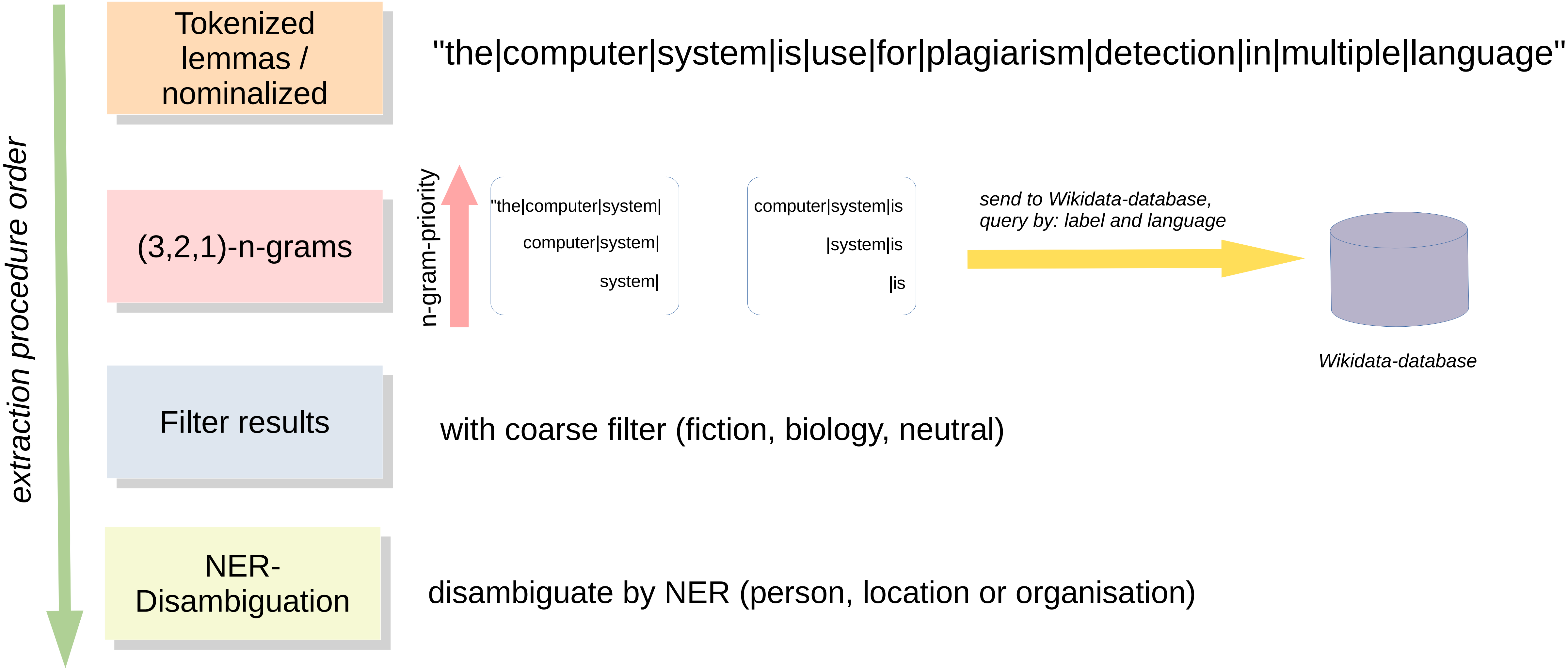
computer/system/use/plagiarism/detection/multiple/language

POS and NER

DT/NN/NN/VBZ/VBN/IN/NN/NN/IN/JJ/NNS

person, location, organisation for NER-type disambiguation

CL-OSA Entity Extraction



extraction procedure order

Tokenized lemmas / nominalized

"the|computer|system|is|use|for|plagiarism|detection|in|multiple|language"

(3,2,1)-n-grams

n-gram-priority

"the|computer|system|
computer|system|
system|

computer|system|is
|system|is
|is

send to Wikidata-database, query by: label and language

Wikidata-database

Filter results

with coarse filter (fiction, biology, neutral)

NER-Disambiguation

disambiguate by NER (person, location or organisation)

CL-OSA Entity Disambiguation

English Document: “**Bass** is a name shared by many species of **fish**.”

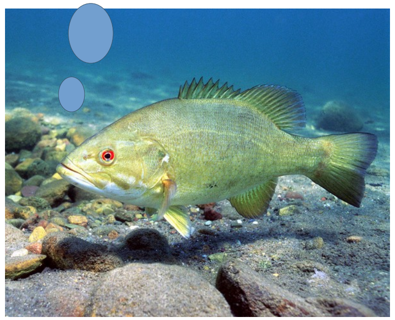
Example Token: “bass”

Label	Entity-id	Ances tor level
“bass”	Q27911	0
“voice”	Q17172850	1
“voice”	Q7390	2
“animal vocalization”	Q97234227	3
…..		

Label	Entity-id	Ances tor level
“bass”	Q1224135	0
“fish”	Q152	1
“aquatic animal”	Q1756633	2
“animal”	Q729	3
…..		

This entity is selected!

The document also contains ‘fish’ so this disambiguation has more weight

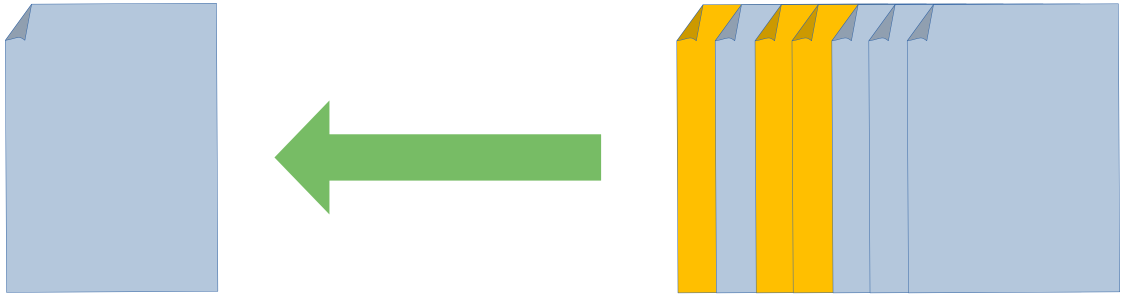


Overview of CL-OSA: Similarity Scoring

- The entities for each document are denoted as vectors
- Scoring similarity is done by applying boolean weight to the term frequency
- The similarity of a candidate to a source vector ($d_{\gg}$ to $d'_{\gg}$) is calculated by cosine-similarity

$$\varphi(\mathbf{d}_{\gg}, \mathbf{d}'_{\gg}) = \frac{\mathbf{d}_{\gg} \cdot \mathbf{d}'_{\gg}}{||\mathbf{d}_{\gg}|| \, ||\mathbf{d}'_{\gg}||}$$

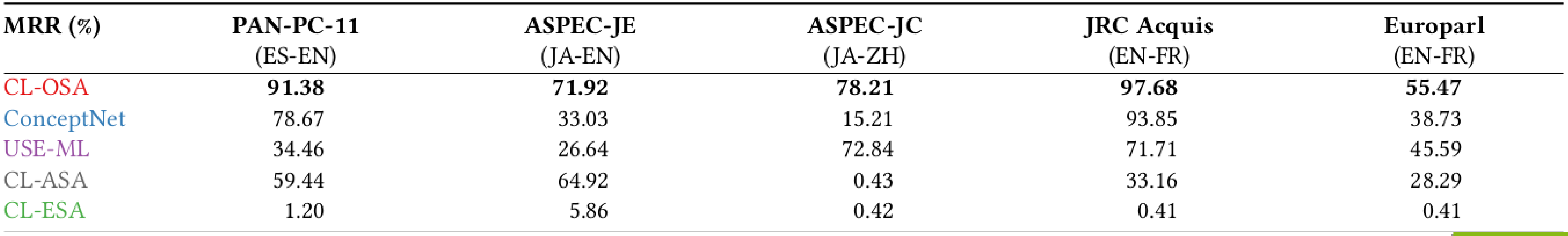
Evaluation: Candidate Retrieval



candidate-document

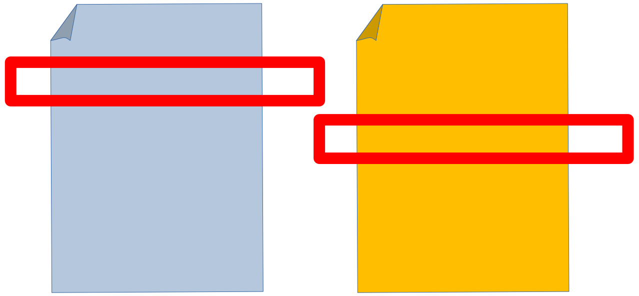
reference corpus of source-documents

- Evaluation of Candidate Retrieval Task
  - MRR: Mean Reciprocal Rank is used as metric
  - Four multilingual corpora in EN, ES, JA, ZH, FR are used



MRR (%)	PAN-PC-11 (ES-EN)	ASPEC-JE (JA-EN)	ASPEC-JC (JA-ZH)	JRC Acquis (EN-FR)	Europarl (EN-FR)
CL-OSA	**91.38**	**71.92**	**78.21**	**97.68**	**55.47**
ConceptNet	78.67	33.03	15.21	93.85	38.73
USE-ML	34.46	26.64	72.84	71.71	45.59
CL-ASA	59.44	64.92	0.43	33.16	28.29
CL-ESA	1.20	5.86	0.42	0.41	0.41

Evaluation: Detailed Analysis

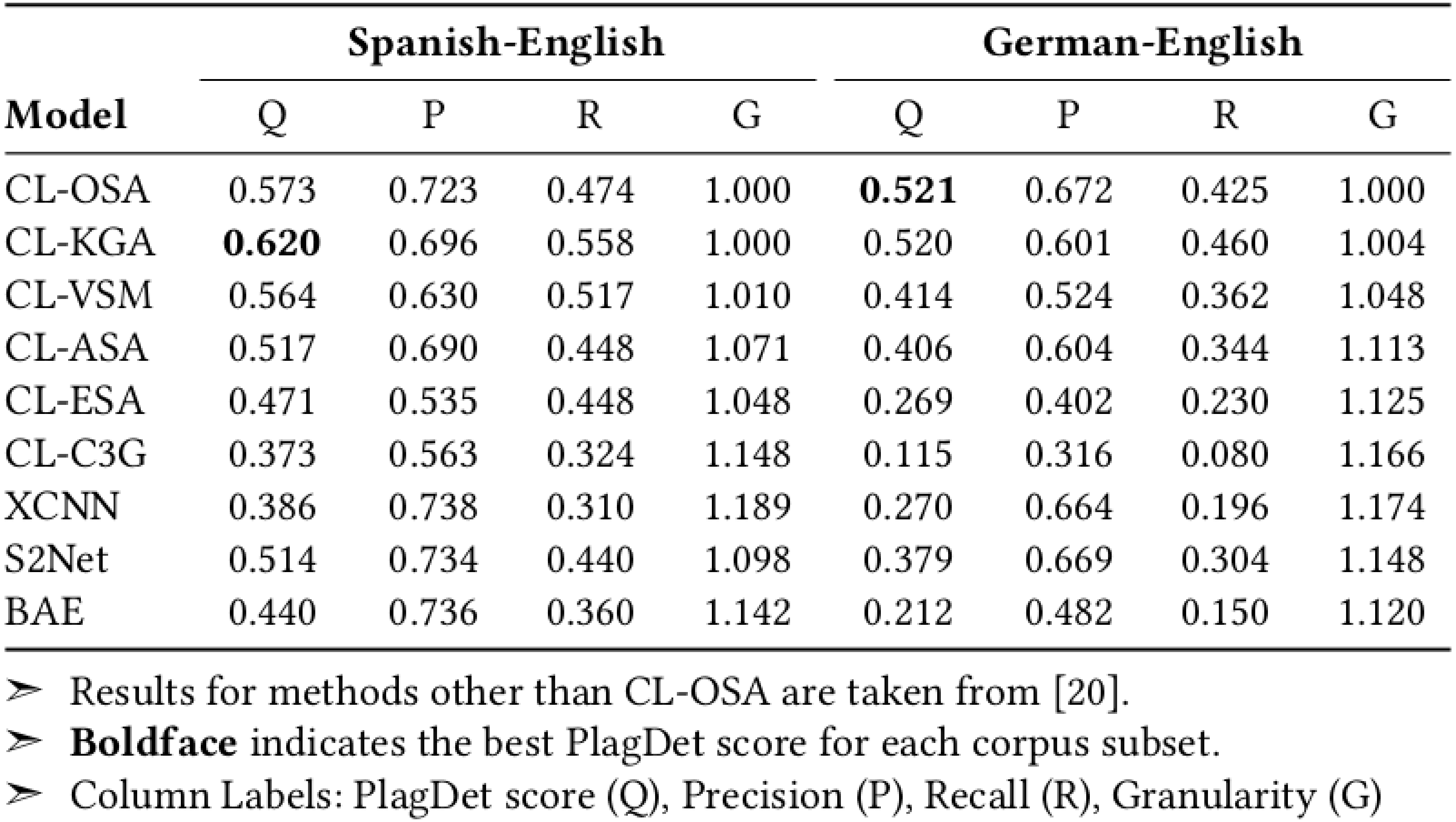


- Evaluation of Detailed Analysis Task
  - The evaluation metric by Salvador et al. [1] (CL-KGA) is used.
  - This uses sliding window with merging algorithm to create consistent plagiarism cases
  - PAN-PC-11 (EN-ES and EN-DE) partitions are used as corpora
  - Metrics are ‘Plagdet’ (Q), Recall (R), Precision (P) and Granularity (G) from PAN-PC evaluation

$$Q = \frac{F_1}{\log_2 (1 + G)},$$

where F_1 represents the harmonic mean of Precision and Recall

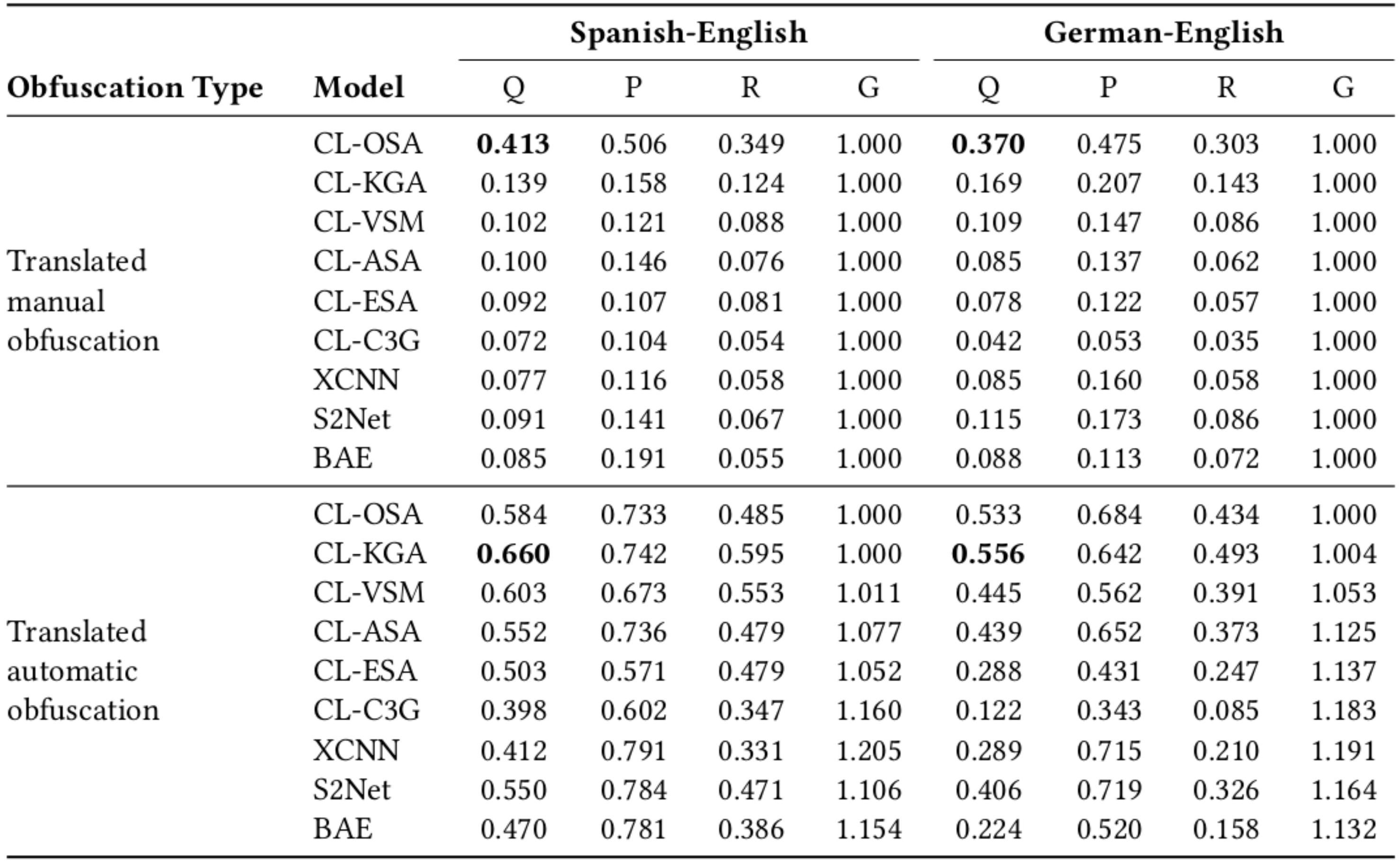
Evaluation: Detailed Analysis Results



	Spanish-English				German-English			
Model	Q	P	R	G	Q	P	R	G
CL-OSA	0.573	0.723	0.474	1.000	**0.521**	0.672	0.425	1.000
CL-KGA	**0.620**	0.696	0.558	1.000	0.520	0.601	0.460	1.004
CL-VSM	0.564	0.630	0.517	1.010	0.414	0.524	0.362	1.048
CL-ASA	0.517	0.690	0.448	1.071	0.406	0.604	0.344	1.113
CL-ESA	0.471	0.535	0.448	1.048	0.269	0.402	0.230	1.125
CL-C3G	0.373	0.563	0.324	1.148	0.115	0.316	0.080	1.166
XCNN	0.386	0.738	0.310	1.189	0.270	0.664	0.196	1.174
S2Net	0.514	0.734	0.440	1.098	0.379	0.669	0.304	1.148
BAE	0.440	0.736	0.360	1.142	0.212	0.482	0.150	1.120

- Results for methods other than CL-OSA are taken from [20].
- **Boldface** indicates the best PlagDet score for each corpus subset.
- Column Labels: PlagDet score (Q), Precision (P), Recall (R), Granularity (G)

Evaluation: Detailed Analysis Results



		Spanish-English				German-English			
Obfuscation Type	**Model**	Q	P	R	G	Q	P	R	G
Translated manual obfuscation	CL-OSA	**0.413**	0.506	0.349	1.000	**0.370**	0.475	0.303	1.000
	CL-KGA	0.139	0.158	0.124	1.000	0.169	0.207	0.143	1.000
	CL-VSM	0.102	0.121	0.088	1.000	0.109	0.147	0.086	1.000
	CL-ASA	0.100	0.146	0.076	1.000	0.085	0.137	0.062	1.000
	CL-ESA	0.092	0.107	0.081	1.000	0.078	0.122	0.057	1.000
	CL-C3G	0.072	0.104	0.054	1.000	0.042	0.053	0.035	1.000
	XCNN	0.077	0.116	0.058	1.000	0.085	0.160	0.058	1.000
	S2Net	0.091	0.141	0.067	1.000	0.115	0.173	0.086	1.000
	BAE	0.085	0.191	0.055	1.000	0.088	0.113	0.072	1.000
Translated automatic obfuscation	CL-OSA	0.584	0.733	0.485	1.000	0.533	0.684	0.434	1.000
	CL-KGA	**0.660**	0.742	0.595	1.000	**0.556**	0.642	0.493	1.004
	CL-VSM	0.603	0.673	0.553	1.011	0.445	0.562	0.391	1.053
	CL-ASA	0.552	0.736	0.479	1.077	0.439	0.652	0.373	1.125
	CL-ESA	0.503	0.571	0.479	1.052	0.288	0.431	0.247	1.137
	CL-C3G	0.398	0.602	0.347	1.160	0.122	0.343	0.085	1.183
	XCNN	0.412	0.791	0.331	1.205	0.289	0.715	0.210	1.191
	S2Net	0.550	0.784	0.471	1.106	0.406	0.719	0.326	1.164
	BAE	0.470	0.781	0.386	1.154	0.224	0.520	0.158	1.132

- Results for methods other than CL-OSA are taken from [20].
- **Boldface** indicates the best PlagDet score for each corpus subset.
- Column Labels: PlagDet score (Q), Precision (P), Recall (R), Granularity (G)

Conclusion and Outlook

- **Benefits of CL-OSA**
  - No machine translation which uses parallel corpora is required
  - It doesn’t require pre trained language models
  - Knowledge-base can be kept up to date, Wikidata license
  - Amount of entities is constantly increasing in most languages

- **Outlook**
  - Investigate performance in Detailed Analysis Task (performance by obfuscation, case-length)
  - Investigate performance in terms of hardware requirements and timings
  - Optimize the weighting scheme for CL-OSA (i.e. TF/IDF instead of binary weights)

References and Sources

[1] PAN-PC-11: Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2011. PAN Plagiarism Corpus 2011 (PAN-PC-11). https://doi.org/10.5281/ZENODO.3250095

[2] CL-KGA and Evaluation Method: Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016.
Cross-Language Plagiarism Detection Over Continuous-Space- and Knowledge Graph-Based Representations of Language. Knowledge-Based Systems 111 (Nov. 2016), 87–99. https://doi.org/10.1016/j.knosys.2016.08.004

[3] Wikidata-Logo https://commons.wikimedia.org/wiki/File:Wikidata-logo-en.svg
[4] Bass on Wikipedia https://www.wikidata.org/wiki/Q1224135#/media/File:Micropterus_dolomieu.jpg

Thank You

Contact:

Johannes Stegmüller

stegmueller@gipplab.org

Twitter: @hyper_node

DATA & KNOWLEDGE ENGINEERING

BERGISCHE UNIVERSITÄT WUPPERTAL