Detecting Cross-Language Plagiarism using Open Knowledge Graphs

Johannes Stegmüller* stegmueller@gipplab.org University of Wuppertal Wuppertal, Germany
Fabian Bauer-Marquart* fabian.marquart@uni-konstanz.de University of Konstanz Konstanz, Germany
Terry Ruas ruas@uni-wuppertal.de University of Wuppertal Wuppertal, Germany
Moritz Schubotz moritz.schubotz@fiz-karlsruhe.de FIZ Karlsruhe Berlin, Germany
Bela Gipp gipp@uni-wuppertal.de University of Wuppertal Wuppertal, Germany

ABSTRACT

Identifying cross-language plagiarism is challenging, especially for distant language pairs and sense-for-sense translations. We introduce the new multilingual retrieval model Cross-Language Ontology-Based Similarity Analysis (CL-OSA) for this task. CL-OSA represents documents as entity vectors obtained from the open knowledge graph Wikidata. Opposed to other methods, CL-OSA does not require computationally expensive machine translation, nor pre-training using comparable or parallel corpora. It reliably disambiguates homonyms and scales to allow its application to Web-scale document collections. We show that CL-OSA outperforms state-of-the-art methods for retrieving candidate documents from five large, topically diverse test corpora that include distant language pairs like Japanese-English. For identifying cross-language plagiarism at the character level, CL-OSA primarily improves the detection of sense-for-sense translations. For these challenging cases, CL-OSA’s performance in terms of the well-established PlagDet score exceeds that of the best competitor by more than factor two. The code and data of our study are openly available.

CCS CONCEPTS

• Information systems → Multilingual and cross-lingual retrieval. Near-duplicate and plagiarism detection.

KEYWORDS

Cross-language plagiarism detection, knowledge graphs, Wikidata

1 INTRODUCTION

Plagiarism is “the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected” [16]. Plagiarism harms scientific discourse, wastes resources, and can unjustifiably benefit the plagiarist if it remains undiscovered [55, p. 22ff]. If researchers revise earlier results in later publications, papers that plagiarized the original findings remain unchanged. Others may spend time and resources trying to replicate such wrong results, or worse, consider them correct and compromise later research or practical applications. Reviewing and sanctioning plagiarized research papers or grant applications often require hundreds of working hours from the reviewers, affected academic institutions, and funding agencies.

The rapid advancement of the Web and information technology have enabled convenient access to vast amounts of information, making plagiarism easier than ever. This development has spurred extensive research on automated methods to identify plagiarized content. Most state-of-the-art plagiarism detection methods analyze lexical, syntactic, and semantic text similarity to identify copied or moderately obfuscated monolingual plagiarism [17].

Detecting cross-language plagiarism remains a significant challenge, despite advances in cross-language information retrieval (CLIR) [17, 55]. Most current cross-language plagiarism detection (CLPD) methods (cf. Section 2) rely on computationally expensive machine translation or learning approaches based on parallel or comparable corpora that are not easily available for many languages. Thus far, few detection methods leverage multilingual knowledge graphs to analyze the deep semantic similarity of documents. This is one of the reasons why current methods can only identify mildly obfuscated cross-language plagiarism reliably [17].

To fill this gap, we propose a new multilingual retrieval model and apply it to CLPD. The main contributions of our work are:

(1) We introduce Cross-Language Ontology-Based Similarity Analysis as a novel CLPD method. CL-OSA identifies the semantic similarity of documents by leveraging multilingual knowledge graphs like Wikidata to extract and compare entities contained in the documents. It models texts as entity vectors and leverages relations between entities for entity disambiguation. CL-OSA is suitable for all topical domains, robust against paraphrasing, and applicable to many close and distant language pairs.

(2) Using documents in Chinese, French, English, Japanese, and Spanish, we show that CL-OSA outperforms state-of-art methods for the two standard sub-tasks in CLPD—candidate retrieval and detailed analysis.

(3) We make our source code and data publicly available.

2 RELATED WORK

Cross-language plagiarism detection is an information retrieval task that methods typically address in two steps [46]. In the candidate retrieval step, the methods use efficient algorithms to retrieve from a large document collection in another language (reference collection) all documents that contain a certain amount of similar content as the

Both lead authors contributed equally to this research and writing the paper.

https://wikidata.org
input document. In the detailed analysis step, the methods perform pairwise comparisons of the input document to each candidate to identify similar segments within the documents at the character level. Hereafter, we summarize CLPD and general cross-language information retrieval approaches relevant to our work.

2.1 Machine Translation

Many CLIR and CLPD methods combine language normalization via machine translation with monolingual similarity analysis [17, 36].

Cross-language Character n-grams (CL-CNG) proposed by McNamie and Mayfield [35] is a vector space retrieval model that uses machine translation to map two documents into a common language, typically English. The method then partitions both documents into character n-grams exclusively consisting of lowercase letters and numbers. CL-CNG computes the cosine measure for the n-gram vectors to determine their similarity. Several studies on CLPD use CL-CNG as a baseline approach, e.g., [4, 18, 20, 46].

Chen et al. [10] combined machine-translation with a vector space model (VSM) for ranked cross-language retrieval of documents in English and Chinese. Their method translates the query using a bilingual dictionary before performing ranked retrieval using the VSM. Their study showed that segmenting Chinese texts is challenging for achieving high retrieval quality when the query is in another language. Franco-Salvador et al. used a similar approach as a baseline in their evaluation [20]. Their Cross-language Vector Space Model (CL-VSM) represents documents in a bilingual form by concatenating tf-idf-weighted vector representations of the original document and its translation obtained using a statistical dictionary. The authors re-weighted the vector representing the translated document using the translation probabilities of words.

Barrón-Cedeño et al. [5] proposed Cross-Language Alignment-based Similarity Analysis (CL-ASA) for the CLPD task. The method uses statistical machine translation based on the IBM alignment model 1 [8]. In a later study performed by the same research group, CL-ASA achieved superior precision over CL-CNG, which achieved the highest recall [4]. CL-ASA is more robust against synonym replacements than CL-CNG because it considers multiple translation candidates and their translation probabilities. However, this approach also causes CL-ASA to be computationally more expensive than CL-CNG. CL-ASA requires computing the similarities between all documents, while CL-CNG is typically implemented using an index, thereby achieving faster query execution.

2.2 Corpus-based Semantics

Corpus-based semantic analysis follows the idea of distributional semantics, i.e., words co-occurring in similar contexts tend to convey similar meaning. Consequently, one assumes that texts with similar word distributions are semantically similar [25]. Word embeddings and Semantic Concept Analysis (SCA) are established corpus-based semantic analysis approaches that researchers applied to CLIR and CLPD, besides many other tasks. The approaches differ in the scope within which they consider co-occurring words.

2.2.1 Word Embeddings. Word embeddings consider the surrounding words to represent a word in a dense, low-dimensional, fixed-size vector space. Words with similar neighboring words should be close to each other in the vector space [6].

Ferrero et al. proposed two CLPD methods based on word embeddings [15]. The first, Cross-Language Conceptual Thesaurus-based Similarity Word Embedding (CL-CTS-WE), represents a word as a bag of words (BOW) consisting of the 10 most similar words according to the embeddings model. The second, Cross-Language Word Embedding Sentence Vector (CL-WES), represents sentences as the sum of the embedding vectors of their constituent words and compares the resulting sentence vectors using the cosine measure. Both methods use Multivec [7] as their pre-trained word embeddings model. Multivec combines word2vec [42], paragraph vectors [30], and bilingual distributed representations [32].

Glavas et al. presented a computationally lightweight method to analyze cross-language similarity for language pairs that lack parallel corpora or named entity recognition [24]. The authors mapped words into a bilingual embedding space by initially creating a monolingual word embedding and then applying a linear function learned from a training corpus.

In our evaluation (cf. Section 4), we use ConceptNet and USE-ML, which are comparable to the methods Ferrero et al. [15] and Glavas et al. [24] proposed, but rely on more recent pre-trained word embedding models. Different from Ferrero et al. [15], we represent the documents in our datasets as the average of their constituent word embeddings from the pre-trained models.

ConceptNet-Numberbatch (ConceptNet) [51] uses traditional word embeddings, such as word2vec [42] and GloVe [45], and the lexical information in ConceptNet2 to derive its semantic vectors.

The Universal Sentence Encoder-Multilingual (USE-ML) [57] offers two architectures to derive its vectors. One is inspired by the Transformer architecture [54] and the other uses Deep Average Networks (DAN) [28].

2.2.2 Semantic Concept Analysis. Semantic concept analysis extends the distributional semantics idea to an external corpus.

Potthast et al. [48] introduced Cross-Language Explicit Semantic Analysis (CL-ESA) as a multilingual generalization of the semantic retrieval model Explicit Semantic Analysis (ESA) proposed by Gabrilovich and Markovitch [22]. ESA and CL-ESA represent documents as vectors in a high-dimensional vector space of semantic concepts, which are explicitly encoded topics in a knowledge base corpus. CL-ESA uses a concept-aligned comparable corpus available in multiple languages. Specifically, Potthast et al. used Wikipedia articles and considered each article available in multiple languages to represent one concept. Each dimension of a document vector represents the tf-idf similarity of the document to one of the concepts. The similarity of document vectors is typically quantified using the cosine measure [22, 48]. Meuschke et al. extended CL-ESA by also considering the order in which concepts occur in the text to identify potentially suspicious patterns [39].

The evaluations of Potthast et al. [48] showed that CL-ESA performs best if the concept space has 100,000 or more dimensions, i.e., if at least 100,000 Wikipedia articles are considered. In this case, CL-ESA achieved a recall above 0.90 for the JRC-Acquis corpus [52]. However, such high dimensionality is computationally expensive. Therefore, Potthast et al. advised that: "if high retrieval speed or a high multilinguality is desired, documents should be represented as 1 000-dimensional concept vectors. At a lower dimension the

2http://conceptnet.io/
retrieval quality deteriorates significantly. A reasonable trade-off between retrieval quality and runtime is achieved for a concept space dimensionality between 1,000 and 10,000.” [48, p. 526f.].

Despite limitations in dimensionality as proposed by Potthast et al., CL-ESA is computationally more expensive than CL-ASA because it requires computing the similarity of the input document to all concepts followed by calculating the similarity of all document vector pairs as is the case for CL-ASA.

### 2.3 Non-textual Content Analysis

Researchers proposed analyzing non-textual content features to overcome the ambiguities of natural language and complement text analysis approaches to improve the detection of concealed plagiarism forms, such as translations. The investigated content elements include academic citations [23, 34, 40], images [11, 12, 37], and mathematical content [38, 41, 50].

### 2.4 Knowledge-based Semantics

Knowledge-based semantic analysis approaches, such as the one we propose, use entities encoded in semantic networks, such as thesauri, ontologies, and knowledge graphs.

Cross-Language Knowledge Graph Analysis (CL-KGA) is a CLPD method proposed by Franco-Salvador et al. and most related to our work [18]. CL-KGA uses sub-graphs of the multilingual semantic network BabelNet\(^\text{1}\) to represent text segments. Specifically, CL-KGA splits documents into segments using a five-sentences-long sliding window with a two-sentences step width, lemmatizes the segments, and performs part-of-speech tagging. By mapping terms in the preprocessed text segments to BabelNet, CL-KGA obtains the subgraph of BabelNet used to represent the segments. Franco-Salvador et al. proposed a graph-based similarity measure that considers the similarity of entities and their relations to compare the entity representations. The authors improved the weighting function in subsequent publications and combined the graph-based representation with neural text representations [19–21].

### 2.5 Neural Networks

Neural text representations and language models have significantly advanced the state of the art for many NLP and CLIR tasks. A comprehensive review would exceed the scope of this paper; therefore, we restrict our description to successful neural CLIR methods, which we use as baselines for our evaluation in Section 4.

The External-data Composition Neural Network (XCNN) is a cross-language continuous space model created by a composition function on top of a deep neural network. In difference to similar approaches, XCNN can be initialized with monolingual data and extended with at least a small set of parallel data. This feature of the network is especially useful for low-resource languages [27].

The Siamese Neural Network (S2Net) trains two identical networks concurrently with parallel data that has to be annotated with a similarity score [20]. The network lends itself for similarity learning in a bilingual use case, where each network reflects data in one of the two languages. For each text input, the networks emit feature vectors representing the input in the respective languages, which can then be compared using the cosine similarity.

\(^{1}\)https://babelnet.org

---

**Figure 1: Number of Wikidata entities (red) and Wikipedia articles (green) per language as of September 2021.**

Bilingual Autoencoders (BAE) are trained using bag-of-words representations of multiple sentences from parallel corpora as input. From the BOW representation in a source language, the encoder creates a BOW representation in the target language. During training, the encoder is optimized by minimizing the reconstruction error between the created representation from the source language and the original target representation [26].

### 2.6 Research Gap

Most CLPD methods rely on machine translation [4, 5, 10, 35] or representations trained using parallel [20, 27] or comparable corpora [27, 48]. These approaches depend on lexical and syntactical similarity and topical homogeneity of the documents in different languages. There is a need for CLPD methods that can analyze a wide variety of topics across academic disciplines. The use of knowledge graphs has been shown to benefit the analysis of semantic document similarity in the monolingual and cross-language setting. However, few studies have investigated the use of knowledge graphs for cross-language plagiarism detection [18–21]. We extend and improve upon this prior work, as we explain hereafter.

### 3 PROPOSED METHOD

Cross-Language Ontology-based Similarity Analysis is a multilingual retrieval model derived from a knowledge graph that includes ontological relations. The method constructs language-independent, semantically-enhanced entity vectors that not only include entities present in the modeled documents but also entities that are hierarchically linked by subclass of and instance of relations.

Three reasons governed our decision to use the open knowledge graph Wikidata to realize CL-OSA, instead of an open encyclopedia like Wikipedia used, e.g., by ESA and CL-ESA, or BabelNet used by CL-KGA. First, the number of entities in Wikidata greatly exceeds the number of Wikipedia articles. ESA and CL-ESA use Wikipedia articles as concepts, which limits their representation. Figure 1 shows the number of Wikidata entities per language (queried from an official JSON dump dated September 2021). There are more than twelve times as many Wikidata entities available for English as there are Wikipedia articles. For Spanish and French, the number of Wikidata entities exceeds the number of Wikipedia articles by factors between six and eighth. For Chinese the factor is four and for Japanese the factor is two.
Second, while Wikipedia exclusively contains cross-references between articles, Wikidata includes property links that express relationships, such as instance of, subclass of, color, or part of. Therefore, Wikidata offers a wider range of typed relationships that are readily accessible for automated processing.

Third, Wikidata offers public domain data with no restrictions on its use. BabelNet, for example, imposes fees for commercial use.\(^4\)

As Figure 1 shows, both the number of Wikipedia articles and Wikidata entities differs greatly between languages. Fewer entities can reduce the detection effectiveness of knowledge-graph-based detection methods like CL-OSA for the respective language. However, as we show in our evaluation (cf. Section 4) CL-OSA and comparable methods already achieve good results for languages with fewer entities, like Japanese and Chinese. Moreover, knowledge bases like Wikidata grow continuously, especially due to significant advances in automated entity extraction and linking.\(^5\)

### 3.1 Retrieval Model

Figure 2 illustrates the three-step process consisting of Text Preprocessing, Entity Extraction, and Similarity Analysis. CL-OSA follows for representing documents as language-independent entity vectors and using them for ranked cross-language retrieval. Hereafter, we formalize the process and present each of its three steps in detail.

Let \( D \) be the set of (suspicious) input documents and \( D' \) be the set of documents in the reference collection, i.e., potential sources for content. The goal is to determine the similarity between an input document \( d \in D \) and a candidate document \( d' \in D' \) denoted as \( \varphi(d, d') \). CL-OSA represents a document \( d = w_1w_2\ldots w_k \) written in language \( L_1 \in \mathcal{L} \) as an hierarchy-enhanced entity vector \( d_s \).

\[
d_s = (q_1, \ldots, q_n).
\]

The elements of \( d_s \) are entities of the knowledge graph occurring in the document and the ontological ancestors of these entities.

A multilingual knowledge graph \( Q \) such as Wikidata is a set of entities \( q \), defined as a tuple \( q = (\Sigma, \mathcal{A}, \Delta, \rightarrow) \) where

- \( \Sigma \) is a set of labels \( \{l_1, \ldots, l_m\} \),
- \( \mathcal{A} \) is a set of alias sets \( \{A_1, \ldots, A_m\} \),
- \( \Delta \) is a set of descriptions \( \{\delta_1, \ldots, \delta_m\} \), and
- \( \rightarrow \subseteq Q \times P \times 2^Q \) is a property relation.

\( P \subseteq Q \), \( \mathcal{L} = \{L_1, \ldots, L_m\} \), and \( T \) denote sets of properties, languages, and topics respectively.

For example, the entity query has the English label query, the English alias database query, the description precise request for information retrieval, and the subclass of property that maps query to the entity for information request, all of which is valuable information when relating entities to each other.

For convenience, we denote property relations as \( q \rightarrow p q' \) instead of \( (q, p, q') \in \rightarrow \) using the following notations:

- \( p_\rightarrow \) denotes the instance of relation
- \( p_\subset \) denotes the subclass of relation
- \( p_\land \), the combined property relation such that \( q \rightarrow p_\land q' = q \rightarrow p_\rightarrow q' \lor q \rightarrow p_\rightarrow q' \)
- \( \rightarrow^* \) denotes the transitive property relation, which we define as \( q \rightarrow^* q' = q \rightarrow^* q_1 \rightarrow^* \cdots \rightarrow^* p \rightarrow q' \)

The transitive property relation lets child entities inherit their parents’ properties. For example, the entity pi is an instance of mathematical constant, which itself is a subclass of number. Therefore, pi is transitively also an instance of number.

### 3.2 Text Preprocessing

The goal of the preprocessing is to avoid typical issues that arise if statistical machine translation or alignment-based retrieval models are employed for the language normalization step of the CLPD process. Translations produced using these state-of-the-art methods often exhibit grammatical errors, syntactical differences, and sub-optimal wording if sufficient domain-specific training data is missing. The availability of training data is often problematic for highly domain-specific texts, such as scientific, technical, and professional documents in languages other than English.

As an example to illustrate these challenges, we use the introductory sentence of the article Volkswirtschaftslehre (macroeconomics) in the German Wikipedia\(^6\):

Die Volkswirtschaftslehre (auch Nationalökonomie oder wirtschaftliche Staatswissenschaften kurz VWL) ist ein Teilgebiet der Wirtschaftswissenschaft.

Translating this sentence to English using Google Translate introduces ambiguity by mapping several words that have a clear distinction in German to “economics”:

Economics (also economics or economic political science short VWL) is a branch of economics.

---

\(^4\)https://babelnet.org/full-license | https://babelscape.com/wordatlas

\(^6\)https://de.wikipedia.org/wiki/Volkswirtschaftslehre
To avoid these issues, CL-OBA first categorizes the document by topic and then extracts terms from the document corresponding to Wikidata entities having the most ancestors. The idea is that in this way, CL-OBA selects entities that are most specific because they have many superordinate entities and are thus representative of the document’s topic. Including superordinate concepts also helps to disambiguate entities in the text, since the included superordinate (potentially ambiguous) entities likely also occur in the document.

Specifically, CL-OBA performs the following steps to prepare documents for being represented as entity vectors.

**Language Detection.** In case the language of a document is unknown, CL-OBA infers it using a language detector. The detector uses the vector space model to compare a document’s n-grams to a pre-trained set of n-grams from various multilingual comparable corpora, and selects the most probable language.

**Topic Detection.** To roughly identify the topical domain of a document, we trained a Bayesian classifier using approximately 100 bag-of-words representations of Wikipedia articles that we hand-assigned to topics relevant for our entity extraction and disambiguation approaches, i.e., biology, fiction, and neutral. Classifying a document yields the topic that has the highest word overlap with the topic-specific articles on which the classifier has been trained. Using only these three categories was sufficient because they help to disregard the largest amount of Wikidata entities that are irrelevant for the extraction task, e.g., movie titles for works not related to fiction, or genes and proteins for works not related to biology.

**Tokenization and Word Segmentation.** Depending on the document’s language, CL-OBA performs tokenization or word segmentation to split the text into a token sequence. Our method uses simple tokenization for white-space separated languages, such as English, Korean, or French, and employs more sophisticated methods, such as a dictionary lookup, for languages lacking a word delimiter, e.g., Chinese or Japanese. The tokenization step keeps stop-words, but strips punctuation.

**Lemmatization.** To exploit entity labels and aliases, which Wikidata contains in their base forms, our method lemmatizes derived and inflected tokens. Additionally, it uses WordNet for mapping verbs to nouns. A fallback procedure if no lemmatizer is available for a language would be using a rule-based stemmer, although this could introduce ambiguity. Therefore, we did not employ stemming. Skipping lemmatization is safe for languages that lack inflection or for which the tokenization step performs inflection removal, e.g., Chinese and Japanese.

**Named Entity Recognition.** To reduce the ambiguity of tokens, CL-OBA performs part-of-speech (POS) tagging and named entity (NE) recognition to annotate the lemmata with POS and proper noun information, such as location, human, or organization.

For tokenization, lemmatization, and named entity extraction, we use Stanford CoreNLP for European languages and Kuromoji for Chinese and Japanese.

### 3.3 Entity Extraction

CL-OBA maps every lemma n-gram with \( n \in \{1, 2, 3\} \) to entity candidates, which it obtains by querying the n-grams to the labels and aliases of the knowledge graph. We apply a coarse filter to disregard entities that are likely irrelevant depending on the document’s topic. Specifically, we require that entity candidates for documents in the category fiction are a subclass of an instance of creative work. For documents in the category biology, entity candidates must be either a subclass of or an instance of gene. Entity candidates originating from longer lemma n-grams take precedence over shorter sequences. Typically, CL-OBA retrieves multiple entity candidates, which it disambiguates as described in the following section.

### 3.4 Entity Disambiguation

CL-OBA disambiguates entity candidates to a single entity using a combination of manually devised filters and mappings of topics to named entities. This procedure removes entities:

- if their original token has a POS tag related to punctuation, prepositions, symbols, markers, or personal pronouns
- that represent a stop-word such as “and”
- that represent a Han character in case of Chinese or Japanese
- that represent a Wikimedia disambiguation page
- that are a subclass of natural number and have numeric labels
- that are not an instance of their named entity types, e.g., human, location, and organization

### Table 1: Entity extraction and disambiguation for a text fragment from an editorial in the English Financial Times.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS</th>
<th>NE</th>
<th>Entity</th>
<th>English label ( l_{ea} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>NNP</td>
<td>LOC</td>
<td>$q_{30}$</td>
<td>United States of America</td>
</tr>
<tr>
<td>tax</td>
<td>NN</td>
<td>O</td>
<td>$q_{161}$</td>
<td>tax</td>
</tr>
<tr>
<td>authorities</td>
<td>NNS</td>
<td>O</td>
<td>$q_{134354}$</td>
<td>authority</td>
</tr>
<tr>
<td>teeth</td>
<td>NNS</td>
<td>O</td>
<td>$q_{553879}$</td>
<td>tooth [heraldic]</td>
</tr>
<tr>
<td>battle</td>
<td>NN</td>
<td>O</td>
<td>$q_{1931249}$</td>
<td>battle [medieval]</td>
</tr>
<tr>
<td>politicians</td>
<td>NNS</td>
<td>O</td>
<td>$q_{82995}$</td>
<td>politician</td>
</tr>
<tr>
<td>Internal Revenue Service</td>
<td>NNP</td>
<td>ORG</td>
<td>$q_{973387}$</td>
<td>Internal Revenue Service</td>
</tr>
<tr>
<td>appears be</td>
<td>VBZ</td>
<td>O</td>
<td>$q_{3620816}$</td>
<td>appearance</td>
</tr>
<tr>
<td>stance</td>
<td>NN</td>
<td>O</td>
<td>$q_{17292}$</td>
<td>stance [martial arts]</td>
</tr>
<tr>
<td>international</td>
<td>JJ</td>
<td>O</td>
<td>$q_{1017212}$</td>
<td>rest of the world</td>
</tr>
<tr>
<td>leaves</td>
<td>VBZ</td>
<td>O</td>
<td>$q_{2479450}$</td>
<td>annual leave</td>
</tr>
<tr>
<td>taxpayers</td>
<td>NNS</td>
<td>O</td>
<td>$q_{532197}$</td>
<td>leave of absence</td>
</tr>
</tbody>
</table>

6https://github.com/optimizae/language-detector
7https://www.atilika.org
Additionally, we exploit the entity hierarchy by disambiguating to the entity candidate that has the most ontological ancestors contained in the text surrounding the entity candidate.

Table 1 illustrates the extraction and disambiguation of entities for the following text fragment taken from the ECCE corpus, which contains Financial Times editorials in English and Chinese [56]:

*US tax authorities are finally finding their teeth. After a long battle with politicians, the Internal Revenue Service appears to be toughening its stance on international tax arbitrage that leaves taxpayers short-changed.*

Table 1 shows the Wikidata entity candidates for every lemma-POS-NE triple extracted from the text. The table omits entities without exit entity candidates for brevity. The final entities after the disambiguation step are underlined.

### 3.5 Similarity Analysis

In the final step, CL-OSA constructs the hierarchy-enhanced entity vector \( \mathbf{d}_q \) by taking all entities \( q \) from the bag-of-entities and applying boolean weighting, i.e., assigning a 1 for entities that occur in the document, and a 0 otherwise. Weighting using the raw term frequency yielded worse results in our experiments.

For example, when analyzing our example sentence from the German Wikipedia article *Volkswirtschaftslehre* (cf. page 4), we obtain the entities *general economics* and *economics* for the German text. For the English text, we only obtain *economics*. However, *general economics* is part of *economics*, and both are transitively instances of *branch of science*. Therefore, as these entities are included in the vector, yet are weighted inversely proportional to their graph-based distance, the similarity score of both sentences increases without introducing too many commonalities.

Furthermore, CL-OSA leverages the relation \( p \) by adding all entities \( q_a \) to \( \mathbf{d}_q \) if they satisfy \( p^m(q_a) = q_a \) for any \( m \in \{1, 2, 3\} \) and assigns the weight \((m + 1)^{-2}\). That is, CL-OSA adds the ancestors of an entity to the vector and assigns an exponentially decreasing weight (inverse quadratic growth) the more distant the ancestors are. Thus, first-level ancestors get a weight of \(1/2^2\), second-level ancestors \(1/3^2\), and so forth. We derived this weighting from the similarity measure by Li et al. [31], which has been shown to reflect semantic similarity in graphs.

CL-OSA compares the resulting vector \( \mathbf{d}_q \) to all vectors \( \mathbf{d}_d \) in the reference collection \( D' \) by computing the cosine similarity

\[
\varphi(d_q, d_d') = \frac{d_q \cdot d_d'}{|d_q||d_d'|}
\]

and uses the scores to rank all reference documents \( d' \in D' \) in decreasing order of their similarity to document \( d \).

### 4 EVALUATION

We evaluate CL-OSA’s performance for the candidate retrieval and detailed analysis tasks of the CLPD process using two distinct experiments, which we present in Section 4.1 and Section 4.2.

The candidate retrieval experiment focuses on covering a wide range of language pairs and corpora. We exclusively include in this evaluation detection methods for which source code or sufficient details for re-implementing the methods are available. For some state-of-the-art methods like CL-KGA, this is not the case, which is why we did not include them in this experiment.

The detailed analysis experiment focuses on comparing CL-OSA to state-of-the-art detection methods, some of which are not available as source code and too complex to be re-implemented. Therefore, we evaluate CL-OSA according to the protocol used in a prior study and compare our results to those reported in this study [20]. The data and source code of our experiments are available at

[https://doi.org/10.5281/zenodo.5159398](https://doi.org/10.5281/zenodo.5159398)

### 4.1 Candidate Retrieval Evaluation

In this evaluation, we compare CL-OSA’s effectiveness in retrieving documents from five multilingual parallel corpora to four state-of-the-art CLPD methods.

#### 4.1.1 Datasets

Using random sampling, we selected 2,000 aligned documents from each of the following five corpora:

**PAN-PC-11 Plagiarism Corpus** [47]. The corpus contains instances of simulated monolingual and cross-language plagiarism that were used for evaluating plagiarism detection methods as part of the workshop series *Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (PAN). Most of the 26,939 documents in the corpus were created by extracting text from openly available books. The documents are partially interspersed with instances of simulated plagiarism that were created and obfuscated automatically or by crowdsourced workers. For the candidate retrieval evaluation, we exclusively sampled test cases from the 2,921 *Spanish-English* aligned document pairs in the corpus.

**ASPEC-JE** [43]. This subset of the Asian Scientific Paper Excerpt Corpus (ASPEC) contains abstracts of approx. two million research papers that were translated manually from *Japanese* to *English*.

**ASPEC-JC** [43]. This subset of the ASPEC corpus contains abstracts and paragraphs from the main text of research papers that were translated manually from *Japanese* to *Chinese*.

**JRC-Acquis** [52]. The corpus consists of legislative texts in 22 languages, which the European Union’s Joint Research Centre (JRC) selected from the cumulative body of EU laws (the so called *Acquis communautaire*). We sampled our test cases from the 10,000 document pairs in the *English-French* subset of the corpus.

**Europarl** [29]. The corpus contains transcripts of European Parliament proceedings in 21 European languages. As for JRC-Acquis, we exclusively sampled test cases from the 9,443 document pairs in the *English-French* subset of the corpus.

We used the subsets of the PAN-PC-11, JRC-Acquis and Europarl corpora that Ferrero et al. [14] pre-selected and provided for the evaluation of cross-language similarity detection methods.

Except for PAN-PC-11, all corpora contain exactly one relevant item for each query. In the PAN-PC-11 corpus, plagiarized text fragments can originate from several source documents.

#### 4.1.2 CLPD Methods

We compare CL-OSA to these methods:

**CL-ASA** implemented according to Potthast et al. [46, p. 9]. For European languages, we derived the translation probabilities from the dictionaries provided by Aker et al. [1]. For the ASPEC corpora (EN-JA and JA-ZH), we used the program *pialign* by Neubig et al.
Table 2: Mean Reciprocal Rank, Recall at Rank, and Average Recall at Rank scores for candidate retrieval.

<table>
<thead>
<tr>
<th>MRR (%)</th>
<th>PAN-PC-11</th>
<th>ASPEC-JE</th>
<th>ASPEC-JC</th>
<th>JRC Acquis</th>
<th>Europarl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(ES-EN)</td>
<td>(JA-EN)</td>
<td>(JA-ZH)</td>
<td>(EN-FR)</td>
<td>(EN-FR)</td>
</tr>
<tr>
<td>CL-OSA</td>
<td>91.38</td>
<td>71.92</td>
<td>78.21</td>
<td>97.68</td>
<td>55.47</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>78.67</td>
<td>33.03</td>
<td>15.21</td>
<td>93.85</td>
<td>38.73</td>
</tr>
<tr>
<td>USE-ML</td>
<td>34.46</td>
<td>26.64</td>
<td>72.84</td>
<td>71.71</td>
<td>45.59</td>
</tr>
<tr>
<td>CL-ASA</td>
<td>59.44</td>
<td>64.92</td>
<td>0.43</td>
<td>33.16</td>
<td>28.29</td>
</tr>
<tr>
<td>CL-ESA</td>
<td>1.20</td>
<td>5.86</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>R@k (%)</td>
<td>R@1</td>
<td>R@2</td>
<td>R@3</td>
<td>R@1</td>
<td>R@2</td>
</tr>
<tr>
<td>CL-OSA</td>
<td>89.65</td>
<td>91.50</td>
<td>92.75</td>
<td>65.95</td>
<td>75.20</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>71.70</td>
<td>80.15</td>
<td>83.65</td>
<td>22.95</td>
<td>35.60</td>
</tr>
<tr>
<td>USE-ML</td>
<td>27.30</td>
<td>33.90</td>
<td>37.50</td>
<td>19.40</td>
<td>24.40</td>
</tr>
<tr>
<td>CL-ASA</td>
<td>53.20</td>
<td>60.15</td>
<td>63.35</td>
<td>56.45</td>
<td>64.90</td>
</tr>
<tr>
<td>CL-ESA</td>
<td>0.45</td>
<td>0.65</td>
<td>0.90</td>
<td>3.15</td>
<td>4.55</td>
</tr>
</tbody>
</table>

Average Recall at Rank

4.1.3 Performance Measures. Recall and the size of the candidate set are essential performance indicators for the candidate retrieval stage. A method’s recall, i.e., which percentage of all source documents the method retrieves among the candidates, is critical because failing to retrieve a source prohibits detecting content that originates from that source in the subsequent detailed analysis stage. The number of candidates necessary to achieve sufficient recall strongly influences the computational effort required for the analysis.

Therefore, we assess the methods’ effectiveness for retrieving candidate documents and ranking them highly (thus enabling small candidate sets) in terms of the Recall at Rank (R@k) and Average Recall at Rank (ARR) measures. For easier comparability of the methods, we report the Mean Reciprocal Rank (MRR) as a single measure, quantifying a method’s overall ranking performance.

4.1.4 Results Candidate Retrieval. Table 2 shows the results for the candidate retrieval task. CL-OSA outperforms the other methods for all corpora, which indicates our method is least affected by the diverse topical domains of the corpora and the lexical and syntactic differences of the languages. CL-OSA is also effective in retrieving documents written in distant languages, such as Japanese and English (cf. ASPEC-JE). All other methods except CL-ASA are significantly less effective for Japanese and English than for closer language pairs like Japanese and Chinese (cf. ASPEC-JC).

All methods exhibit a significant drop in their effectiveness for the Europarl corpus. The likely reason is that the transcripts of political proceedings in this corpus often contain boilerplate text, i.e., frequent words that do not convey additional meaning, such as parliament, resumption of the session, or declare.

CL-ESA performs the poorest for all corpora. A likely reason is that the dimensionality of the concept space (20,000) is too low, although it exceeds the recommendation of Potthast et al.[48, p. 526f.] that 5,000 to 10,000 Wikipedia articles represents a reasonable trade-off between time and retrieval quality. In an experiment by Ashgaria et al. [3], CL-ESA achieved similar results.

CL-OSA’s advantage is particularly strong for the PAN-PC-11 corpus, which is designed to test plagiarism detection methods. CL-OSA outperforms the second-best method (ConceptNet) by 12.71% in terms of MRR and 17.95% in terms of R@1. This result shows the suitability of CL-OSA for the CLPD task.

4.2 Detailed Analysis Evaluation

This evaluation quantifies the effectiveness of CL-OSA in aligning plagiarized text fragments and their sources at the level of characters. We compare CL-OSA’s results to those of eight state-of-the-art CLPD methods reported by Franco-Salvador et al. in the most comprehensive evaluation of CLPD methods to date [20].
Table 3: Detailed analysis results for entire corpus subsets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Spanish-English</th>
<th>German-English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q</td>
<td>P</td>
</tr>
<tr>
<td>CL-OSA</td>
<td>0.573</td>
<td>0.723</td>
</tr>
<tr>
<td>CL-KGA</td>
<td><strong>0.620</strong></td>
<td>0.696</td>
</tr>
<tr>
<td>CL-VSM</td>
<td>0.564</td>
<td>0.630</td>
</tr>
<tr>
<td>CL-ASA</td>
<td>0.517</td>
<td>0.690</td>
</tr>
<tr>
<td>CL-ESA</td>
<td>0.471</td>
<td>0.535</td>
</tr>
<tr>
<td>CL-C3G</td>
<td>0.373</td>
<td>0.563</td>
</tr>
<tr>
<td>XCNN</td>
<td>0.386</td>
<td>0.738</td>
</tr>
<tr>
<td>S2Net</td>
<td>0.514</td>
<td>0.734</td>
</tr>
<tr>
<td>BAE</td>
<td>0.440</td>
<td>0.736</td>
</tr>
</tbody>
</table>

- 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)

4.2.2 CLPD Methods. We compare CL-OSA to eight CLPD methods evaluated by Franco-Salvador et al. [20], which cover all prominent approaches to CLPD discussed in Section 2:

- Machine Translation: CL-ASA, CL-CNG (specifically cross-language character 3-grams CL-C3G), CL-VSM
- Corpus-based Semantics: CL-ESA
- Knowledge-based Semantics: CL-KGA
- Neural Networks: BAE, S2Net, XCNN

4.2.3 Methodology. To enable a comparison of our results to those reported in the study by Franco-Salvador et al., we adhere to the methodology of Experiment B of the previous study [20, p. 94ff.].

Aligning plagiarized text segments in a document with their source segments requires the computation of similarity scores at the sub-document level. Therefore, all documents in the PAN-PC-11 subsets involved in cross-language plagiarism (both suspicious and source documents) were split into fragments. Subsequently, the evaluated CLPD methods were applied to compute the similarity scores for all possible fragment pairs.

CL-OSA splits documents into fragments using a sliding window with a length of six sentences and a step-width of three sentences. Thus, consecutive fragments have an overlap of three sentences, which aids in identifying plagiarism that spans multiple fragments. We use each fragment of a suspicious document containing cross-language plagiarism as a query. For each query, we retrieve from the set of fragments obtained from the respective PAN-PC-11 subset the five fragments with the highest CL-OSA similarity score.

To identify plagiarism that spans multiple fragments, the affected fragments need to be merged. For merging and classifying fragments as plagiarized, we used Algorithm 1 proposed by Franco-Salvador et al. [20, p. 89]. The algorithm checks if the character distance between two query fragments and their potential source fragments retrieved by a CLPD method (in our case, CL-OSA) are below a certain threshold. If so, the fragments are merged and their similarity scores accumulated. If the accumulated similarity scores of the merged fragments are above a certain threshold, the affected text segment is marked as plagiarized. We determined the best-performing thresholds for merging and classifying fragments as plagiarized via parameter tuning runs.

4.2.4 Performance Measures. For the detailed analysis evaluation, we use the performance measures Potthast et al. defined for this
### Table 5: Detailed analysis results by plagiarism case length.

<table>
<thead>
<tr>
<th>Case Length</th>
<th>Model</th>
<th>Spanish-English</th>
<th></th>
<th>German-English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q</td>
<td>P</td>
<td>R</td>
<td>G</td>
</tr>
<tr>
<td>Long cases (x &gt; 5,000 chars.)</td>
<td>CL-OSA</td>
<td>0.366</td>
<td>0.773</td>
<td>0.240</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CL-KGA</td>
<td>0.406</td>
<td>0.414</td>
<td>0.398</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CL-VSM</td>
<td>0.399</td>
<td>0.416</td>
<td>0.391</td>
<td>1.016</td>
</tr>
<tr>
<td></td>
<td>CL-C3G</td>
<td>0.299</td>
<td>0.467</td>
<td>0.269</td>
<td>1.207</td>
</tr>
<tr>
<td></td>
<td>XCNN</td>
<td>0.327</td>
<td>0.587</td>
<td>0.368</td>
<td>1.145</td>
</tr>
<tr>
<td></td>
<td>S2Net</td>
<td>0.369</td>
<td>0.631</td>
<td>0.314</td>
<td>1.200</td>
</tr>
<tr>
<td></td>
<td>BAE</td>
<td>0.369</td>
<td>0.631</td>
<td>0.314</td>
<td>1.200</td>
</tr>
<tr>
<td>Medium cases (700 ≤ x ≤ 5,000 chars.)</td>
<td>CL-OSA</td>
<td>0.054</td>
<td>0.062</td>
<td>0.048</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CL-KGA</td>
<td>0.012</td>
<td>0.009</td>
<td>0.021</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CL-VSM</td>
<td>0.009</td>
<td>0.006</td>
<td>0.014</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>CL-C3G</td>
<td>0.006</td>
<td>0.003</td>
<td>0.009</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>XCNN</td>
<td>0.005</td>
<td>0.004</td>
<td>0.006</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>S2Net</td>
<td>0.008</td>
<td>0.007</td>
<td>0.010</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>BAE</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- 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)

Table 5 presents a more fine-grained analysis of the results reported in Table 3 by distinguishing the length of plagiarism cases. All methods perform better for longer cases than for shorter ones. This result is intuitive, since longer cases offer more data usable for the similarity analysis. Notably, CL-OSA performs better than all other methods in detecting short and medium cases, which are more challenging to identify. The merging algorithm described in Section 4.2.3 greatly improves CL-OSA’s effectiveness for medium and long cases. The larger a plagiarism case, the more fragments will the algorithm merge, and the more likely the accumulated similarity score will be above the reporting threshold.

Table 6 presents another breakdown of the results in Table 3 according to the obfuscation applied to plagiarism cases. The results confirm that identifying manually obfuscated cases, i.e., sense-for-sense translations, is more challenging for all methods, as intended by the creators of the PAN-PC-11 dataset [47]. That most of the manually obfuscated cases are short further increases the difficulty of detecting them [20]. CL-OSA outperforms all other methods for manually obfuscated plagiarism cases in both corpus subsets. Notably, CL-OSA’s PlagDet score exceeds that of the conceptually similar method CL-KGA by a factor of 2.97 for Spanish-English and 2.19 for German-English cases. The deep semantic analysis...
Table 6: Detailed analysis results by obfuscation type.

<table>
<thead>
<tr>
<th>Obfuscation Type</th>
<th>Model</th>
<th>Spanish-English</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>German-English</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q</td>
<td>P</td>
<td>R</td>
<td>G</td>
<td></td>
<td>Q</td>
<td>P</td>
<td>R</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>CL-OSA</td>
<td></td>
<td>0.413</td>
<td>0.506</td>
<td>0.349</td>
<td>1.000</td>
<td></td>
<td>0.370</td>
<td>0.475</td>
<td>0.303</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CL-KGA</td>
<td></td>
<td>0.139</td>
<td>0.158</td>
<td>0.124</td>
<td>1.000</td>
<td></td>
<td>0.169</td>
<td>0.207</td>
<td>0.143</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CL-VSM</td>
<td></td>
<td>0.102</td>
<td>0.121</td>
<td>0.088</td>
<td>1.000</td>
<td></td>
<td>0.109</td>
<td>0.147</td>
<td>0.086</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Translated</td>
<td></td>
<td>0.092</td>
<td>0.146</td>
<td>0.076</td>
<td>1.000</td>
<td></td>
<td>0.085</td>
<td>0.137</td>
<td>0.062</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>manual</td>
<td></td>
<td>0.072</td>
<td>0.104</td>
<td>0.054</td>
<td>1.000</td>
<td></td>
<td>0.042</td>
<td>0.053</td>
<td>0.035</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>obfuscation</td>
<td></td>
<td>0.091</td>
<td>0.141</td>
<td>0.067</td>
<td>1.000</td>
<td></td>
<td>0.115</td>
<td>0.173</td>
<td>0.086</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CL-C3G</td>
<td></td>
<td>0.077</td>
<td>0.116</td>
<td>0.058</td>
<td>1.000</td>
<td></td>
<td>0.085</td>
<td>0.160</td>
<td>0.058</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>XCNN</td>
<td></td>
<td>0.077</td>
<td>0.116</td>
<td>0.058</td>
<td>1.000</td>
<td></td>
<td>0.085</td>
<td>0.160</td>
<td>0.058</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>S2Net</td>
<td></td>
<td>0.091</td>
<td>0.141</td>
<td>0.067</td>
<td>1.000</td>
<td></td>
<td>0.115</td>
<td>0.173</td>
<td>0.086</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BAE</td>
<td></td>
<td>0.085</td>
<td>0.191</td>
<td>0.055</td>
<td>1.000</td>
<td></td>
<td>0.088</td>
<td>0.113</td>
<td>0.072</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>automatic</td>
<td></td>
<td>0.584</td>
<td>0.733</td>
<td>0.485</td>
<td>1.000</td>
<td></td>
<td>0.533</td>
<td>0.684</td>
<td>0.434</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CL-OSA</td>
<td></td>
<td>0.600</td>
<td>0.742</td>
<td>0.595</td>
<td>1.000</td>
<td></td>
<td>0.556</td>
<td>0.642</td>
<td>0.493</td>
<td>1.004</td>
<td></td>
</tr>
<tr>
<td>CL-KGA</td>
<td></td>
<td>0.603</td>
<td>0.673</td>
<td>0.553</td>
<td>1.011</td>
<td></td>
<td>0.445</td>
<td>0.562</td>
<td>0.391</td>
<td>1.053</td>
<td></td>
</tr>
<tr>
<td>CL-VSM</td>
<td></td>
<td>0.552</td>
<td>0.736</td>
<td>0.479</td>
<td>1.077</td>
<td></td>
<td>0.439</td>
<td>0.652</td>
<td>0.373</td>
<td>1.125</td>
<td></td>
</tr>
<tr>
<td>Translated</td>
<td></td>
<td>0.503</td>
<td>0.571</td>
<td>0.479</td>
<td>1.052</td>
<td></td>
<td>0.288</td>
<td>0.431</td>
<td>0.247</td>
<td>1.137</td>
<td></td>
</tr>
<tr>
<td>manual</td>
<td></td>
<td>0.398</td>
<td>0.602</td>
<td>0.347</td>
<td>1.160</td>
<td></td>
<td>0.122</td>
<td>0.343</td>
<td>0.085</td>
<td>1.183</td>
<td></td>
</tr>
<tr>
<td>obfuscation</td>
<td></td>
<td>0.412</td>
<td>0.791</td>
<td>0.331</td>
<td>1.205</td>
<td></td>
<td>0.289</td>
<td>0.715</td>
<td>0.210</td>
<td>1.191</td>
<td></td>
</tr>
<tr>
<td>CL-C3G</td>
<td></td>
<td>0.350</td>
<td>0.784</td>
<td>0.471</td>
<td>1.106</td>
<td></td>
<td>0.406</td>
<td>0.719</td>
<td>0.326</td>
<td>1.164</td>
<td></td>
</tr>
<tr>
<td>XCNN</td>
<td></td>
<td>0.470</td>
<td>0.781</td>
<td>0.386</td>
<td>1.154</td>
<td></td>
<td>0.224</td>
<td>0.520</td>
<td>0.158</td>
<td>1.132</td>
<td></td>
</tr>
<tr>
<td>S2Net</td>
<td></td>
<td>0.470</td>
<td>0.781</td>
<td>0.386</td>
<td>1.154</td>
<td></td>
<td>0.224</td>
<td>0.520</td>
<td>0.158</td>
<td>1.132</td>
<td></td>
</tr>
</tbody>
</table>

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)

5 CONCLUSION & FUTURE WORK

We introduced CL-OSA—a novel method that uses open knowledge graphs for cross-language plagiarism detection. CL-OSA sets itself apart from many state-of-the-art methods by performing a deep semantic analysis of documents using entities and relationships obtained from Wikidata. Our method creates a language-independent semantic representation of documents that allows assessing the documents’ similarity for many languages. CL-OSA does not require machine translation, which is a drawback of several existing methods, whose effectiveness strongly depends on the availability and quality of parallel corpora.

We evaluated CL-OSA for the candidate retrieval and detailed analysis tasks in cross-language plagiarism detection. In the candidate retrieval experiment, CL-OSA outperformed state-of-the-art CLPD methods for all five multilingual test corpora. The difference in CLPD effectiveness was most evident for the PAN-PC-11 corpus, which is tailored to the evaluation of plagiarism detection methods and includes manually translated test cases. CL-OSA’s performance was unaffected by topical domains or the lack of lexical and syntactic similarities among languages. Our method also achieved excellent results for assessing the similarity of documents written in distant language pairs, such as English and Japanese, which represent a major challenge for other CLPD methods.

In the detailed-analysis experiment, CL-OSA and the conceptually similar method CL-KGA outperform all other methods. Considering the entire test corpora, CL-KGA is slightly more effective than CL-OSA. However, our method performs significantly better than CL-KGA in detecting manually obfuscated cases of plagiarism, which are particularly challenging to identify.

Given these results, we consider CL-OSA a promising approach to detect the highly obfuscated cross-language plagiarism we expect of researchers with strong incentives to mask wrongdoing.

In our future work, we plan to further increase the effectiveness of CL-OSA by investigating in more detail which characteristics of CL-KGA cause its performance advantage for long and automatically obfuscated cases. Moreover, we intend to optimize CL-OSA’s weighting scheme for entity types. We hypothesize that using contextual information at the level of documents and fragments instead of the current boolean weighting of the term frequency will improve the selection of relevant concepts and the identification of suspicious cross-language similarity.

REFERENCES

