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

Norman Meuschke
meuschke@uni-wuppertal.de
University of Wuppertal
Wuppertal, 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[1] 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.

[1] 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 McNamee 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].

Glavaš 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 Glavaš 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 ConceptNet[2] 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

[2] http://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*[3] 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.

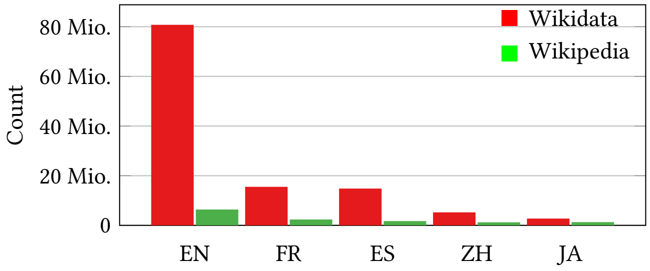


Figure 1: Number of Wikidata entities (red) and Wikipedia articles (green) per language as of September 2021.

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.

[3] https:/babelnet.org

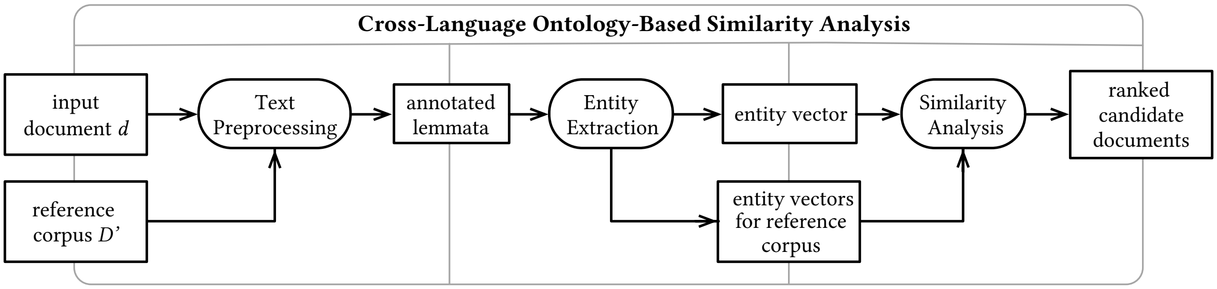


Figure 2: Overview of Cross-Language Ontology-based Similarity Analysis (CL-OSA).

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 [2].

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 $d = w_1 w_2 \dots w_k$ written in language $L_1 \in \mathcal{L}$ as an hierarchy-enhanced entity vector $\mathbf{d}_{\gg}$.

$$\mathbf{d}_{\gg} = (q_1, \dots, q_n).$$

The elements of $\mathbf{d}_{\gg}$ 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 $q = (\Sigma, \mathcal{A}, \Delta, \longrightarrow)$ where

- Σ is a set of labels $\{l_1, \dots, l_m\}$,
- $\mathcal{A}$ is a set of alias sets $\{A_1, \dots, A_m\}$,
- Δ is a set of descriptions $\{\delta_1, \dots, \delta_m\}$, and
- $\longrightarrow \subseteq Q \times P \times 2^Q$ is a property relation.

$P \subsetneq Q$, $\mathcal{L} = \{L_1, \dots, L_m\}$, and $\mathcal{T}$ 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 $q \xrightarrow{p} q'$ instead of $(q, p, q') \in \longrightarrow$ using the following notations:

- $p_{\rightarrow}$ denotes the *instance of* relation
- $p_{\triangleright}$ denotes the *subclass of* relation
- $p_{\gg}$ the *combined* property relation such that $q \xrightarrow{p_{\gg}} q' = q \xrightarrow{p_{\rightarrow}} q' \vee q \xrightarrow{p_{\triangleright}} q'$
- $\longrightarrow^*$ denotes the *transitive* property relation, which we define as $q \xrightarrow{p}{}^* q' = q \xrightarrow{p_{\gg}} q_1 \xrightarrow{p_{\gg}} \dots \xrightarrow{p} q'$

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[5]:

> *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

[5] https://de.wikipedia.org/wiki/Volkswirtschaftslehre

To avoid these issues, CL-OSA 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-OSA 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-OSA 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-OSA infers it using a language detector[6]. 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-OSA 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 [13] 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-OSA 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 [33] for European languages and Kuromoji[7] for Chinese and Japanese.

3.3 Entity Extraction

CL-OSA 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 or 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-OSA retrieves multiple entity candidates, which it disambiguates as described in the following section.

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

Lemma	POS	NE	Entity	English label $l_{\mathbf{en}}$
US	NNP	LOC	$\underline{q_{30}}$	United States of America
			$q_{64142888}$	User Systems (United States)
tax	NN	O	q_{8161}	tax
authorities	NNS	O	$\underline{q_{174834}}$	authority
			$q_{13424378}$	authority [rulership]
			$q_{59646503}$	authority record
			$q_{15708736}$	public authority
teeth	NNS	O	$q_{55347892}$	tooth [heraldic]
			$q_{47450777}$	tooth [gear]
			$\underline{q_{553}}$	tooth [jaw]
			$q_{15043709}$	tine
battle	NN	O	$q_{4869972}$	battle [medieval]
			$\underline{q_{178561}}$	battle
politicians	NNS	O	q_{82955}	politician
Internal Revenue Service	NNP	ORG	q_{973587}	Internal Revenue Service
appears be	VBZ	O	$\underline{q_{3620816}}$	appearance
			$q_{4207474}$	semblance
stance	NN	O	$\underline{q_{172378}}$	stance [martial arts]
			$q_{7598021}$	stance [football]
			$q_{48302332}$	stance [tennis]
			$q_{17052364}$	stance [linguistics]
international	JJ	O	$\underline{q_{1072012}}$	international
			$q_{61029267}$	rest of the world
leaves	VBZ	O	$q_{24759450}$	leave
			$q_{19279529}$	go
			$\underline{q_{5338673}}$	annual leave
			$q_{13561011}$	leave of absence
taxpayers	NNS	O	$q_{25211970}$	taxpayer [building]
			$\underline{q_{1938414}}$	taxpayer

3.4 Entity Disambiguation

CL-OSA 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*

[6] https://github.com/optimaize/language-detector
[7] https://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 triples without 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 $\mathbf{d}_{\ggg}$ 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 $\mathbf{d}_{\ggg}$ if they satisfy $p^m(q) = q_a$ 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 $\mathbf{d}_{\ggg} = (q_1, \ldots, q_n)$ to all vectors $\mathbf{d}'_{\ggg}$ in the reference collection D' by computing the cosine similarity

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

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

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 **P**lagiarism Analysis, **A**uthorship Identification, and **N**ear-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*[8]). 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:

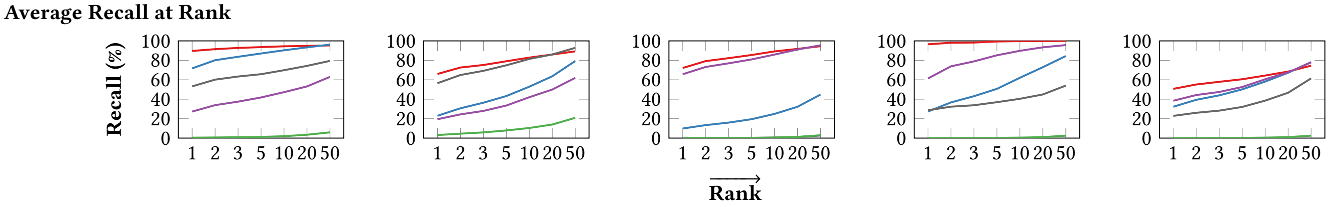
CL-ASA implemented according to Potthast et al. [46, p. 9]. For European languages, we derived the translation probabilities from the dictionaries provided by Aker et al. [1]. For the ASPEC corpora (EN-JA and JA-ZH), we used the program *pialign* by Neubig et al.

[8] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

Table 2: Mean Reciprocal Rank, Recall at Rank, and Average Recall at Rank scores for candidate retrieval.

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

R@k (%)	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
CL-OSA	**89.65**	**91.50**	**92.75**	**65.95**	**72.60**	**75.20**	**72.10**	**79.20**	**82.25**	**96.50**	**98.10**	**98.30**	**50.65**	**55.15**	**57.90**
ConceptNet	71.70	80.15	83.65	22.95	30.65	36.40	9.85	13.35	15.95	27.35	36.70	43.10	32.27	39.38	43.97
USE-ML	27.30	33.90	37.50	19.40	24.40	27.95	65.70	73.20	76.95	61.35	73.85	78.90	38.45	44.30	47.45
CL-ASA	53.20	60.15	63.35	56.45	64.90	69.20	0.05	0.11	0.16	28.66	32.33	33.79	22.85	26.05	28.25
CL-ESA	0.45	0.65	0.90	3.15	4.55	5.85	0.05	0.15	0.15	0.05	0.10	0.15	0.05	0.10	0.15



Average Recall at Rank

[44] to train the probabilities on the Tanaka corpus [53] and the TED English Chinese Parallel Corpus of Speech [9], respectively.

CL-ESA [48] uses a comparable corpus of 20,000 Wikipedia articles, i.e., twice the upper boundary of the dimensionality interval that Potthast et al. reported achieving a good trade-off between retrieval quality and computing time [48, p. 526f.].

ConceptNet [51] refers to the pre-created set of vectors from ConceptNet-Numberbatch available on GitHub[9].

USE-ML [57] refers to the pre-trained model Universal Sentence Encoder-Multilingual based on the Transformer architecture introduced by Yang et al. [57] and available in TensorFlow[10]. This model was trained with the Stanford Natural Language Inference corpus.

The vectorized documents for ConceptNet and USE-ML are available in our Zenodo repository.

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].

4.2.1 Datasets. In accordance with the experiments by Franco-Salvador et al. [20], we used the *English-Spanish* and *English-German*

[9] https://github.com/commonsense/conceptnet-numberbatch
[10] https://tfhub.dev/google/universal-sentence-encoder-multilingual/2

Table 3: Detailed analysis results for entire corpus subsets.

	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)

subsets of the PAN-PC-11 corpus [47]. To our knowledge, these are the only datasets that offer the necessary ground-truth information on "plagiarized" segments at the level of characters. Opposed to our evaluation of the candidate retrieval task (Section 4.1), for which we reused a sample of the PAN-PC-11 corpus provided by Ferrero et al. [14], we extracted the two cross-language subsets directly from the original corpus [49].

The subsets consist of simulated cross-language plagiarism instances of different lengths embedded into topically related text. Most of the "plagiarized" text segments that were taken from documents in the other language were machine-translated. Additionally, hired workers obfuscated approx. 1% of those machine-translated segments manually to increase their obfuscation and make them more challenging to detect [47]. Table 4 summarizes the two datasets.

Table 4: Overview of the German-English (DE-EN) and Spanish- English (ES-EN) subsets of the PAN-PC-11 corpus.

German-English documents	
Suspicious	251
Source	348
Spanish-English documents	
Suspicious	304
Source	202
Plagiarism cases (DE-EN, ES-EN combined)	
Case length	
• Long cases	1,506
• Medium cases	2,118
• Short cases	1,951
Obfuscation	
• Machine translation	5,142
• Machine translation + manual obfuscation	433

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.

		Spanish-English				German-English			
Case Length	**Model**	Q	P	R	G	Q	P	R	G
Long cases (x > 5,000 chars.)	CL-OSA	0.366	0.773	0.240	1.000	0.331	0.737	0.214	1.000
	CL-KGA	**0.406**	0.414	0.398	1.000	**0.366**	0.392	0.347	1.006
	CL-VSM	0.399	0.416	0.391	1.016	0.320	0.386	0.300	1.077
	CL-ASA	0.411	0.535	0.375	1.106	0.339	0.513	0.299	1.168
	CL-ESA	0.351	0.388	0.352	1.076	0.220	0.329	0.198	1.176
	CL-C3G	0.299	0.467	0.269	1.207	0.090	0.275	0.064	1.227
	XCNN	0.327	0.655	0.271	1.253	0.230	0.619	0.170	1.234
	S2Net	0.411	0.587	0.368	1.145	0.322	0.589	0.269	1.212
	BAE	0.369	0.631	0.314	1.200	0.178	0.449	0.127	1.159
Medium cases (700 ≤ x ≤ 5,000 chars.)	CL-OSA	**0.317**	0.713	0.204	1.000	**0.284**	0.659	0.180	1.000
	CL-KGA	0.224	0.224	0.225	1.000	0.211	0.231	0.193	1.000
	CL-VSM	0.205	0.215	0.196	1.000	0.155	0.183	0.134	1.000
	CL-ASA	0.174	0.224	0.142	1.000	0.149	0.204	0.117	1.000
	CL-ESA	0.164	0.174	0.156	1.000	0.092	0.113	0.078	1.000
	CL-C3G	0.131	0.175	0.105	1.000	0.041	0.070	0.029	1.000
	XCNN	0.127	0.221	0.089	1.000	0.096	0.204	0.063	1.000
	S2Net	0.176	0.240	0.139	1.000	0.135	0.217	0.098	1.000
	BAE	0.148	0.241	0.107	1.000	0.072	0.126	0.051	1.000
Short cases (x < 700 chars.)	CL-OSA	**0.054**	0.062	0.048	1.000	**0.053**	0.069	0.043	1.000
	CL-KGA	0.012	0.009	0.021	1.000	0.011	0.008	0.018	1.000
	CL-VSM	0.009	0.006	0.014	1.000	0.007	0.005	0.011	1.000
	CL-ASA	0.006	0.005	0.009	1.000	0.006	0.005	0.009	1.000
	CL-ESA	0.009	0.006	0.015	1.000	0.005	0.003	0.008	1.000
	CL-C3G	0.005	0.004	0.006	1.000	0.004	0.003	0.005	1.000
	XCNN	0.006	0.006	0.006	1.000	0.009	0.009	0.009	1.000
	S2Net	0.008	0.007	0.010	1.000	0.008	0.006	0.010	1.000
	BAE	0.003	0.003	0.004	1.000	0.005	0.004	0.007	1.000

➢ 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)

task as part of the PAN-PC competition series, i.e., Precision (P), Recall (R), Granularity (G), and PlagDet score (Q) [47]. Precision is the fraction of characters pertaining to a plagiarism case and the characters a method reports as plagiarized. Recall quantifies the share of all plagiarized characters a method identifies correctly. Granularity indicates whether a method reports multiple detections for a coherent plagiarism case, or yields overlapping detections, both of which are undesirable. The granularity score is in the interval $[0, 1]$, with $G = 1$ reflecting the best-possible case, i.e, the method reports each plagiarism case as one detection. The PlagDet score combines P, R and G into a single score

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

where F_1 represents the harmonic mean of Precision and Recall.

4.2.5 Results Detailed Analysis. Table 3 shows the results of the detailed analysis evaluation on the full corpus subsets. For the Spanish-English subset, CL-OSA outperforms seven of the eight comparison methods. Only CL-KGA, which is conceptually similar to our method, achieves a slightly higher PlagDet score. For the German-English subset, CL-OSA performs marginally better than CL-KGA and significantly better than the other methods.

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.

		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)

capabilities of CL-OSA seem to provide a significant benefit for identifying these challenging plagiarism cases.

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

[1] Ahmet Aker, Monica Paramita, Marcis Pinnis, and Robert Gaizauskas. 2014. Bilingual Dictionaries for All EU Languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 2839–2845. http://www.lrec-conf.org/proceedings/lrec2014/summaries/803.html

[2] Tareq Al-Moslmi, Marc Gallofre Ocana, Andreas L. Opdahl, and Csaba Veres. 2020. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* 8 (2020), 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

[3] Habibollah Asghari, Omid Fatemi, Salar Mohtaj, Heshaam Faili, and Paolo Rosso. 2019. On the Use of Word Embedding for Cross Language Plagiarism Detection. *Intelligent Data Analysis* 23, 3 (April 2019), 661–680. https://doi.org/10.3233/ida-183985

[4] Alberto Barrón-Cedeño, Parth Gupta, and Paolo Rosso. 2013. Methods for Cross-Language Plagiarism Detection. *Knowledge-Based Systems* 50 (Sept. 2013), 211–217. https://doi.org/10.1016/j.knosys.2013.06.018

[5] Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In *Proceedings of the 2008 International Conference on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN)*, Vol. 377 CEUR WS. CEUR-WS.org, Aachen, Germany, 9–14. http://ceur-ws.org/Vol-377/paper1.pdf

[6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3 (March 2003), 1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

[7] Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: A Multilingual and Multilevel Representation Learning Toolkit for NLP. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), Portorož, Slovenia, 4188–4192. https://aclanthology.org/L16-1662

[8] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16, 2 (Aug. 1990), 79–85. https://doi.org/10.3115/991365.991407

[9] Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Trento, Italy, 261–268. https://aclanthology.org/2012.eamt-1.60

[10] Shijie Chen, Tao Zhang, and Yuejie Zhang. 2005. English-Chinese Cross-Language Information Retrieval using Lucene Toolkit. In *Proceedings of the International Conference on Chinese Computing (ICCC)*. Chinese and Oriental Languages Information Processing Society (COLIPS), Singapore, 1–6. http://www.colips.org/conferences/iccc2005/papers/ICCC-05-121.pdf

[11] Taiseer Eisa, Naomie Salim, and Salha Alzahrani. 2017. Figure Plagiarism Detection Using Content-Based Features. In *Proceedings of the International Conference on Intelligent Computing, Communication and Devices (ICCD)*, Srikanta Patnaik and Florin Popentiu-Vladicescu (Eds.), Vol. 555 AISC. Springer, Singapore, 17–20. https://doi.org/10.1007/978-981-10-3779-5_3

[12] Taiseer Abdalla Elfadil Eisa, Naomie Salim, and Abdelzahir Abdelmaboud. 2020. Content-Based Scientific Figure Plagiarism Detection Using Semantic Mapping. In *Emerging Trends in Intelligent Computing and Informatics*, Faisal Saeed, Fathey Mohammed, and Nadhmi Gazem (Eds.). Vol. 1073. Springer International Publishing, Cham, 420–427. https://doi.org/10.1007/978-3-030-33582-3_40

[13] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, Mass.

[14] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. 2016. A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Association for Computational Linguistics, Portorož, Slovenia, 4162–4169. https://www.aclweb.org/anthology/L16-1657

[15] Jérémy Ferrero, Frédéric Agnes, Laurent Besacier, and Didier Schwab. 2017. Using Word Embedding for Cross-Language Plagiarism Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2: Short Papers. Association for Computational Linguistics, Valencia, Spain, 415–421. https://doi.org/10.18653/v1/e17-2066

[16] Teddi Fishman. 2009. "We Know It When We See It" Is Not Good Enough: Toward a Standard Definition of Plagiarism That Transcends Theft, Fraud, and Copyright. In *Proceedings 4th Asia Pacific Conference on Educational Integrity (4APCEI)*. University of Wollongong, Wollongong, Australia, 1–5. https://www.bmartin.cc/pubs/09-4apcei/4apcei-Fishman.pdf

[17] Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *Comput. Surveys* 52, 6 (Oct. 2019), 112:1–112:42. https://doi.org/10.1145/3345317

[18] Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2013. Cross-Language Plagiarism Detection Using a Multilingual Semantic Network. In *Proceedings of the 35th European Conference on IR Research (ECIR)*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.), Vol. 7814 LNCS. Springer, Berlin, Heidelberg, 710–713. https://doi.org/10.1007/978-3-642-36973-5_66

[19] Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2014. Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. In *Bridging Between Information Retrieval and Databases: Revised Tutorial Lectures of the Promise Winter School 2013*, Nicola Ferro (Ed.). Vol. 8173 LNCS. Springer-Verlag, Berlin, Heidelberg, 227–236. https://doi.org/10.1007/978-3-642-54798-0_12

[20] 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

[21] Marc Franco-Salvador, Paolo Rosso, and Manuel Montes-y-Gómez. 2016. A Systematic Study of Knowledge Graph Analysis for Cross-Language Plagiarism Detection. *Information Processing & Management* 52, 4 (July 2016), 550–570. https://doi.org/10.1016/j.ipm.2015.12.004

[22] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 7. Morgan Kaufmann Publishers Inc., Hyderabad, India, 1606–1611. https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-259.pdf

[23] Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus. *Journal of the Association for Information Science and Technology* 65, 8 (Aug. 2014), 1527–1540. https://doi.org/10.1002/asi.23228

[24] Goran Glavaš, Marc Franco-Salvador, Simone P. Ponzetto, and Paolo Rosso. 2018. A Resource-Light Method for Cross-Lingual Semantic Textual Similarity. *Knowledge-Based Systems* 143 (March 2018), 1–9. https://doi.org/10.1016/j.knosys.2017.11.041

[25] Wael H. Gomaa and Aly A. Fahmy. 2013. A Survey of Text Similarity Approaches. *International Journal of Computer Applications* 68, 13 (April 2013), 13–18. https://doi.org/10.5120/11638-7118

[26] Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query Expansion for Mixed-Script Information Retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM Press, Gold Coast, Queensland, Australia, 677–686. https://doi.org/10.1145/2600428.2609622

[27] Parth Gupta, Rafael E. Banchs, and Paolo Rosso. 2017. Continuous Space Models for CLIR. *Information Processing & Management* 53, 2 (March 2017), 359–370. https://doi.org/10.1016/j.ipm.2016.11.002

[28] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1681–1691. https://doi.org/10.3115/v1/p15-1162

[29] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit*. Association for Machine Translation in the Americas, Phuket, Thailand, 79–86. https://www.statmt.org/europarl/

[30] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14, Vol. 32)*. JMLR.org, Beijing, China, II–1188–II–1196. arXiv:1405.4053 http://proceedings.mlr.press/v32/mittelman14.pdf

[31] Peipei Li, Haixun Wang, Kenny Q. Zhu, Zhongyuan Wang, and Xindong Wu. 2013. Computing Term Similarity by Large Probabilistic isA Knowledge. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*. ACM Press, San Francisco, California, USA, 1401–1410. https://doi.org/10.1145/2505515.2505567

[32] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics, Denver, Colorado, 151–159. https://doi.org/10.3115/v1/w15-1521

[33] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*. The Association for Computer Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/p14-5010

[34] Nikolay A. Mazov, Vadim N. Gureev, and D. V. Kosyakov. 2016. On the Development of a Plagiarism Detection Model Based on Citation Analysis Using a Bibliographic Database. *Scientific and Technical Information Processing* 43, 4 (Oct. 2016), 236–240. https://doi.org/10.3103/s0147688216040092

[35] Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7, 1-2 (Jan. 2004), 73–97. https://doi.org/10.1023/b:inrt.0000009441.78971.be

[36] Norman Meuschke and Bela Gipp. 2013. State of the Art in Detecting Academic Plagiarism. *International Journal for Educational Integrity* 9, 1 (June 2013), 50–71. https://doi.org/10.5281/zenodo.3482941

[37] Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel Keim, and Bela Gipp. 2018. An Adaptive Image-Based Plagiarism Detection Approach. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. ACM Press, Fort Worth, USA, 131–140. https://doi.org/10.1145/3197026.3197042

[38] Norman Meuschke, Moritz Schubotz, Felix Hamborg, Tomas Skopal, and Bela Gipp. 2017. Analyzing Mathematical Content to Detect Academic Plagiarism. In

Proceedings ACM Conference on Information and Knowledge Management (CIKM). ACM, Singapore, 2211–2214. https://doi.org/10.1145/3132847.3133144

[39] Norman Meuschke, Nicolas Siebeck, Moritz Schubotz, and Bela Gipp. 2017. Analyzing Semantic Concept Patterns to Detect Academic Plagiarism. In *Proceedings of the International Workshop on Mining Scientific Publications (WOSP) co-located with the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, Toronto, Canada, 46–53. https://doi.org/10.1145/3127526.3127535

[40] Norman Meuschke, Vincent Stange, Moritz Schubotz, and Bela Gipp. 2018. HyPlag: A Hybrid Approach to Academic Plagiarism Detection. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM Press, Ann Arbor, MI, USA, 1321–1324. https://doi.org/10.1145/3209978.3210177

[41] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 2019. Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. In *Proceedings of the Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Xplore, Urbana-Champaign, Illinois, USA, 120–129. https://doi.org/10.1109/jcdl.2019.00026

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*. Curran Associates Inc., Lake Tahoe, CA, USA, 3111–3119. arXiv:1310.4546 http://arxiv.org/abs/1310.4546

[43] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, Portorož, Slovenia, 2204–2208. https://www.aclweb.org/anthology/L16-1350

[44] Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An Unsupervised Model for Joint Phrase Alignment and Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 632–641. https://aclanthology.org/P11-1064

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/d14-1162

[46] Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language Plagiarism Detection. *Language Resources and Evaluation* 45, 1 (March 2011), 45–62. https://doi.org/10.1007/s10579-009-9114-z

[47] Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Overview of the 3rd International Competition on Plagiarism Detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF)*, Forner, Pamela, Navigli, Roberto, Tufis, Dan, and Ferro, Nicola (Eds.), Vol. 1177 CEUR WS. CEUR-WS.org, Amsterdam, Netherlands, 1–10. http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-PotthastEt2011a.pdf

[48] Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-based Multilingual Retrieval Model. In *Proceedings of the 30th European Conference on Advances in Information Retrieval (ECIR)*. Springer, Berlin, Heidelberg, 522–530. https://doi.org/10.1007/978-3-540-78646-7_51

[49] 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

[50] Moritz Schubotz, Olaf Teschke, Vincent Stange, Norman Meuschke, and Bela Gipp. 2019. Forms of Plagiarism in Digital Mathematical Libraries. In *Proceedings International Conference on Intelligent Computer Mathematics*, Vol. 11617 LNCS. Springer, Czech Republic, 258–274. https://doi.org/10.1007/978-3-030-23250-4_18

[51] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, San Francisco, California, USA, 4444–4451. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[52] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), Genoa, Italy, 2142–2147. http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf

[53] Yasuhito Tanaka. 2001. Compilation of A Multilingual Parallel Corpus. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING)*. Pacific Association for Computational Linguistics, Kitakyushu, Japan, 1–4. http://www.afnlp.org/archives/pacling2001/pdf/tanaka.pdf

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[55] Debora Weber-Wulff. 2014. *False Feathers: A Perspective on Academic Plagiarism*. Springer Berlin Heidelberg, Berlin. https://doi.org/10.1007/978-3-642-39961-9

[56] Linwei Yang. 2017. ECCE 2.0: The English Chinese Corpus of Editorials. http://corpus.bfsu.edu.cn/info/1070/1415.htm

[57] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12