big data and cognitive computing

Article

OTNEL: A Distributed Online Deep Learning Semantic Annotation Methodology

Christos Makris * and Michael Angelos Simos *

Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece

* Correspondence: makri@ceid.upatras.gr (C.M.); asimos@ceid.upatras.gr (M.A.S.);
Tel.: +30-2610-996-968 (C.M.)

Received: 15 September 2020; Accepted: 27 October 2020; Published: 29 October 2020

Abstract: Semantic representation of unstructured text is crucial in modern artificial intelligence and information retrieval applications. The semantic information extraction process from an unstructured text fragment to a corresponding representation from a concept ontology is known as named entity disambiguation. In this work, we introduce a distributed, supervised deep learning methodology employing a long short-term memory-based deep learning architecture model for entity linking with Wikipedia. In the context of a frequently changing online world, we introduce and study the domain of online training named entity disambiguation, featuring on-the-fly adaptation to underlying knowledge changes. Our novel methodology evaluates polysemous anchor mentions with sense compatibility based on thematic segmentation of the Wikipedia knowledge graph representation. We aim at both robust performance and high entity-linking accuracy results. The introduced modeling process efficiently addresses conceptualization, formalization, and computational challenges for the online training entity-linking task. The novel online training concept can be exploited for wider adoption, as it is considerably beneficial for targeted topic, online global context consensus for entity disambiguation.

Keywords: named entity disambiguation; text annotation; word sense disambiguation; ontologies; Wikification; neural networks; machine learning

1. Introduction and Motivation

Named entity disambiguation (NED) is a process involving textual mention resolution and assignment to predefined concepts from a knowledge base or concept ontology. The deterministic identification and linking of semantically dominant entity mentions, based on contextual information available, is not trivial in most cases; ambiguity is common on unstructured corpora, as homonymy and polysemy phenomena are inherent to natural languages.

Advances in the domains of artificial intelligence, information retrieval, and natural language processing, outlining the requisition of semantic knowledge input, such as common paradigms like the bag of words representation, are proven inefficient for deeper knowledge extraction and, hence, higher accuracy. As a result, NED is a common step for many relevant tasks including information retrieval [1], data mining [2,3], and web and semantic search [4–7], consequently being a vital component of the artificial intelligence (AI), internet and information industries.

One of the basal challenges in NED involves the maintenance of knowledge resources, especially as new domains and concepts arise or change dynamically over time. In recent years, Wikipedia has been leveraged as a knowledge base and concept universe due to its online nature. The introduction of Wikipedia in the domain derived noteworthy leaps in classic challenges such as knowledge acquisition and adversarial knowledge resolution, as its articles tend to summarize widespread and commonly accepted concepts.

Deep learning architectures have recently been established in several scientific fields including machine translation, computer vision, medical image analysis, speech recognition, audio recognition, social networks, bioinformatics, and material inspection. As such, methodologies successfully modeled high-level abstraction patterns, leveraging deep multilevel transformations, and several approaches successfully addressed the NED problem. However, a series of factors constitute a challenging background for the task.

The engagement of deep learning architectures preconditions the dimensionality projection to lower-dimension spaces for the training input as computational challenges with large-scale training datasets arise. Consequently, the training input entropy is abstracted during a dimensionality reduction process aiming at fitting sparse input spanning plenteous domains to predefined sized dimension spaces, mainly for computational feasibility purposes. As a result of this computational complexity and accuracy trade-off, the inference of semantics deviant from the dominant patterns is burdensome. In addition, the extensive training process required in the scope of a general-purpose NED application is demanding from a computational complexity perspective as outlined in [8]. Our introduced methodology employs an alternative modeling and dimensionality reduction approach method, detailed in Section 3.

Another fundamental adversity for the task resides in knowledge acquisition, including adversarial knowledge resolution and the impact of noise in the training input. Successful deep learning applications require vast training sets. As the task is based on facts for semantic acquisition of pertinent sense representation associations in the available context, the intricacy of semantic attainment, defined as *knowledge acquisition bottleneck* in [9], is dominant. Consequently, the attainment of high-quality data at a scale for wide semantic coverage is not trivial. Similar works as detailed in [8] often rely on diffusive sources ranging from structured ontologies to unstructured corpora, for example, by inducting context with unsupervised techniques for the inference of co-reference information. The impact of noise in the training input is critical for attaining high accuracy at scale. On the contrary, uncontrolled data cleansing approaches aiming at eliminating anomalies on the input training sets could result in substantial information loss for the resolution of more intricate and less frequent senses of a polysemous anchor.

In this work, we propose a novel approach for efficient NED. In particular, by employing divergent thinking on the main task impediments described above, we propose a model for dimensionality reduction according to topical confinement in the context of online training. We focus on minimizing the impact of input data loss and simplifying the task by leveraging topical inference using a semantic ontology information network representation of Wikipedia.

The remainder of this manuscript is organized as follows: the necessary background and related work are presented in Section 2. Our methodology and implementation details are presented in Section 3. The experimental process is described and assessed in Section 4. Our final conclusions are presented in Section 5, along with potential improvements and future work.

2. Background and Related Work

The NED task requisites a knowledge base or concept ontology as its foundation for the identification of named entities, to resolve text segments to a predefined concept or sense universe. Human domain experts also need such a knowledge ontology for identifying the appropriate sense of a polysemous mention within a context. As the creation of knowledge resources by human annotators is an expensive and time-consuming task, facing implications as new concepts or domains emerge or change eventually, the knowledge acquisition issue has been pervasive in the field. The maintainability, coverage, and knowledge acquisition challenges have been outlined on several manually created ontologies applied to the NED task. As a result, attempts for unifying such ontologies emerged; however, they encountered accuracy issues throughout the unification process.

As Wikipedia is an online crowdsourcing encyclopedia with millions of articles, it constitutes one of the largest online open repositories of general knowledge. Wikipedia articles are created and maintained by a multitudinous and highly active community of editors. As a result, widespread and commonly accepted textual descriptions are created as derivatives of a diverse knowledge convergence process in real time. Each article can be interpreted as a knowledge entity. As Wikipedia's online nature inherits the main principles of the web in a wide and highly active user base, named entity linking with Wikipedia is among the most popular approaches in several similar works. The rich contextual and structured link information available in Wikipedia along with its online nature and wide conceptual coverage can be leveraged toward successful high-performance named entity linking applications.

2.1. Named Entity Disambiguation Approximations

Among natural language processing domain tasks, NED and word sense disambiguation (WSD) are acknowledged as challenging for a diversity of aspects. WSD was defined as AI-complete in [8]. AI-completeness is defined by analogy to the nondeterministic polynomial completeness (NP-completeness) concept in complexity theory.

Several formalization approaches have been applied at entity linking coarseness scopes ranging from specific sense ontological entities to generic domains or topics. The disambiguation coverage spans from the disambiguation of one to several senses per sentence. Domain confinement assumptions may also be present on the entity universe.

According to [8], WSD and, hence, NED approaches may be broadly classified into two major categories:

Supervised machine-learning methodologies are used for inferring candidate mentions on the basis of knowledge inference from labeled training sets, usually via classification techniques.

Unsupervised methodologies are based on unstructured corpora for the inference of semantic context via unsupervised machine-learning techniques.

A second level further distinction according to knowledge sources involved can be made as follows:

- knowledge-based, also known as knowledge-rich, relying on lexical resources such as ontologies, machine-readable dictionaries, or thesauri;
- corpus-based, also known as knowledge-poor, which do not employ sense-labeled knowledge sources.

Supervised knowledge-based NED methodologies are in the limelight of current research focus. Wikipedia is commonly employed for underlying knowledge base representation as an entity linking ontology.

2.2. Early Approaches

The pioneering works on the NED problem using Wikipedia for the entity linking approach were [9–11]. The works proposed methods for semantic entity linking to Wikipedia. Those early methods clearly captured the technical impediments of the task, while proposing some effective early solutions. Foundations for future work were placed by the establishment of the commonness feature value for the task.

In [12,13], a more sophisticated approach to the task led to the introduction of relatedness among Wikipedia articles as an invaluable measure of semantic compatibility. Relatedness was defined as the inbound link overlap between Wikipedia articles. The coherence of input text anchors disambiguated with unambiguous mentions of the input was used as the core of the introduced models. Specifically, ambiguous mentions were ranked on the basis of a global score formula involving statistics, relatedness, and coherence for the final selection.

The segmentation of the ranking scores to local and global resulted in further improvements in [14]. Local scores were leveraged for the contribution representation of contextual content surrounding an anchor being processed. The consensus among every input anchor disambiguation within the full frame of the input was modeled as a global score. The problem was formalized as a ranking selection and a quadratic assignment problem, aiming at the approximation of an entity mention for each anchor on the basis of a linear summation of local and global scores.

Another suite with a focus on accuracy and computational efficiency of short input was introduced in [15]. The work is particularly popular and established as a baseline to date. Relatedness, commonness, and other Wikipedia statistics were combined in a voting schema for the selection of the top scoring candidate annotation for a second-step evaluation and selection process.

An alternate modeling approximation was used by [16,17]. An undirected weighed graph was used for the knowledge base representation. The graph nodes were used to model entity annotations or candidate entities. The weighted edges among entities of the graph were used for representing relatedness. Weighted edges among mentions and entities of the graph were used to model contextual similarities. These representations were referred to as the mention-entity graph in [16], and a dense subgraph search approximation was used for the selection of a subgraph of anchor nodes, each containing a unique mention-entity edge. In [17], the representation was referred to as a referent graph, and the methodology employed was based on the PageRank algorithm.

In [18], some innovative approaches for text annotation and entity linking were contributed. Voting schema approximations were introduced, along with a novel method inspired by the human interpretation process on polysemous contexts. An iterative method approach was employed for the modeling process. The iteration converged to proposed annotations while balancing high accuracy with performance, leveraging established metrics derived from the Wikipedia graph representation.

A graph knowledge base representation was employed by [19], and a Hyperlink-Induced Topic Search (HITS) algorithm variant using a breadth first search traversal was evaluated with the candidate entities of the input text anchors as initial seed nodes. Coreference resolution heuristics, extension of surface forms, and normalization contributions to the system constituted the core of this work.

The architecture of [15] was refined and redesigned in WAT [20], as several methodology variants were introduced for experimental assessment. The three-component architecture was revisited by some PageRank and HITS algorithm-based approaches. The main components were thoroughly assessed, and results for a series of methodologies were contributed to the community.

2.3. Recent Deep Learning Approaches

A leading deep learning approximation for the problem was presented in [21]. A vector space representation was used for modeling entities, context, and mentions. The core methodology architecture consisted of a convolutional neural network, in various context windows, for the projection of anchors on the continuous vector space. Finally, a computationally demanding methodology employing a tensor network introduced context and mention interaction information. A similar vector space representation approach of mentions and entities was also employed in [22]. The core disambiguation methodology extended the skip gram model using a knowledge base graph. At a second level, the vector space proximity optimization of vectors representing coherent anchors and entities was used for concluding the process

The authors of [23] introduced a suite combining established approaches, such as graph representation and knowledge base statistics, with deep learning benefits, involving an ensemble consensus disambiguation step. Specifically, variable sized context windows were used by a "neural attention mechanism" with an entity embedding representation.

As most systems rely on heuristics or knowledge-based approaches for conducting semantic relation evaluations, such as coreference, relatedness, or commonness for the conceptual compatibility assessment, the authors of [24] followed a neural entity linking approach, modeling relations as latent variables. Specifically, they extracted semantic relations in an unsupervised manner using an end-to-end optimization methodology for selecting the optimal mentions. The proposed multi-relational model exhibited high performance throughout an experimental evaluation process.

The problem was also addressed in [25] by leveraging a knowledge graph representation. This work was based on the observation that the link density on the representation graph plays a key role as the vertex degree had a major impact to the selection of strongly coherent nodes. To that end, their methodology induced a density enhancement step on the graph representation on the basis of cooccurrence statistics from an unstructured text for relational inference. A training step of entity embeddings was employed for extracting similarity results for the linking step. As anticipated, the system presented exceptional results for the less densely interconnected concepts on the input, resulting in high performance throughout the experimental assessment through a simple, yet effective method.

The authors of [26] attempted to address weaknesses in previous global models. Specifically, by filtering inconsistent candidate entity annotations, they successfully simplified their proposed model while reducing noise on data input. The task was treated as a sequence decision problem, as a sequential approach of exploiting disambiguated mentions during the disambiguation of subsequent anchors was applied. Finally, a reinforcement learning model was used, factoring in a global context. The experimental results outlined accuracy and high generalization performance.

2.4. Conclusions and Current Limitations

Following the success of deep learning methodologies on AI tasks, several similar research endeavors approached the NED, using deep neural network architectures, furthering the outstanding research works outlined above. However, the input dimensionality challenges placed considerable impediments of production-ready, computationally efficient methodologies as outlined in [27]. The complexity of recent approximations employing deep learning architectures led to several recent works, including [28] which queried whether deep neural network NED methodologies are currently applicable for industry-level big data AI applications compared to simpler and more scalable approaches. Current methods focus more and more on accuracy instead of run-time performance, neglecting the options for complexity reduction in many cases, by focusing on input dimensionality for complexity reduction. To that end, systems, like RedW employ a performance-oriented approach relying on graph and statistical analysis features, questioning deep neural network approaches at scale. As deep learning methodologies have been established in terms of knowledge inference and enhanced modeling capabilities, a computationally efficient approach bridging complexity and performance would be propitious for wide, industry adoption.

3. Materials and Methods

3.1. Notations and Terminology

For readability enhancement purposes, this section presents a terminology and notation summary. The terminology in use is aligned with widely adopted previous works in the domain.

- The Wikipedia articles are also referred to as Wikipedia entities, denoted as p.
- A text hyperlink to a Wikipedia page is denoted as a *mention.*
- Text hyperlink anchors within Wikipedia pointing to another page or article are referred to as anchors and denoted as a. Indices are used for referral to specific items in the anchor sequence as follows: a_0 is the first anchor, $i + 1$ and so on. The number of anchors of a text input is cited as m.
- The notation p_a refers to one of the candidate Wikipedia page senses of the anchor a.
- The set of linkable Wikipedia entities to an anchor a is denoted as *Pg(a).*
- The ensemble of inbound links to a given Wikipedia entity p is represented using *in(p).*
- The size of the Wikipedia entities ensemble is cited as $|W|$.
- *link(a)* refers to the cardinality of the count of an anchor's indices as a mention.
- *freq(a)* denotes the total occurrence count of an anchor text within a corpus, including free text and hyperlinks.
- *lp* denotes the link probability of a text segment.

A formal definition of the task on the basis of the above notation can be summarized as the selection of the best fit mention to a p_a from *Pg(a)* for each anchor a_i from the set of identified anchors of a text input.

3.2. Knowledge Extraction

Knowledge is fundamental in an NED task. The current work relies on semantic information from hyperlink anchors to Wikipedia entities. Our methodology supports knowledge acquisition by incorporating any Wikipedia annotated corpus as a training set. In the scope of this work, we leverage the corpus of Wikipedia and the annotated inter-wiki hyperlinks for composing the mention universe ensemble. This ensemble of potential anchor senses grows in parallel with Wikipedia's corpus link structure and its semantic coverage, and it can be considered a sound foundation for our knowledge acquisition process, due to the collaborative online nature of the encyclopedia. Wikipedia entities are widely adopted as an underlying representation ontology for the task due to their commonly accepted textual descriptions.

The population of the mention universe requires the ensemble of Wikipedia pages in MediaWiki article namespace, i.e., pages in namespace with identifier 0. Redirect page unification is carried out for the inclusion of the redirect link context. This involves following the redirect chains and accordingly updating Wikipedia hyperlinks to the corresponding direct link entity IDs. As in [10,11,14,15,20], a preprocessing of a Wikipedia snapshot can be used for the initial extraction of the mention universe, which can remain up to date in syndication with the Wikimedia Update Service [29]. The process involves harvesting the following:

- *Anchor ID*: by keeping an identifier encoding for each text segment encountered as a hyperlink on the processed Wikipedia snapshot.
- *Mention entity ID*: the Wikipedia ID pointed to by a mention. Maintaining this information is necessary for deriving relatedness and commonness statistics.
- *Source article ID*: the Wikipedia article ID where an individual mention is encountered. This is necessary for relatedness calculations.

The above structures constitute the core of the knowledge acquisition for the extraction, transformation, and loading of our training dataset universe, effectively composing a graph representation of Wikipedia. In addition, an *anchor occurrence count* dictionary was extracted and maintained for link probability calculations via a second parse of the corpus for implementation simplicity. An appropriate indexing mechanism can be implemented for avoiding this second parse.

The mitigation of noise impact during the knowledge acquisition phase is crucial to the success of our NED methodology and any deep learning model. In the first stage, following an approach inspired by [25], we performed a mention dictionary density enhancement, by incorporating redirect title information. Specifically, page and redirect titles were treated as hyperlinks in a special source article ID. In the next step, unlike many recent deep learning approaches employing a coarser approximation, we applied filtering rules for ignoring low-frequency mention anomalies, with a relative threshold up to 0.5%, at a minimum of one occurrence. Common preprocessing rules for discarding stop-words, punctuation, single characters, and special symbols were also applied to the extracted mention vocabulary, as established in similar works [12–20]. The knowledge extraction phase was straightforward for both the initial loading and the online syndication of the mention universe, as real-time updates were performed in the structures outlined above.

As an outcome from the knowledge extraction process, Wikipedia was encoded in a mention database, enabling the next steps.

3.3. Methodology

The focus of this work was oriented toward the named entity disambiguation task. The task prerequired anchor extraction. The entity universe to be linked by our system was derived as detailed in the previous section for creating a mention database. In the first step, an extraction, transformation, and loading process was carried out on the unstructured text input for disambiguation (Section 3.3.1). As a next step, we applied a topical coherence-based pruning technique for the confinement of the entity linking scope to coherent topics in a given context (Section 3.3.2). Then, we employed a novel deep learning model for the selection of candidate mentions of an anchor on a local context window, modeling the problem as a classification problem (Section 3.3.3). Finally, a quantification of uncertainty scoring step followed for the confidence evaluation of outcome predictions (Section 3.3.4). Figure 1 outlines our methodology. In the remainder of this manuscript, our methodology is referred to as OTNEL (Online Training Named Entity Disambiguation).

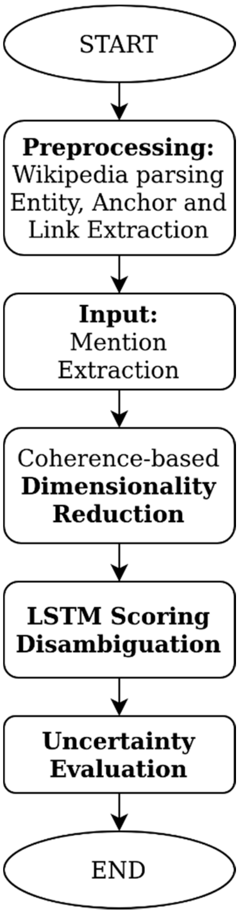

Figure 1. OTNEL (Online Training Named Entity Disambiguation) methodology flowchart.

3.3.1. Extraction, Transformation, and Loading

In the first stage, the unstructured text input was parsed for extracting candidate anchors along with their candidate mentions for further evaluation in the following steps. The input underwent a tokenization process for composing candidate n-grams, with n sized from 1–6. The candidate n-grams were initially evaluated for existence in our mention database as in similar works [10,13,15,20]. A n-gram present in the database could be identified as an anchor for annotation. However, the case of overlapping n-grams needed further examination. The *link probability* of a mention as outlined below by Equation (1) was basial in this examination process.

$$lp(a) = \frac{link(a)}{freq(a)}. \tag{1}$$

Link probability expresses the probability of a word or text fragment occurring as a hyperlink within a corpus. As expressed above, *link(a)* denotes the number of occurrences of anchor *a* as a link. The notation *freq(a)* depicts the occurrence frequency of *a* in a corpus. To preserve meaningful mentions and filter semantically meaningless mentions, *n*-grams with link probability less than 0.001 were pruned similarly with the corresponding knowledge extraction process. As link probability indicates link worthiness, in cases of overlap, the *n*-gram with the highest link probability was selected. Stop-words, punctuation, special symbols, and characters were ignored as they did not return matches in the mention database not being present in the mention universe due to the relevant filtering during the knowledge extraction phase. The specific *n*-gram length was selected in accordance with the maximum size of links on our dataset. Larger *n*-gram lengths would have no effect. Smaller lengths would confine the maximum token length of detected anchors. After this step, unstructured text segments were converted to sequences of semantically significant anchors.

3.3.2. Coherence-Based Dimensionality Reduction

Generic deep learning approximations to the problem face feasibility intricacies at scale for the NED task. On the other hand, in the similar task of named entity recognition, the problem space is limited for the NED task. Specifically, the problem space dimension spans to the mention universe registered on the underlying knowledge base. A perusal of the Wikipedia knowledge graph representation delineates relevant topic coherence with a high degree of reciprocity that can be exploited for discarding incoherent entity mentions from further evaluation in a given context.

As the current work constitutes online semantic annotation, we applied topical confinement for our predictions in terms of the online training process. Specifically, in the first stage, the knowledge graph was pruned to the candidate mentions set of identified input anchors. In the next step, we recalled a training set consisting of mentions within that specific subgraph. This process could be iterated until a wider subgraph was covered, forming a clique from the knowledge graph. As our aim was a vast reduction in the dimensionality space involved, enabling on-the-fly training, a single iteration was performed. The trained model for the specific topical confinement could be persisted for future predictions or on-the-fly training expansion as training input becomes available. As a result, our methodology's dynamic adaptation to rapid underlying knowledge changes as twofold. Firstly, as described in Section 3.2, our mention universe and knowledge graph remained up to date in near real time in syndication with the Wikimedia Update Service [29]. Secondly, the online training process enabled eventual adaptation on potential knowledge changes or updates within the scope of the topical confinement.

It is worth noting that existing approaches to the problem are entrenched by focusing on the generality in a shared vector space with a unified model across topics. However, our approach efficiently confines the training scope required to explicitly generate coherent topical context. The application of similar works following unified approaches would be prohibitive in the online scope of the NED task.

3.3.3. Named Entity Disambiguation

For unambiguous anchors identified by the extraction step, a single entity annotation is available and can be used for linking, i.e., $|Pg(a)| = 1$. For cases where $|Pg(a)|$ has more than one entry, a compatible mention evaluation is needed. Polysemous anchors may have several candidate entities for linking derived from the relevant mention ensemble of the knowledge base. However, using the knowledge base entity dimension as our output dimension would place exorbitant barriers on performance. For named entity disambiguation, we modeled the process as a feature-based classification problem, leveraging an architecture of long short-term memory (LSTM) cells in an artificial recurrent neural network (RNN). Specifically, the problem of selecting a coherent mention for a polysemous anchor in a context was modeled as a binary classification problem as described below.

For every candidate mention of an anchor, we evaluated the classification to the following complementary classes:

- class 1: the compatibility of the mention in the given context;
- class 0: the incompatibility of that mention in the given context.

In the next phase, we could utilize the penultimate deep learning model layer scoring for depicting class predictions probabilities. We selected the highest scoring class 1, i.e., compatible mention, as the disambiguation result.

For maintaining a low input dimension in our model, in the performance context of the online scope of the problem, we provided a set of three features at the input layer, summarizing the gist of topical semantic information. Those features were as follows:

An *inter-wiki Jaccard index average*, as shown in Equation (2). This formula expresses the reciprocity of inbound mentions. The feature of Jaccard similarity was established in [20] as a strong Wikipedia entity coherence measure.

$$avg\ interwiki\ Jaccard\ index(a_i) = \sum\nolimits_{k=0}^{k=i-1} \frac{\left|in\left(p_{a_i}\right) \cap in\left(p_{a_k}\right)\right|}{\left|in\left(p_{a_i}\right) \cup in\left(p_{a_k}\right)\right|} / \mathrm{m} + \sum\nolimits_{k=i+1}^{k=m} \frac{\left|in\left(p_{a_i}\right) \cap in\left(p_{a_k}\right)\right|}{\left|in\left(p_{a_i}\right) \cup in\left(p_{a_k}\right)\right|} / \mathrm{m}. \quad (2)$$

Relatedness is an established measure of semantic entity relatedness. The feature has been used as a core disambiguation primitive in several works [13–15]. In this case, we applied an *average relatedness* feature as depicted by Equation (3).

$$avg\ relatedness(a_i) = \sum\nolimits_{k \in \{p_{a_0} \dots p_{a_m}\} - \{p_{a_i}\}} \frac{\log\left(\max\left(\left|in\left(p_{a_i}\right)\right|, \left|in\left(p_{a_k}\right)\right|\right)\right) - \log\left(\left|in\left(p_{a_i}\right) \cap in\left(p_{a_k}\right)\right|\right)}{\log(|W|) - \log\left(\min\left(\left|in\left(p_{a_i}\right)\right|, \left|in\left(p_{a_k}\right)\right|\right)\right)} / \mathrm{m}. \quad (3)$$

Commonness as defined by Equation (4) is the prior probability of an anchor pointing to a specific Wikipedia entity. Commonness was broadly used in similar works, contributing significant statistical information to the model.

$$Commonness(p_k, a_i) = P(p_k | a_i). \quad (4)$$

Figure 2 presents the deep learning layers of our classifier's distributed architecture. The classifier received a three-dimensional vector of feature scores as input, summarizing the contextual compatibility of an evaluated candidate mention for an anchor. This evaluation was derived as a classification score for the binary output compatible/incompatible classes. As more than one or (in rare cases) even none of the candidate mentions were classified as compatible in a context, we exploited the penultimate layer score for deriving a relative prediction to select the most coherent mention.

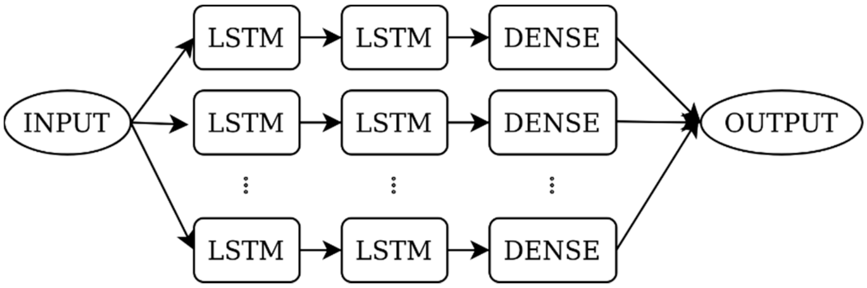

Figure 2. The proposed methodology classifier layer architecture.

The model's input dimensionality was intentionally maintained low through the employment of established context features. Leveraging Keras' [30] and Tensorflow's [31] distributed execution, the input was equally split and fed to the distributed deep learning pipelines. Each individual output was combined to form our classifier output. The complexity was maintained as simple as possible for computational performance reasons.

LSTM was established for addressing RNN's limitations [32]. General RNNs exhibit the "vanishing gradient problem", resulting in a declining effect of the contribution of previous training inputs. The LSTM layer building blocks comprise cells featuring typical input and output gates, along with a forget gate enabling a connected cell state and output recurring feedback from previous states. The *sigmoid* function is commonly adopted as an activation function in LSTM architectures [33] for the input, forget, and output gates, efficiently exhibiting the characteristics and mathematical assets of an effective activation function. The *Tanh* function, a hyperbolic tangent function, is commonly applied as the LSTM candidate and output hidden state.

Our model consisted of two stacked LSTM cell layers after the input followed by a dense layer, producing the output summary. In our model implementation, the LSTM cell size used for the first and second layer was 3, thereby maintaining a low training complexity, yet a high degree of nonlinear feature relation adaptivity. The *MSE* loss function and *Adam* optimizer were used during the model training phase. The *Tanh* activation function and *sigmoid* recurrent activation were employed for the LSTM layers parameters.

The intuition behind the specific multilevel LSTM layer architecture was to involve enhanced semantic relation persistence from the topically confined training sequence. In addition to a clear architecture, simplicity, and modeling and computational efficacy, the methodology enables enhanced prediction strength via exploiting a rich set of both positive and negative linking training examples.

As depicted in the comparative results on a different domain by [33], several activation function options may be explored and compared, contributing intriguing results even in the case of simple classification problems and LSTM architectures. However, in the scope of the current work, we focused on the general approach of a model for the problem, applying established activation functions that experimentally exhibit efficient results for a range of relevant domains. Further exploration of tunning options for our deep learning architectural approach in the specific domain is among our plans for future work.

3.3.4. Quantification of Uncertainty

The NED task is quite demanding, with several cases of variant semantics, insufficient underlying information, or highly ambiguous context. Absolute accuracy may be considered unattainable even for humans on the task. As a result, the confidence evaluation of a named entity prediction is momentous for the development of successful applications. At the pre-output layer of our deep learning model architecture, we could fruitfully exploit the output score as a quality indication for the predicted positive linking compatibility class outcome.

For an anchor a, the candidate mention set size is $|Pg(a)| = k$. Let *compatibility score(m)* denote the compatibility score of mention m. Let the candidate mentions set for anchor a in $Pg(a)$ be denoted as $\{m_1, m_2 \ldots m_k\}$.

Hence, the uncertainty quantification formula can be defined as follows:

$$U(a, m_a) = \frac{compatibility\ score(m_i) - compatibility\ score\left(m_j\right)}{compatibility\ score(m_i)},$$
$$m_i : max(compatibility\ score(m_i \in \{m_1,\ m_2 \ldots m_k\})),$$
$$m_j :\ max(compatibility\ score(m_j \in \{m_1,\ m_2 \ldots m_k\} - \{m_i\})\,.) \tag{5}$$

Equation (5) is an expression of the semantic distance between the selected mention for an anchor annotation and the next most coherent available mention for that annotation in a specific context. This metric was proven as a good uncertainty indication throughout our experimental evaluation.

3.4. Evaluation Process

The evaluation analysis focused on the entity linking disambiguation process, delineating the benefits of our novel methodology. The uncertainty score introduced for our methodology was thoroughly validated as a confidence indicator for the outcome prediction. For performance comparison, a classic precision, recall, and F1 measure assessment was carried out. Specifically, we evaluated precision by depicting the ratio of valid anchor mention annotations over the size of the identified mention ensemble.

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{6}$$

We evaluated recall on the basis of the number of correctly annotated predictions divided by the total predictions made.

$$\text{Recall} = \frac{TP}{|mentions|}. \tag{7}$$

Lastly, the F1 score outlined a harmonic mean between recall and precision.

$$\text{F1} = 2 * \frac{Precision \times Recall}{Precision + Recall}. \tag{8}$$

The wiki-disamb30 dataset, introduced by [15], was utilized by several works in the domain, including [15,19,20,30], and it is generally accepted as a baseline for the task. Our methodology evaluation process was based on segments of the wiki-disamb30 dataset for a thorough performance analysis. This dataset contains approximately 1.4 million short input texts up to 30 words, incorporating at least one ambiguous anchor hyperlink along with its corresponding Wikipedia entity. As the dataset target entity links correspond to an old instance of Wikipedia, some processing is required for updating the references to the current changes. The dataset in use features extensive context variability, as the text segments cover a wide range of topics, making it ample for a thorough assessment.

As this work introduces and studies a specific NED problem, namely, online training NED, our main focus was the evaluation of our methodology, using precision, recall, and F1 measures. However, the established baseline methodologies from [15] along with the systems proposed in [34] were included for an incommensurate yet indicative performance comparison, outlining the performance of our methodology under a common evaluation dataset.

The first baseline, TAGME [15], is a particularly popular and relatively simple to implement methodology featuring computational efficiency. Relatedness, commonness, and other Wikipedia statistics were combined in a voting schema for the selection of the top scoring candidate annotation for a second-step evaluation and selection process. We aimed to extract insights from a comparison with classic baseline high-performance approaches. The second baseline employed was the Clauset–Newman–Moore (CNM) methodology from [34]. This approach introduced community detection algorithms for semantic segmentation of the Wikipedia knowledge graph into densely connected subgraphs, achieving high accuracy. A classification model approach was employed for the task, using established features along with community coherence information derived by the Clauset–Newman–Moore algorithm.

4. Results

Our methodology assessment was performed using Wikipedia snapshots for the knowledge extraction process. Specifically, the enwiki-20191220-pages-articles dump of pages from 20 December 2019 [35] was employed for an extraction, transformation, and loading process in the WikiText [36] format using the Apache Spark framework [37]. Big data techniques were eventful for the process, deriving more than 37,500,000 individual anchors and over 19,300,000 entities, composing a graph with over 1,110,000,000 edges. The distributed architecture and processing model of our implementation can handle a far larger scale, and its scalability capabilities allow following the growth rates of Wikipedia datasets. For the NED disambiguation process implementation, we used Keras [30] and TensorFlow [31]. Our experiments employed a distributed 192vCPU and 720 GB memory Google Cloud Platform setup.

4.1. Experimental Analysis Discussion

Our OTNEL implementation was experimentally evaluated as dominant for high precision performance, as outlined in the comparative results of Figure 3 and Table 1. Specifically, the methodology indicatively outperformed the baseline methodologies at full recall by 7%. The inclination of precision–recall in similar works in the generic NED scope was impetuous at recall levels above 0.6. This fact was interpreted as the inadequacy of those methodologies to fit a generic model for low-frequency or poorly linked entities in the knowledge graph. Conversely, in the case of our methodology, we observed a more gradual decline in precision to the point of 0.9 recall levels. This not only justified the overall precision of our methodology but also the high-performance certainty evaluation metric employed for recall adjustment, along with the improved modeling capabilities of OTNEL, due to its topical online training. The recall area (0.9, 1] of our method evaluation framed an elevated negative inclination as anticipated toward absolute recall. This was mainly interpreted as knowledge deficiency and a latent modeling approximation of deviating cases.

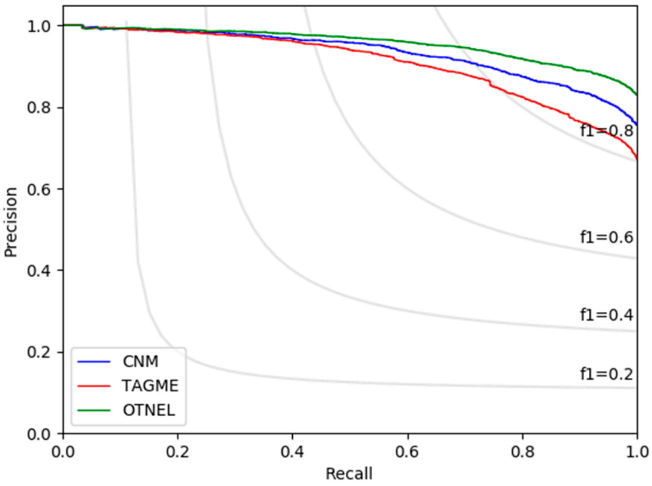

Figure 3. Precision–recall of OTNEL method, compared with Clauset–Newman–Moore (CNM) and TAGME baselines.

Table 1. OTNEL, TAGME, and CNM precision and F1 scores at varying recall levels.

Recall	1.0	0.9	0.8	0.6	0.4	0.2
CNM [1] (Precision)	0.7554	0.8362	0.8738	0.9337	0.9678	0.9866
CNM [1] (F1)	0.8606	0.8667	0.8351	0.7287	0.5659	0.3325
TAGME (Precision)	0.6720	0.7640	0.8242	0.9101	0.9619	0.9832
TAGME (F1)	0.8038	0.8264	0.8118	0.7222	0.5650	0.3323
OTNEL [2] (Precision)	0.8290	0.8897	0.9180	0.9589	0.9789	0.9896
OTNEL [2] (F1)	0.9065	0.8948	0.8548	0.7381	0.5678	0.3327

[1] WSD methodology based on Clauset-Newman-Moore Community detection; [2] Online Training Named Entity Linking.

Overall, our deep learning architecture consisted of a multilevel LSTM network. The recurrent learning selection was driven through a delayed reward with a global context within the topic confinement. Furthermore, the utilization of our penultimate layer score apparently yielded considerable insights into the success of certainty scoring, contributing to a progressive precision recall inclination. Again, we could observe high precision even at high recall over 0.9. For a dataset featuring such context variability, as training was conducted using Wikipedia, the extraordinary performance and potential of our approximation is profound.

The F1 score had a local maximum of approximately 91% of recall for the OTNEL model, as shown in Figure 4. The influence of topical segmentation introduced by our online training methodology, in conjunction with the high-performance indicator of linking certainty in the big data scale of the evaluation, emphasizes a consistently high performance, as clearly illustrated for the area over 0.8 of the recall axis. The value of our modeling approach is emphasized by the impressive accuracy even at high recall levels.

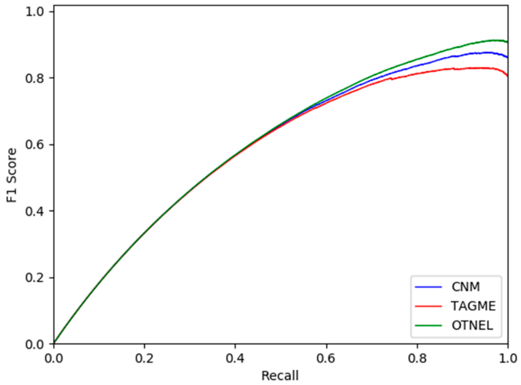

Figure 4. F1 score of OTNEL method, compared with CNM and TAGME baselines.

4.2. *Quantification of Certainty Evaluation*

Certainty was modeled as a measure of confidence for a mention selection. The correlation of certainty score and prediction score are outlined in the two-dimensional (2D) histograms in Figures 5 and 6. In Figure 5, we can observe a dense distribution of high certainty scores and a strong correlation of high prediction scores with high certainty. On the contrary, Figure 6 presents a less dense distribution of low certainty, in the areas of certainty below 0.6 and prediction score below 0.5. This intriguing observation can be interpreted as a knowledge deficit in the knowledge acquisition process, probably due to the coverage of our training set. Another reading of Figure 6 could delineate the presence of outliers; however, the apparent correlation of low certainty with low prediction score clearly indicates our model's advanced capabilities. At this point, it is worth noting that the analogy of valid (true positive) and invalid (false positive) entity link predictions was highly inclined toward true positives, as outlined on Figures 3 and 4, and the visualization of certainty and prediction scores validates our intuitions. Overall, the certainty metric performance as a measure of confidence for the validity of a linked entity was outstanding.

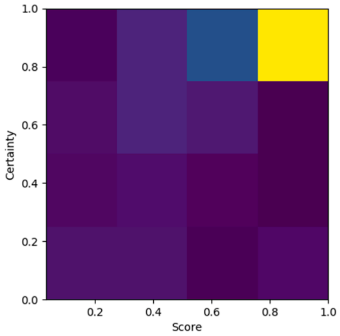

Figure 5. Prediction score–certainty score two-dimensional (2D) histogram: true positive distribution.

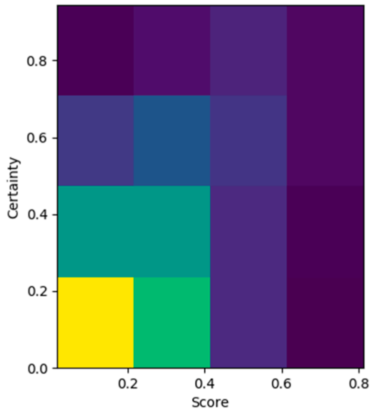

Figure 6. Prediction score–certainty score 2D histogram: false positive distribution.

5. Conclusions and Future Work

The current work proposed an innovative methodology featuring a deep learning classification model for the NED task, introducing the novel concept of online topical training for attaining high performance whilst maintaining rich semantic input specificity. This work introduced and studied the domain of online training NED. Moreover, to the best of our knowledge, this is the first approximation of Wikification and NED leveraging online topical training, introducing a stacked LSTM deep learning architecture model for the task. Our thorough experimental assessment revealed astounding performance in terms of precision, at moderate computational requirements, due to our simplicity-oriented dimensionality and modeling approach.

Our overall deep learning architecture permeates nonlinear input relation modeling, as the LSTM layers involved enable the exploitation of a dynamically changing contextual window over the input sequence history during the online topical training process. As a result, the use of a limited set of established features from works in the domain was adequate for attaining superior deep semantic inference capabilities with a topical focus, successfully addressing high-dimensional-space performance difficulties on a challenging task.

Among our plans for future enhancement of the current work's promising results, further analysis and experimentation in the quest for a more accurate architecture will be considered. A noteworthy advantage of the proposed neural network architecture is its understandability and neural network opacity via a simple model for delineating the benefits of the topical confinement concept in the online training NED task. As entity linking and NED tasks are based on knowledge, their underlying adversity is discerned in the absence of semantically linked corpora, namely, the knowledge acquisition bottleneck. An unsupervised machine learning knowledge expansion approximation could lead to more accurate results and, thus, knowledge acquisition closure from both structured and unstructured knowledge sources and corpora. The incorporation of an unstructured knowledge source via an ensemble learning approach for mitigating the impact of superinducing noise in the knowledge acquisition phase is among our plans.

In this article, our primary focus was the evaluation of new concepts for lowering the computational feasibility barrier to employing deep learning architectures in the NED task, while maintaining input semantic entropy by avoiding vast input transformations and granularity loss. Our extensive experimentation revealed propitious results, placing the introduced methodology in the limelight for further study and broad adoption.

Author Contributions: Conceptualization, C.M. and M.A.S.; methodology, M.A.S.; software, M.A.S.; validation, M.A.S.; formal analysis, M.A.S.; investigation, M.A.S.; resources, C.M.; data curation, M.A.S.; writing—original draft preparation, M.A.S.; writing—review and editing, C.M. and M.A.S.; visualization, M.A.S.; supervision, M.A.S.; project administration, C.M.; funding acquisition, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: Christos Makris was co-financed by the European Union and the Greek national funds through the Regional Operational Program "Western Greece 2014–2020", under the Call "Regional Research and Innovation Strategies for Smart Specialization—RIS3 in Information and Communication Technologies" (project: 5038701 entitled "Reviews Manager: Hotel Reviews Intelligent Impact Assessment Platform").

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khalid, M.A.; Jijkoun, V.; de Rijke, M. The impact of named entity normalization on information retrieval for question answering. In *Advances in Information Retrieval*; Macdonald, C., Ounis, I., Eds.; Springer: Berlin, Germany, 2008; Volume 4956, pp. 705–710.
2. Chang, A.X.; Valentin, I.S.; Christopher, D.M.; Eneko, A. A comparison of Named-Entity Disambiguation and Word Sense Disambiguation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, 23–28 May 2016; pp. 860–867.
3. Dorssers, F.; de Vries, A.P.; Alink, W. Ranking Triples using Entity Links in a Large Web Crawl—The Chicory Triple Scorer at WSDM Cup 2017. Available online: https://arxiv.org/abs/1712.08355 (accessed on 28 August 2020).
4. Artiles, J.; Amigó, E.; Gonzalo, J. The role of named entities in web people search. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; Volume 2, pp. 534–542.
5. Blanco, R.; Ottaviano, G.; Meij, E. Fast and Space-Efficient Entity Linking for Queries. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15), Shanghai, China, 31 January–6 February 2015; pp. 179–188.
6. Dietz, L.; Kotov, A.; Meij, E. Utilizing Knowledge Graphs in Text-centric Information Retrieval. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM'17), Cambridge, UK, 6–10 February 2017; pp. 815–816.
7. Chair-Carterette, B.G.; Chair-Diaz, F.G.; Chair-Castillo, C.P.; Chair-Metzler, D.P. Entity linking and retrieval for semantic search. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM'14), New York, NY, USA, 24–28 February 2014; pp. 683–684.
8. Navigli, R. Word sense disambiguation. *ACM Comput. Surv.* **2009**, *41*, 1–69. [CrossRef]
9. Gale, W.A.; Church, K.W.; Yarowsky, D. A method for disambiguating word senses in a large corpus. *Lang. Resour. Eval.* **1992**, *26*, 415–439. [CrossRef]
10. Mihalcea, R.; Csomai, A. Wikify! Linking Documents to Encyclopedic Knowledge. In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisbon, Portugal, 6–10 November 2007; pp. 233–242.
11. Silviu, C. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 28–30 June 2007; pp. 708–716.
12. Milne, D.N.; Witten, I.H. Learning to link with wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08), Hong Kong, China, 2–6 November 2008; pp. 509–518.
13. Milne, D.; Witten, I.H. An Effective, Low-Cost Measure of Semantic Relatedness obtained from Wikipedia Links. In Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI), Chicago, IL, USA, 13 July 2008; pp. 25–30.
14. Sayali, K.; Amit, S.; Ganesh, R.; Soumen, C. Collective annotation of Wikipedia entities in web text. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09), Paris, France, 28 June–1 July 2009; pp. 457–466.

15. Paolo, F.; Ugo, S. TAGME: On-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10), Toronto, Canada, 26–30 October 2010; pp. 1625–1628.
16. Johannes, H.; Mohamed, A.Y.; Ilaria, B.; Hagen, F.; Manfred, P.; Marc, S.; Bilyana, T.; Stefan, T.; Gerhard, W. Robust Disambiguation of Named Entities in Text. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh, UK, 27–31 July 2011; pp. 782–792.
17. Han, X.; Sun, L.; Zhao, J. Collective entity linking in web text: A graph-based method. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11), Beijing, China, 25–29 July 2011; pp. 765–774.
18. Makris, C.; Simos, M.A. Novel Techniques for Text Annotation with Wikipedia Entities. In Proceedings of the Artificial Intelligence Applications and Innovations Evaluation—AIAI 2014, Rhodes, Greece, 19–21 September 2014.
19. Ricardo, U.; Axel-Cyrille, N.N.; Michael, R.; Daniel, G.; Sandro, A.C.; Sören, A.; Andreas, B. AGDISTIS—Agnostic Disambiguation of Named Entities Using Linked Open Data. In Proceedings of the Twenty-first European Conference on Artificial Intelligence, Prague, Czech Republic, 18–24 August 2014; pp. 1113–1114.
20. Piccinno, F.; Ferragina, P. From TagME to WAT: A new entity annotator. In Proceedings of the First International Workshop on Entity Recognition & Disambiguation (ERD'14), Gold Coast, Queensland, Australia, 11 July 2014; pp. 55–62.
21. Sun, Y.; Lin, L.; Tang, D.; Yang, N.; Ji, Z.; Wang, X. Modeling mention, context and entity with neural networks for entity disambiguation. In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15), Buenos Aires, Argentina, 25–31 July 2015; pp. 1333–1339.
22. Ikuya, Y.; Hiroyuki, S.; Hideaki, T.; Yoshiyasu, T. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 250–259.
23. Ganea, O.-E.; Hofmann, T. Deep joint entity disambiguation with local neural attention. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
24. Ivan, T.; Phong, L. Improving Entity Linking by Modeling Latent Relations between Mentions. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 1595–1604.
25. Priya, R.; Partha, T.; Vasudeva, V. ELDEN: Improved entity linking using densified knowledge graphs. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 1844–1853.
26. Fang, Z.; Cao, Y.; Li, Q.; Zhang, D.; Zhang, Z.; Liu, Y. Joint Entity Linking with Deep Reinforcement Learning. In Proceedings of the World Wide Web Conference (WWW'19), San Francisco, CA, USA, 13–17 May 2019; pp. 438–447.
27. Avirup, S.; Gourab, K.; Radu, F.; Wael, H. Neural Cross-Lingual Entity Linking. In Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018; pp. 5464–5472.
28. Ilya, S.; Liat, E.-D.; Yosi, M.; Alon, H.; Benjamin, S.; Artem, S.; Yoav, K.; Dafna, S.; Ranit, A.; Noam, S. Fast End-to-End Wikification. Available online: https://arxiv.org/abs/1908.06785 (accessed on 28 August 2020).
29. Wikimedia Update Feed Service. Available online: https://meta.wikimedia.org/wiki/Wikimedia_update_feed_service (accessed on 28 August 2020).
30. Keras: The Python Deep Learning API. Available online: https://keras.io (accessed on 28 August 2020).
31. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
32. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

33. Farzad, A.; Mashayekhi, H.; Hassanpour, H. A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Comput. Appl.* **2017**, *31*, 2507–2521. [CrossRef]
34. Christos, M.; Georgios, P.; Michael, A.S. Text Semantic Annotation: A Distributed Methodology Based on Community Coherence. *Algorithms* **2020**, *13*, 160. [CrossRef]
35. Index of /Enwiki/. Available online: https://dumps.wikimedia.org/enwiki (accessed on 28 August 2020).
36. Specs/wikitext/1.0.0 MediaWiki. Available online: https://www.mediawiki.org/wiki/Specs/wikitext/1.0.0 (accessed on 28 August 2020).
37. Matei, Z.; Mosharaf, C.; Michael, J.F.; Scott, S.; Ion, S. Spark: Cluster computing with working sets. In Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10), Boston, MA, USA, 22 June 2010.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.