Smart Innovation, Systems and Technologies 98

Giuseppe De Pietro · Luigi Gallo
Robert J. Howlett · Lakhmi C. Jain
Ljubo Vlacic *Editors*

Intelligent Interactive Multimedia Systems and Services

Proceedings of 2018 Conference

International

Springer

Smart Innovation, Systems and Technologies

Volume 98

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

More information about this series at http://www.springer.com/series/8767

Giuseppe De Pietro · Luigi Gallo
Robert J. Howlett · Lakhmi C. Jain
Ljubo Vlacic
Editors

Intelligent Interactive Multimedia Systems and Services

Proceedings of 2018 Conference

Springer

Editors
Giuseppe De Pietro
Istituto Di Calcolo E Reti Ad Alte Prestazioni (Icar)
National Research Council
Rome, Italy

Luigi Gallo
Istituto Di Calcolo E Reti Ad Alte Prestazioni (Icar)
National Research Council
Rome, Italy

Robert J. Howlett
Bournemouth University
Poole, UK

and

KES International
Shoreham-by-Sea, UK

Lakhmi C. Jain
Centre for Artificial Intelligence, Faculty of Engineering and Information Technology
University of Technology Sydney
Sydney, NSW, Australia

and

Faculty of Science, Technology and Mathematics
University of Canberra
Canberra, ACT, Australia

and

KES International
Shoreham-by-Sea, UK

Ljubo Vlacic
Griffith Sciences - Centres and Institutes
Griffith University
South Brisbane, QLD, Australia

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-3-319-92230-0 ISBN 978-3-319-92231-7 (eBook)
https://doi.org/10.1007/978-3-319-92231-7

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume presents a series of carefully selected papers on the theme of Intelligent Interactive Multimedia Systems and Services (IIMSS-18), but also includes contributions on Innovation in Medicine and Healthcare (InMed-18) and Smart Transportation Systems (STS-18).

The papers were presented at the Smart Digital Futures 2018 multi-theme conference, which grouped conferences on Agent and Multi-Agent Systems: Technologies and Applications (AMSTA-18), Intelligent Decision Technologies (IDT-18), Innovation in Medicine and Healthcare (InMed-18), Smart Education and E-Learning (SEEL-18), Smart Transportation Systems (STS-18) together with IIMSS-18 in one venue in Gold Coast, Australia, during 20–22 June 2018.

IIMSS-18 included sessions on 'Cognitive Systems and Big Data Analytics', 'Data Processing and Secure Systems', 'Innovative Information Services for Advanced Knowledge Activity', 'Autonomous System' and 'Image Processing'. InMed-18 papers cover major areas of 'Digital Architecture for Internet of Things, Big data, Cloud and Mobile IT in Healthcare' and 'Advanced ICT for Medical and Healthcare'. STS-18 papers provide a comprehensive overview of various aspects of current research into intelligent transportation technology.

We would like to acknowledge and thank all those who made the conference possible through their hard work. We are grateful to the Programme Co-Chairs, the General Track Chairs, the International Programme Committee members and reviewers for their valuable efforts in the review process, thereby helping us to guarantee the highest quality possible for the conference. We would also like to thank the organisers and chairs of the special sessions which make an essential contribution to the success of the conference.

Lastly, we would like to thank all the authors, presenters and delegates for their valuable contribution in making this an extraordinary event.

We hope and intend that this volume will make a significant contribution to international research in the leading edge topics of the conferences.

R. J. Howlett
L. Gallo
Y.-W. Chen
X. Qu
L. C. Jain

Contents

Intelligent Interactive Multimedia: Systems and Services (KES-IIMSS-18) Introduction

Intelligent Interactive Multimedia: Systems and Services (KES-IIMSS-18) Introduction

We introduce to you a series of carefully selected papers presented during the 11th KES International Conference on Intelligent Interactive Multimedia Systems and Services (IIMSS-18).

At a time when computers are more widespread than ever, and computer users range from highly qualified scientists to non-computer expert professionals, Intelligent Interactive Systems are becoming a necessity in modern computer systems. The solution of "one-fits-all" is no longer applicable to wide ranges of users of various backgrounds and needs. Therefore, one important goal of many intelligent interactive systems is dynamic personalization and adaptivity to users. Multimedia Systems refer to the coordinated storage, processing, transmission and retrieval of multiple forms of information, such as audio, image, video, animation, graphics, and text. The growth rate of multimedia services has become explosive, as technological progress matches consumer needs for content.

The conference took place as part of the Smart Digital Futures 2018 multi-theme conference, which groups AMSTA, IDT, InMed, SEEL, STS with IIMSS in one venue. It was a forum for researchers and scientists to share work and experiences on intelligent interactive systems and multimedia systems and services. It included a general track and four invited sessions.

The invited session "Cognitive Systems and Big Data Analytics" (Chaps. 1–3) specifically focuses on Big Data security, word representations in a machine-readable way, and interactive virtual environments. The invited session "Data Processing and Secure Systems" (Chaps. 4–6) focuses on models, techniques, and algorithms capable of analysing, mining and processing both critical and social data. Differently, the invited session "Innovative Information Services for Advanced Knowledge Activity" (Chaps. 7–13) discusses models, techniques and algorithms for the detection and recognition of human activities, as well as sensing architectures for control systems. The invited session "Autonomous System" (Chaps. 14–16) considers theoretical and practical issues in the design of intelligent and autonomous systems. Finally, the general track (Chapter 17) focuses on image processing, specifically on novel approaches to saliency detection.

Our gratitude goes to many people who have greatly contributed to putting together a fine scientific program and exciting social events for IIMSS 2018. We acknowledge the commitment and hard work of the program chairs and the invited session organizers. They have kept the scientific program in focus and made the discussions interesting and valuable. We recognize the excellent job done by the program committee members and the extra reviewers. They evaluated all the papers on a very tight schedule. We are grateful for their dedication and contributions. We could not have done it without them. More importantly, we thank the authors for submitting and trusting their work to the IIMSS conference.

We hope that readers will find in this book an interesting source of knowledge in fundamental and applied facets of intelligent interactive multimedia and, maybe, even some motivation for further research.

Giuseppe De Pietro
Luigi Gallo
Robert J. Howlett
Lakhmi C. Jain
Ljubo Vlacic

Organization

IIMSS-18 International Programme Committee

Awais Ahmad	Yeungnam University, Korea
Flora Amato	Università degli Studi di Napoli Federico II, Italy
Marco Anisetti	University of Milan, Italy
Koichi Asakura	Daido University, Japan
Monica Bianchini	Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, Università degli Studi di Siena, Italy
Francesco Bianconi	University of Perugia, Italy
Alfredo Cuzzocrea	University of Trieste and ICAR-CNR, Italy
Ernesto Damiani	Khalifa University, Abu Dhabi, UAE
Dinu Dragan	University of Novi Sad, Faculty of Technical Sciences, Serbia
Massimo Esposito	Institute for High Performance Computing and Networking (ICAR) of the National Research Council of Italy (CNR)
Margarita Favorskaya	Reshetnev Siberian State University of Science and Technology, Russia
Christos Grecos	Central Washington University, USA
Katsuhiro Honda	Osaka Prefecture University, Japan
Ignazio Infantino	ICAR-CNR, Italy
Gwanggil Jeon	Incheon National University, Korea
Dimitris Kanellopoulos	University of Patras, Greece
Chengjun Liu	New Jersey Institute of Technology, USA
Antonino Mazzeo	University of Naples Federico II, Italy
Giovanni Luca Christian Masala	Plymouth University, UK
Marian Cristian Mihăescu	University of Craiova, Romania
Aniello Minutolo	National Research Council of Italy, Italy
Vincenzo Moscato	University of Naples Federico II, Italy
Francesco Piccialli	University of Naples Federico II, Italy
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Antonio M. Rinaldi	Università degli Studi di Napoli Federico II, Italy
Milan Simic	RMIT University, School of Engineering, Australia
Maria Spichkova	RMIT University, School of Science, Australia
Toyohide Watanabe	Nagoya Industrial Science Research Institute, Japan

Big Data Security on Cloud Servers Using Data Fragmentation Technique and NoSQL Database

Nelson Santos(✉) and Giovanni L. Masala

Big Data Group, School of Computing, Electronics and Mathematics, Plymouth University, Plymouth PL4 8AA, UK
nelson.santos@students.plymouth.ac.uk, giovanni.masala@plymouth.ac.uk

Abstract. Cloud computing has become so popular that most sensitive data are hosted on the cloud. This fast-growing paradigm has brought along many problems, including the security and integrity of the data, where users rely entirely on the providers to secure their data. This paper investigates the use of the pattern fragmentation to split data into chunks before storing it in the cloud, by comparing the performance on two different cloud providers. In addition, it proposes a novel approach combining a pattern fragmentation technique with a NoSQL database, to organize and manage the chunks. Our research has indicated that there is a trade-off on the performance when using a database. Any slight difference on a big data environment is always important, however, this cost is compensated by having the data organized and managed. The use of random pattern fragmentation has great potential, as it adds a layer of protection on the data without using as much resources, contrary to using encryption.

Keywords: Cloud security · Data fragmentation · NoSQL database
Big data

1 Introduction

Cloud computing can be considered one of the most promising technology for IT applications. It is defined by NIST [1] as the model that enables on-demand access to a pool of resources (e.g., networks, storage, applications, and services) that can be rapidly provisioned with minimal effort from the service provider. This technology is growing in such a way that most modern applications are delivered as hosted services. Such services are divided into Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). This scenario has two main cornerstones: virtualization and distributed computing. They provide many benefits including terms of flexibility, elasticity and resource management. Big data is a big adept of this technology, as customers take advantage of the features offered to utilize and pay the resources needed to accommodate the business model and extend such resources when required [2]. This allows the customers to reduce the cost of the storage and computing clusters, as well deviate from the maintenance of the infrastructure and shift all the focus to the development [3].

G. De Pietro et al. (Eds.): KES-IIMSS-18 2018, SIST 98, pp. 5–13, 2019.
https://doi.org/10.1007/978-3-319-92231-7_1

Despite its benefits, cloud computing also brings many challenges. Among them, is the protection of the data and the privacy of the user. In cloud computing, the user's information is handed to the cloud provider and they are responsible for the storage and safekeeping of the data, often without disclosing their procedures to the end-user [2–5]. Furthermore, storing all the data with a single provider, along with the large number of mining algorithms available, leaves users susceptible to mining attacks from attackers with unauthorized access to the cloud and escalated privileges [6].

This paper investigates the use of random pattern fragmentation [7, 8] on different cloud providers, to add a layer of security on the data, by measuring the performance to fragment, send and retrieve the data. In addition, a novel approach of managing the fragmented information on a NoSQL (Not Only SQL) database is proposed, with its performance also measured and compared. It will start by investigating the state of the art (Sect. 2), followed by the methodology in Sect. 3. Afterwards, in Sect. 4 the results will be displayed and discussed and compared to similar approaches, to provide a better evaluation of the performance, as well as a better understanding of the benefits and disadvantages of data protection by means of random pattern fragmentation.

2 State of the Art

Encryption schemes present a satisfactory solution to the data privacy problem, however, they are very complex and computationally expensive [9, 10]. Therefore, research has been shifting towards other alternatives. Kapusta et al. [11] attempted to avoid encryption by splitting information on two distinct groups and provide different protection, according to the sensitivity of the data. Dev et al. [6], approached the problem by categorizing and fragmenting data into chunks and store them in different providers, to avoid mining from providers, as well as attackers. Bahramim et al. [9] proposed a lightweight modality for mobile phones, where random pattern fragmentation, based on chaos system, is used to split a JPEG file and store in multi cloud systems. Bahramim et al. [10] investigates the use of databases to store and manage chunks created with the same method and adding a layer of encryption to the database. Lentini et al. [12] measured the performance of different fragmentation techniques on Amazon Web Services and compared them with the AES cryptography.

However, to improve the organization and overall management of the data in the server, it is imperative to use a database. Rafique et al. [13] proposed a mapping strategy that leverages columnar NoSQL databases to perform data encryption at various levels of granularity dynamically. Alsirhanni et al. [14] proposed a technique that stores data in different providers, by splitting into a master cloud that contains indexes of the fragments, and various slave clouds that store the data encrypted in columnar databases. Masala et al. [15] proposed data fragmentation on the cloud environment using a NoSQL approach, based on MongoDB [16] to take advantage of the highly scalable distributed architecture, which is the main characteristic of NoSQL.

The aim of this paper is the comparison of a novel approach (RPFNoSQLDB), having a mixed solution between a random pattern fragmentation approach and a NoSQL database, with a random pattern fragmentation approach (RPF). The NoSQL solution adds a management layer on the scrambled data, offering therefore better scalability.

3 Methodology

3.1 Random Pattern Fragmentation

In the random pattern fragmentation (RPF), originally proposed by [9, 10], but referencing the version implemented in [12], the original file is divided into N chunks and the pattern indexes are created with a random function, in other words, a random permutation of N elements before being stored in split files. The split files, are then saved on a cloud instance. The pattern indexes get stored in the client's machine, to reconstruct the original file when needed. With this technique, the attacker does not possess the knowledge of the random order and therefore cannot reconstruct the file. In the Fig. 1 the method is shown.

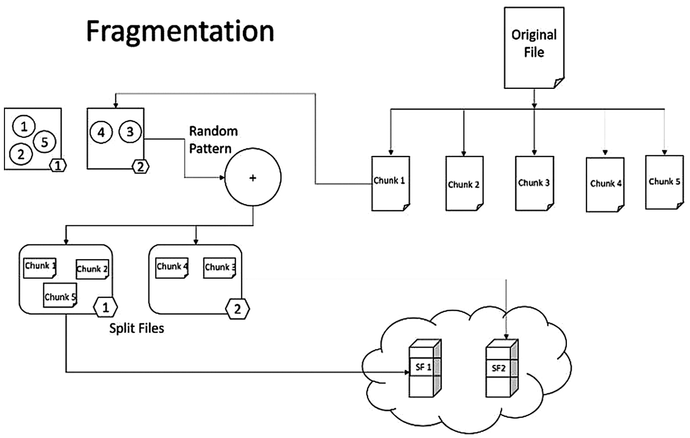

Fig. 1. The process of random pattern fragmentation. The original file gets split into chunks that are stored in split files. The split files are then saved on the cloud server.

In the reconstruction phase (Fig. 2), the split files get downloaded from the cloud and reconstructed using the dictionary format, by combining the stored indexes on the client machine to the different chunks inside the split file. The chunks are then reshuffled back into the original order before being stored back into the client's device.

3.2 The Use of a NoSQL Database Combined with the Random Pattern Fragmentation

We propose a novel approach (Fig. 3) where we combine the use of the combination of the random pattern fragmentation with a NoSQL database (RPFNoSQLDB), where the original file gets split into chunks and those chunks are then inserted to split files. The chosen database management system was CouchDB, version 2.1.1 [17].

The split files are then stored inside the NoSQL database that resides inside an instance on a cloud provider. The data is secured in transit with the use of the virtual

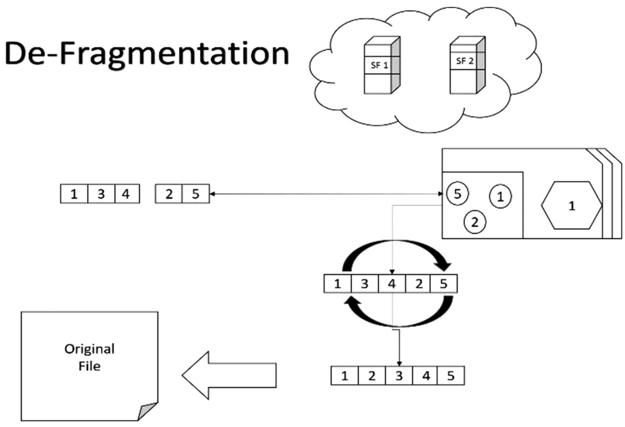

Fig. 2. The process of reconstructing the chunks back to the original file. The file is downloaded from the server and reconstructing using the indexes on the client's machine via a dictionary data structure. After the reconstruction, the file is stored on the client's machine.

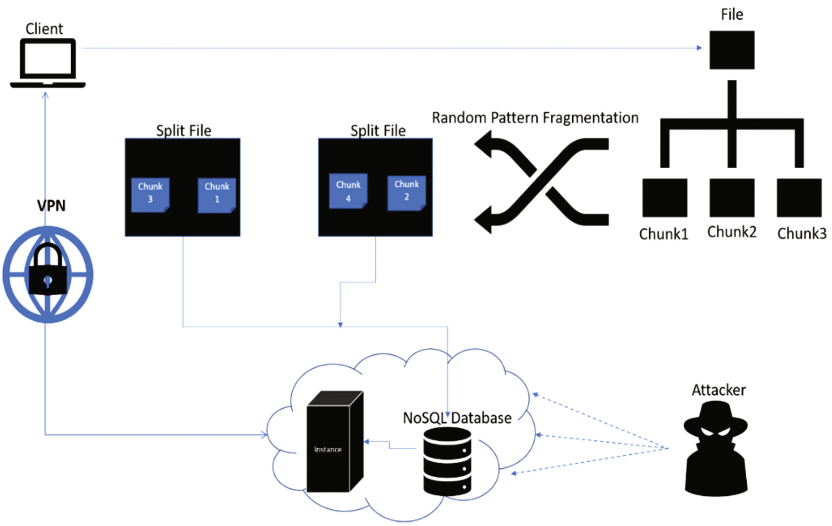

Fig. 3. Proposed model that uses random pattern fragmentation and stores the random chunks in split files, which are then stored on a NoSQL database.

private network (VPN) [18], and in case an attacker accesses the database, the chunks are in a random order, discouraging therefore any attempts to reconstruct the data. The details of the patterns are stored in the client's machine, which are then used to reconstruct the original file.

Using NoSQL to presents an advantage over relational databases, as the files are not structured, making the process of analyzing and retrieving the files faster. In the

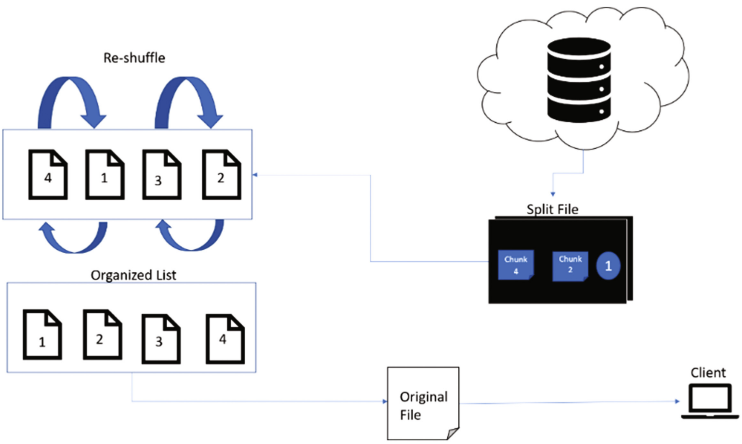

Fig. 4. Process of retrieving and reconstructing the original file. The chunks are sent from the database to the client via a VPN. In the client program, the chunks get re-arranged in a dictionary manner, where the client machine holds the indexes of all the chunks in the correct order.

reconstruction phase, a method based on a dictionary is used, where the client machine uses the stored indexes, combined with the downloaded split files, to re-shuffle the chunks into the correct order, as shown in the Fig. 4.

4 Results and Discussion

In the first part of this paper we are aiming to analyze the performance of using data fragmentation on different cloud providers, as well as the performance of the connection type. This work investigates the performance of the most promising pattern fragmentation technique [12] in a virtual machine hosted by Amazon Web Services (AWS) [19], in comparison with the cloud offered by Microsoft Azure [20]. During the investigation, we always consider sending the files to a single provider via a secure connection. The single provider is the worst-case scenario, as the entire data is available, providing a single point of attack for attackers to mine the data. Nevertheless, we are considering the typical scenario, related to the public cloud.

We are presenting different experiments, using the same algorithm and database in [12], with three different file types (.docx, .jpg, and .pdf), all with 100 KB of size. The result presented in [12] determines that the random pattern fragmentation is faster than the traditional AES encryption [21]. As a result, we are exploring the use of the random pattern fragmentation in the cloud environment.

In the first experiment, we test the random pattern fragmentation approach on a virtual machine in AWS [19] and Azure [20]. The time of splitting a file, storing in a virtual machine, retrieving and reconstructing back to the original file is compared between both providers, in Fig. 5. The communication between the client and the instance is done via tunnel-SSH. In addition, the time of sending a single .docx file, without fragmentation, is highlighted to compare the performance of using the

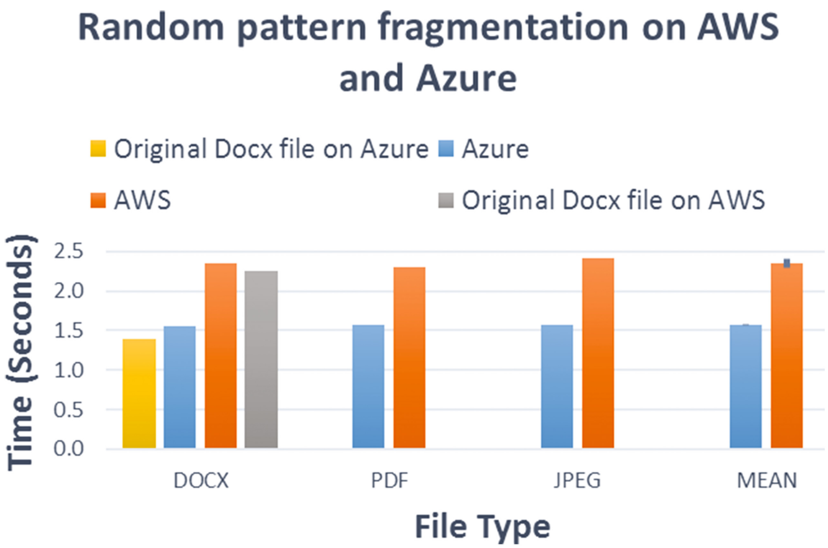

Fig. 5. Performance using tunnel-SSH on two different cloud providers. In the docx file is shown also the difference between sending the original file (called original DOCX) without fragmentation in both providers. On the mean bar is indicated also the standard deviation.

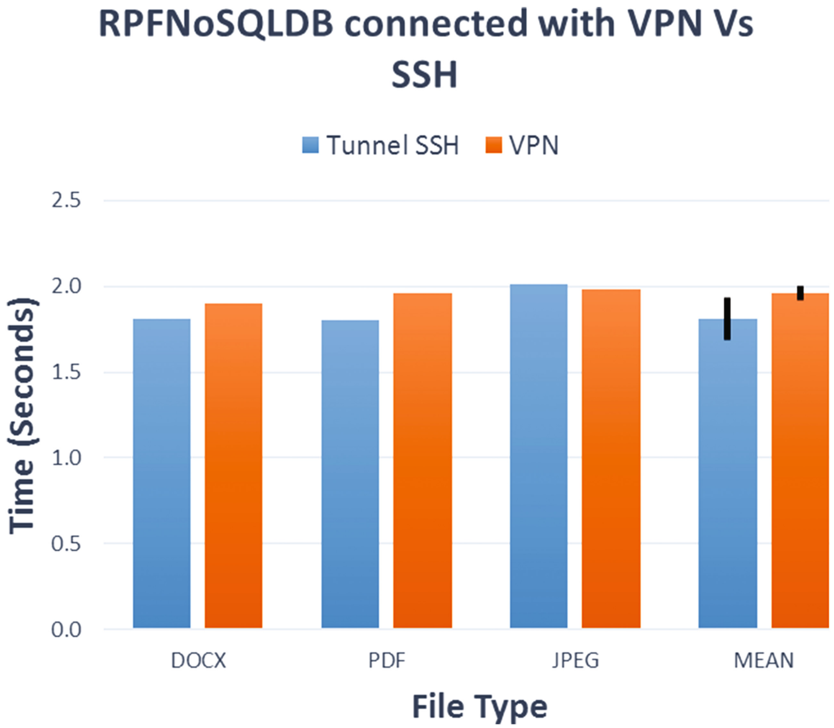

Fig. 6. Performance of the RPFNoSQLDB connecting Azure cloud with tunnel SSH vs VPN.

fragmentation. It is visible, in Fig. 5, that Azure performs better than AWS, with an average of just above 1.5 s (i.e. considering also the sending of the original file without fragmentation).

In the second experiment, we tested the proposed approach RPFNoSQLDB on two different scenarios, regarding the connection between the cloud and the client application. The chosen cloud environment to test the use of the database was Azure. On one hand the program connected to the database using tunnel-SSH, and on the other hand the program interacted with the database using an encrypted Point-to-Site VPN. The results are displayed in Fig. 6. Tunnel-SSH displays slightly better results than its counterpart, however, given the standard deviation calculated in the mean, the difference can be considered neglectable. Nevertheless, using a VPN allows a clear communication channel between the cloud and the client, whereas with the SSH tunnel the client is opening a single connection to the host, complicating the process of transferring multiple files, as well as having multiple users on the application. In addition, with SSH the files are sent sequentially or with multiple connections from the same client, consuming therefore more resources from the server.

In the last experiment, using Azure cloud, in Fig. 7, our proposed method RPFNoSQLDB was compared with the RPF, which does not contain a database. Further details are also published on Table 1.

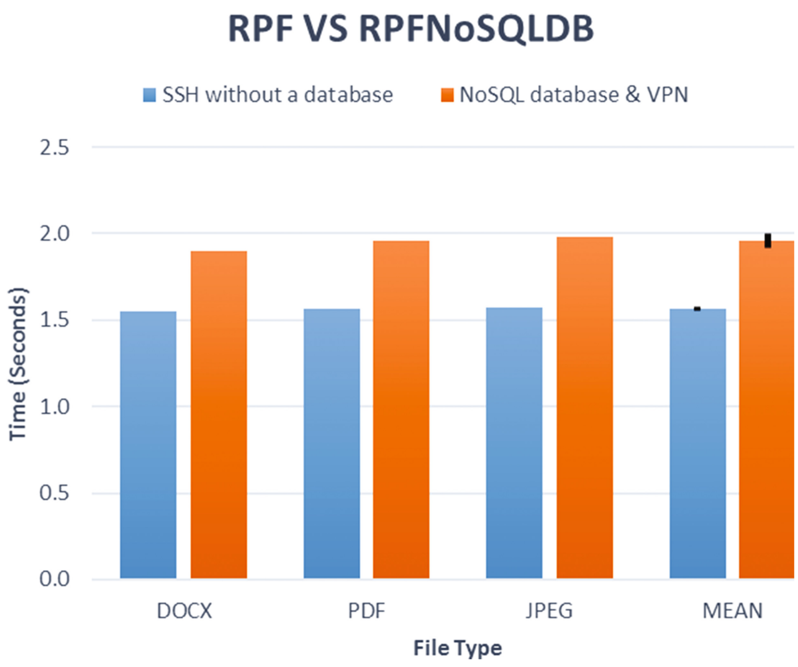

Fig. 7. Comparison of the proposed method RPFNoSQLDB (random pattern fragmentation + NoSQL database), which uses a VPN, with respect to the RPF (random pattern fragmentation without database), with uses a SSH connection.

Table 1. Evaluation of the performance of using RPFNoSQLDB with VPN over sending the files to the instance with respect to the RPF using a SSH connection. It encompasses the time to fragment the file, upload it, download and reconstructing the original file.

File Type	RPFNoSQLDB with VPN	RPF with SSH	Length chunks
100 KB	Time (Seconds)	Time (Seconds)	Bytes
DOCX	1.90	1.55	**1000**
PDF	1.95	1.57	**1000**
JPEG	1.98	1.57	**1000**
MEAN	1.96	1.57	**1000**
ST. DEV	±*0.04*	±*0.01*	**1000**

It can be derived from the Fig. 7 and Table 1 that using a database to manage the fragments affects the performance. On the base of the first two experiments the results don't depend by the connection used (SSH or VPN). Such performance costs are relevant on a big data environment; however, this tradeoff compensates by having the data organized and structured, facilitating the management of the data.

5 Conclusion

Cloud computing offers many advantages in terms of flexibility, scalability and reliability. Nevertheless, it also brings new challenges on security, data privacy and protection. We compared the use of splitting files and shuffling chunks on different cloud environments.

We also proposed a novel method of combining random pattern fragmentation and a NoSQL database (RPFNoSQLDB), to facilitate the organization and management of the data. When applying RPFNoSQLDB, trough the database structure, there is a trade-off on the performance, and the difference is compensated by having the data stored in an organized manner.

Furthermore, the use of a VPN creates a direct channel of communication between the client and the server, encrypted with IPsec, compared to SSH, where the different connections need to be created, to send the fragments without affecting the performance. Future work would include the use of columnar databases and storing the split files in different environments, and in binary large object formats, instead of using document-oriented databases, which store the information in JSON. These techniques show potential to the data security problem, as they add a further layer of security, without using many computing resources, which is not the case when traditional encryption methods like AES are applied.

References

1. NIST, Definition of Cloud Computing. National Institute of Standards and Technology (2011)
2. Cloud Security Alliance: Top threats to cloud computing. Version 1.0 (2010)

3. Bahrami, M., Singhal, M.: The role of cloud computing architecture in Big Data. In: Pedrycz, W., Chen, S.-M. (eds.) Information Granularity, Big Data, and Computational Intelligence. SBD, vol. 8, pp. 275–295. Springer, Cham (2015)
4. Kumar, P., Raj, H., Jelciana, P.: Exploring data security issues and solutions in cloud computing. Procedia Comput. Sci. **125**, 691–697 (2018)
5. Hegarty, R., Haggerty, J.: Extrusion detection of illegal files in cloud-based systems. Int. J. Space Based Situ. Comput. **5**(3), 150–158 (2015)
6. Dev, H., Sen, T., Basak, M., Ali, M.: An approach to protect the privacy of cloud data from data mining based attacks. In: Companion: High Performance Computing, Networking Storage and Analysis, pp. 1006–1115. IEEE, Salt Lake City (2012)
7. Chakraborty, D., Sarkar, P.: A new mode of encryption providing a tweakable strong pseudo-random permutation. In: Robshaw, M. (ed.) FSE 2006. LNCS, vol. 4047, pp. 293–309. Springer, Heidelberg (2006)
8. Gharajedaghi, J.: Systems Thinking: Managing Chaos and Complexity: A Platform for Designing Business Architecture. Elsevier, Boston (2011)
9. Bahramim, M., Singhal, M.: A light-weight permutation based method for data privacy in Mobile Cloud Computing. In: 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, pp. 189–198. IEEE, San Francisco (2015)
10. Bahrami, M., Singhal, M.: CloudPDB: a light-weight data privacy schema for cloud-based databases. In: 2016 International Conference on Computing, Networking and Communications (ICNC), pp. 1–5, Kauai (2016)
11. Kapusta, K., Memmi, G.: Data protection by means of fragmentation in distributed storage systems. In: 2015 International Conference on Protocol Engineering (ICPE) and International Conference on New Technologies of Distributed Systems (NTDS), pp. 1–8. IEEE, Paris (2015)
12. Lentini, S., Grosso, E., Masala, G.: A comparison of data fragmentation techniques in cloud servers. In: Proceedings of International Conference on Emerging Internet, Data and Web Technologies (EIDWT 2018), Tirana (2018)
13. Rafique, A., Van Landuyt, D., Reniers, V., Joosen, W.: Leveraging NoSQL for scalable and dynamic data encryption in multi-tenant SaaS. In: 2017 IEEE Trustcom/BigDataSE/ICESS, pp. 885–892, Sydney (2017)
14. Alsirhani, A., Bodorik, P., Sampalli, S.: Improving database security in cloud computing by fragmentation of data. In: International Conference on Computer and Applications, pp. 43–49. IEEE, Dubai (2017)
15. Masala, G.L., Ruiu, P., Grosso, E.: Biometric authentication and data security in cloud computing. In: Daimi, K. (eds.) Computer and Network Security Essentials. Springer, Heidelberg (2018)
16. MongoDB Homepage. https://www.mongodb.com/. Accessed 3 Mar 2018
17. CouchDB Homepage. http://couchdb.apache.org/. Accessed 3 Mar 2018
18. SANS Institute: Extending your business network through a virtual private network (VPN). SANS Infosec Reading room (2016)
19. AWS Amazon Homepage. https://amazon.com. Accessed 2 Mar 2018
20. Microsoft Azure Homepage. https://azure.microsoft.com/en-gb/. Accessed 2 Mar 2018
21. Federal Information. Announcing the Advanced Encryption Standard (AES). Federal Information Processing Standards Publication 197 (2001)

A Comparison of Character and Word Embeddings in Bidirectional LSTMs for POS Tagging in Italian

Fiammetta Marulli(✉), Marco Pota, and Massimo Esposito

Institute for High Performance Computing and Networking - National Research Council of Italy, Via Pietro Castellino 111, 80131 Naples, Italy
{fiammetta.marulli,marco.pota,
massimo.esposito}@icar.cnr.it

Abstract. Word representations are mathematical items capturing a word's meaning and its grammatical properties in a machine-readable way. They map each word into equivalence classes including words sharing similar properties. Word representations can be obtained automatically by using unsupervised learning algorithms that rely on the distributional hypothesis, stating that the meaning of a word is strictly connected to its context in terms of surrounding words. This assessed notion of context has been recently reconsidered in order to include both distributional and morphological features of a word in terms of characters co-occurrence. This approach has evidenced very promising results, especially in NLP tasks, e.g, POS Tagging, where the representation of the so-called Out of Vocabulary (OOV) words represents a partially solved issue. This work is intended to face the problem of representing OOV words for a POS Tagging task, contextualized to the Italian language. Potential benefits and drawbacks of adopting a Bidirectional Long Short Term Memory (bi-LSTM) fed with a joint character and word embeddings representation to perform POS Tagging also considering OOV words have been investigated. Furthermore, experiments have been performed and discussed by estimating qualitative and quantitative indicators, and, thus, suggesting some possible future direction of the investigation.

Keywords: Deep neural network · Natural Language Processing
POS tagging · Character and word embeddings

1 Introduction

Inspired by deep hierarchical structures of human speech perception and analysis systems, the concept of deep learning algorithms was introduced and widely applied to Natural Language Processing (NLP), currently reaching very surprising and promising results. Among the variety of NLP tasks, text classification and Part of Speech (POS) Tagging took much advantage from adopting deep neural network schemes and methods, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and recursive neural networks. As very recent text classification accuracy-augmenting strategies, current research is exploring the possibilities enabled

G. De Pietro et al. (Eds.): KES-IIMSS-18 2018, SIST 98, pp. 14–23, 2019.
https://doi.org/10.1007/978-3-319-92231-7_2

by joining the exploitation of a Bidirectional Long Short Term Memory (bi-LSTM) fed with both distributional semantic information and morphological features of the processed words.

Recent works and results obtained investigating in this direction are suggesting that this kind of joint approach could reveal particularly effective and useful in the analysis of morphological rich languages, among which the Italian. This kind of hybrid approach could respond to the effective need to fill and reduce the gap in the accuracy when textual snippets and corpus under a real domain analysis include a relevant amount of unfrequently occurring and very specific words. The problem of accounting these words, better known as Out of Vocabulary words (OOV) is relevant in NLP tasks. Typically, a deep neural networks exploit a distributional word representation model built over large but generic language corpora. However, they are trained for text classification tasks on very specific domains, such as cultural heritage, medicine or law and justice, with a notable amount of words for which no word representation exists, implying a worsening in the classification performance.

In this perspective, the main contribution of this work is to perform an accurate evaluation over the potential deep neural network based solutions addressing the general problem of POS tagging accuracy and the particular problem of improving the POS tagging performance when OOV words are present. By extending the experiences documented in the state of the art for POS Tagging, this exploration is addressed to the case of study of the Italian Language.

To this aim, two different strategies for calculating distributional word representations, typically named word embeddings, have been adopted, namely word2vec [1, 2] and FastText [3], both retrained on a textual snippet of Wikipedia Dump dataset for Italian including a collection of about 500.000 words. Moreover, representations of words at a character level, also called character embeddings, have been calculated by adopting a further Bi-LSTM network. The two typologies of word embeddings and the character embeddings have been combined by concatenation to jointly exploit both distributional semantic and morphological features in the POS tagging task. A set of experiments has been arranged and performed aimed at evaluating the effectiveness of applying a bi-LSTM network for POS tagging on the Universal Dependencies v 2.1 (UD) [4] Dataset for Italian, by considering different possible configurations combining word and character embeddings.

The structure of the paper is as follows: Sect. 2 describes the most recent work related to the topic discussed. Section 3 introduces the most significant concepts (the fundamentals) concerning distributed representations for words and characters, simply known as word and character embeddings. Section 4 introduces the system architecture and the POS tagging model adopted for the aim of comparing the joint exploitation of distributed representations at word and character level in a Bidirectional LSTM network. Sections 5 and 6 describe, respectively, the experiments performed and the results obtained.

2 Related Work

Deep learning methods employ multiple processing layers to learn hierarchical representations of data, and have produced state-of-the-art results in many domains. Recently, a variety of model designs and methods have blossomed in the context of NLP. An interesting and comprehensive overview concerning current methods and strategies exploiting deep learning and deep neural networks applied to numerous NLP tasks has been provided in [5]. In this work, authors also provides a walk-through of their evolution.

In [6], a simple deep learning framework is demonstrated to outperform most state-of-the-art approaches in several NLP tasks such as Named Entity Recognition (NER), Semantic Role Labeling (SRL), and POS tagging. Since then, numerous complex deep learning based algorithms have been proposed to solve difficult NLP tasks.

In particular, bi-LSTM networks have been recently employed for many NLP tasks, reaching very promising results in terms of accuracy improvement in POS tagging [7, 8] and transition-based dependency parsing [9, 10]. LSTMs are a particular kind of multilayer RNNs designed for preventing the vanishing gradients problem. bi-LSTMs, indeed, implement a bidirectional process consisting in a backward and forward pass through the sequence before passing on to the next layer, as described in [11].

Another recent trend that has proven successful in POS tagging and word level classification modeling is represented by the proposal discussed in [12] of adopting a multi-level bi-LSTM network: in a first step, word embeddings are built according a compositional criteria employing an inner bi-LSTM working on the character level and, in a second step, the tagging is performed by an outer bi-LSTM over words.

In [13], authors describe several experiments performed in order to evaluate bi-LSTMs with word, characters and Unicode byte embeddings for a POS tagging task, reaching very high accuracy levels over a set of 22 different languages, including the Italian. Moreover, authors adopted a dynamic modeling approach in designing their bi-LSTM, based on the exploitation of DyNet Libraries [14], introducing the possibility of switching the POS tagging loss function with an auxiliary loss function for managing OOV words.

As regard to the Italian Language, TINT [15] can be currently considered as the state of the art tool for POS tagging and dependency parsing tasks. Indeed, the POS tagging model adopted in TINT is based on the Stanford NLP Core, proposed in [16], showing a quite different but performing approach if compared to other current neural network based solutions. Anyway, no mention is made about how to handle OOV or rare words.

In [17], some experiments aimed to compare TINT and Google SyntaxNet [18] approach for Italian, evidenced that, for a particular category of words, that is verbal forms including enclitics pronouns (very usual in the Italian language), TINT failed much more times than SyntaxNet in assigning the correct POS tag to this type of words. Furthermore, the latest version of SyntaxNet, released as ParseySaurus [19], adopts a bi-LSTM network, both for the POS tagger and the dependency parser.

With respect to the problem of handling OOV words in NLP tasks, in [20] was provided an interesting approach aiming to aimed to build, by adopting a miming inverse function, at training time, word embeddings for unknown words. Furthermore, adapting the basic idea to consider sub-word level information as proposed in [12, 13] a dynamic character-based bi-LSTM model is proposed. This allowed learning a compositional mapping from character to word embeddings, thus tackling the OOV problem.

In [13, 20], some potential solutions are discussed to explicitly address efficiently the problem of rare or OOV words, but the common idea to all the current approaches is to take advantage from joining character level and word level information in order assign a weighted representation in the word space distribution to these words instead of a random representation.

These observations have been source of inspiration for this work to further investigate in this direction for Italian language that is currently missing. More in particular, to the best of our knowledge, no approach exists in literature, where word embeddings including distributional semantic and morphological features and character embeddings have been jointly exploited in a bi-LSTM for POS tagging, also effectively handling the problem of OOV words in Italian.

3 Distributed Representations

3.1 Word Embeddings

Word embeddings [5, 21] essentially express the distributional hypothesis, according to which words with similar meanings tend to occur in similar context. The main advantage of distributional vectors is that they capture similarity between the neighbors of a word. Similarity between vectors can be measured in more than one way; the cosine similarity represents one of the most effectively employed measure. Furthermore, word embeddings are often used as the first data processing layer in a deep learning model.

Typically, word embeddings are pre-trained by optimizing an auxiliary objective in a large unlabeled corpus [21, 22] and the learned word vectors can capture general syntactical and semantic information. These embeddings have proven to be proficient in capturing context similarity, analogies and due to its smaller dimensionality, are characterized by fastness and efficiency in computing core NLP tasks.

The first work showing the utility of pre-trained word embeddings is described in [7]. The authors proposed a neural network architecture that forms the foundation to many current approaches. The work also establishes word embeddings as a useful tool for NLP tasks. However, the immense popularization of word embeddings was arguably due to [1, 2] who proposed the continuous bag-of-words (CBOW) and skipgram models to efficiently construct high-quality distributed vector representations.

The most popular word embedding methods were represented, until the few past years, by the Mikolov's word2vec [1, 2] and by Pennington's Glove [23], which is essentially a "count-based" model.

3.2 Character Embeddings

Character embeddings [5] allow to represent each word as no more than a composition of individual letters. Differently from word embeddings, which are able to capture syntactic and semantic information, character embeddings can capture intra-word morphological and shape information. Generally speaking, building NLP systems at the character level has attracted certain research attention [12, 24]. Better results on morphologically rich languages are reported in certain NLP tasks.

In [9], authors exhibit positive results on building a neural language model using only character embeddings. Bojanowski et al. [3] also tried to improve the representation of words by using character-level information in morphologically rich languages. They approached the skipgram method by representing words as bag-of-characters ngrams. Their work thus had the effectiveness of the skip-gram model along with addressing some persistent issues of word embeddings. The method was also fast, which allowed training models on large corpora quickly. Popularly known as FastText, such a method stands out over previous methods in terms of speed, scalability, and effectiveness.

4 System Architecture

The deep neural network architecture here proposed for the POS tagging task resembles the one proposed in [13, 20]. It is essentially composed by two bi-LSTM networks, used in a cascade configuration, as shown in Fig. 1. In particular, a first bi-LSTM network is responsible of calculating the character embedding for each word in input whereas a second bi-LSTM network is charge of receiving, for each word in input, both the calculated character embedding and the word embedding, combining them and performing the whole POS tagging process.

In more detail, the first bi-LSTM network calculates the character embedding for each input word by employing the compositional model proposed in [12]. More formally, the input of this network is a single word w, and the output is a d-dimensional vector representing w. Additionally, an alphabet of characters C is defined, containing an entry for each uppercase and lowercase letter as well as numbers and punctuation.

The input word w is decomposed into a sequence of characters $c_1, \ldots . c_m$, where m is the length of w. Each c_i is defined as a one hot vector 1_{ci}, with one on the index of c_i in vocabulary C. A projection layer $P_C \epsilon \mathbb{R}^{dC x |C|}$ is defined, where $\mathrm{d_C}$ is the number of parameters for each character in the character set C.

This is, of course, just a character lookup table, and is used to capture similarities between characters in a language (e.g., vowels vs. consonants). For this reason, the projection of each input character c_i can be written as $e_{ci} = P_c \cdot 1_{ci}$.

Thus, given a sequence of character representations $e^C_{c1}, \ldots . e^C_{cm}$ as input, the forward LSTM, yields the state sequence $s^f_0, \ldots . s^f_m$, while the backward LSTM receives as input the reverse sequence, and yields states $s^b_m, \ldots . s^b_0$. Both LSTMs use a different set of

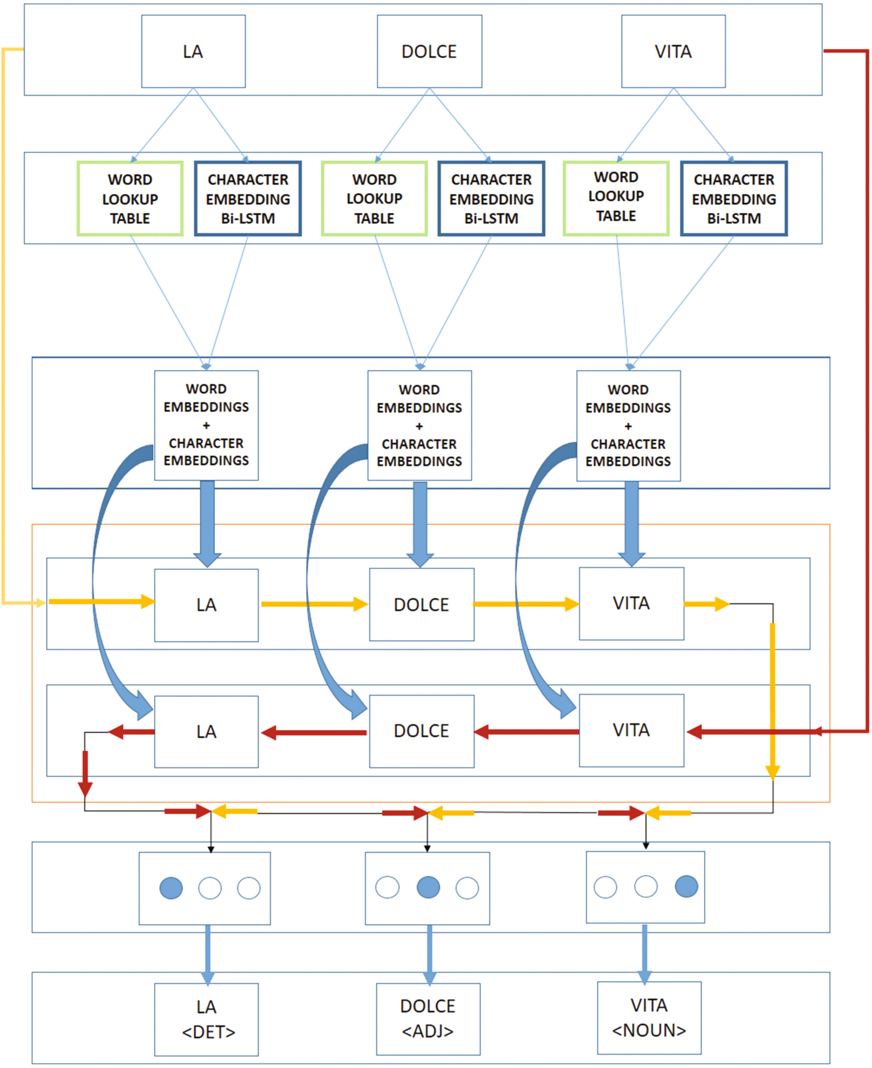

Fig. 1. Deep neural network architecture for POS Tagging.

parameters W^f and W^b. The representation of the word w is obtained by combining the forward and backward states:

$$e_W^C = \mathbf{D}^f s_m^f + \mathbf{D}^b s_0^b + \mathbf{b}_d$$

where D^f, D^b and b_d are parameters that determine how the states are combined.

On the other hand, the second bi-LSTM network operates on a sequence of embeddings $f(w_1), \ldots f(w_n)$, one for each word in input. Each embedding is equal to the one calculated by the previous bi-LSTM, i.e. $f(w_i) = e_{wi}^C$, in case of OOV words, or