Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 115-118 recibido 23-11-2018 revisado 07-01-2019 aceptado 12-02-2019

Cross-view Embeddings para la Recuperación de Información

Cross-view Embeddings for Information Retrieval

Parth Gupta
Pattern Recongnition Human Language Technology (PRHLT) Research Center
Universitat Politècnica de València
Camino de Vera s/n, 46022. Valencia, Spain
pgupta@dsic.upv.es

Resumen: Tesis doctoral en Informática realizada por Parth Gupta bajo la supervisión del Dr. Paolo Rosso (Universitat Politècnica de València) y el Dr. Rafael E. Banchs (Institute for Infocomm Research, Singapore). La tesis se defendió en Valencia (España) el 26 de enero de 2017. El comité de doctorado estuvo compuesto por los siguientes doctores: Eneko Agirre (Universidad del País Vasco), Julio Gonzalo (Universidad Nacional de Educación a Distancia) y Jaap Kamps (Universidad de Amsterdam). La tesis obtuvo la calificación de sobresaliente Cum Laude.
Palabras clave: Recuperación de información, multilingüe, aprendizaje profundo

Abstract: Ph.D. thesis in Computer Science written by Parth Gupta under the supervision of Dr. Paolo Rosso (Universitat Politècnica de València) and Dr. Rafael E. Banchs (Institute for Infocomm Research, Singapore). The thesis was defended in Valencia (Spain) on January 26, 2017. The doctoral committee comprised of the following doctors: Eneko Agirre (University of the Basque Country), Julio Gonzalo (Universidad Nacional de Educación a Distancia) and Jaap Kamps (University of Amsterdam). The thesis got the grade of outstanding *Cum Laude.*
Keywords: Information retrieval, cross-lingual, deep learning

1 Introduction

In this dissertation, we dealt with the cross-view tasks related to information retrieval using embedding methods. Paired instances of data which provide the same information about each datum in different modalities are referred to as cross-view data. For example, parallel sentences are two different views of a sentence in different languages. A word and its transliteration can be seen as two different views of the same word in different scripts. In cross-view tasks, instances of different views are not directly comparable. Under this terminology, CLIR and mixed-script information retrieval (MSIR) can be seen as cross-view retrieval tasks. Broadly, there are two approaches to cross-view tasks: (*i*) translation; and (*ii*) cross-view projection. In translation approaches, one-view is translated into the other view using a translation model and the retrieval is carried using the other view. While, in cross-view projection approaches, data in both views are projected to an abstract common space using dimensionality reduction techniques, where they can be compared. Such representation is also referred to as embeddings. Though translation based approaches provide very rich representation of the data, such approaches are mainly devised for actual translation task such as machine translation (MT) of text from one language to the other. On the other hand, the projection methods provide a representation which may not be interpreted clearly, but provide more flexibility in obtaining representation pertaining to a particular task. For example, it is straight-forward to induce an objective function directly related to the task at hand in the learning mechanism *e.g.* increase cosine similarity between similar documents for a retrieval task. In this dissertation, we explore the cross-view embedding models for cross-view retrieval tasks.

We formally introduced the concept of mixed-script IR, which deals with the challenges faced by an IR system when a lan-

ISSN 1135-5948. DOI 10.26342/2019-62-16

guage is written in different scripts because of various technological and sociological factors. Mixed-script terms are represented by a small and finite feature space comprised of character n-grams. We proposed the cross-view autoencoder (CAE) to model such terms in an abstract space and CAE provides the state-of-the-art performance.

We studied a wide variety of models for cross-language information retrieval (CLIR) and propose a model based on compositional neural networks (XCNN) which overcomes the limitations of the existing methods and achieves the best results for many CLIR tasks such as ad-hoc retrieval, parallel sentence retrieval and cross-language plagiarism detection.

We also explored an effective method to incorporate contextual similarity for lexical selection in machine translation. Concretely, we investigated a feature based on context available in source sentence calculated using deep autoencoders. The proposed feature exhibited statistically significant improvements over the strong baselines for English-to-Spanish and English-to-Hindi translation tasks.

Finally, we explored the methods to evaluate the quality of autoencoder generated representations of text data and analyse its architectural properties. For this, we proposed two metrics based on reconstruction capabilities of the autoencoders: structure preservation index (SPI) and similarity accumulation index (SAI). We also introduced a concept of critical bottleneck dimensionality (CBD) below which the structural information is lost and present analyses linking CBD and language perplexity.

2 Research Questions

In this dissertation, we concretely investigated the following research questions.

RQ1 To what extent mixed-script IR is prevalent in web-search and what is the best way to model terms for it? [Chapter 5]

RQ2 How effective is text representation obtained using external data composition neural network for cross-language IR applications? [Chapter 6]

RQ3 How cross-view autoencoder is useful for lexical selection issue in machine translation? [Chapter 6]

RQ4 How should the number of dimensions in the lowest-dimensional representation of a deep neural network autoencoder be chosen? [Chapter 7]

3 Thesis Overview

The dissertation is organised into four broad blocks: (*i*) we first introduce the background of the main topics of the thesis (Chapters 1, 2 & 3); (*ii*) we present the theoretical models proposed in this dissertation (Chapter 4); (*iii*) we present the evaluation results and analyses for the proposed models on cross-view tasks (Chapters 5 & 6); (*iv*) finally, we present analyses on structural properties for a proposed model (Chapter 7). More details about the organisation of each chapter is presented below.

Chapter 1 presents the introduction and motivation of the thesis. It also highlights the research questions investigated in the thesis along with contributions.

Chapter 2 discusses the theoretical background on information retrieval and dimensionality reduction. It also presents the main challenges and current state-of-the-art around these topics.

Chapter 3 presents necessary background on neural networks, Boltzmann machines, autoencoders and the optimisation methods to understand the technical details of the proposed models.

Chapter 4 presents the main technical contributions of the dissertation and explains the necessary details of the proposed models. We present the proposed cross-view autoencoder based framework to model mixed-script terms and the details of the external-data compositional neural network (XCNN) model.

Chapter 5 presents the details of the mixed-script information retrieval. We first formally define the problem of mixed-script information retrieval with research challenges. We further analyse the query logs of the Bing search engine to understand better the mixed-script queries and their distributions. Finally, we present extensive performance evaluation of the proposed model based on cross-view autoencoder on a standard collection along with other state-of-the-art methods and present insightful analyses.

Chapter 6 presents the evaluation results of the proposed models on cross-language information retrieval tasks such as CL ad-hoc retrieval, parallel sentence retrieval, cross-language plagiarism detection and source context modelling for machine translation. For each application, we first give the description of the problem statement followed by the details of the existing methods. Finally, the comparative evaluation on standard benchmark collections is presented with necessary analysis.

In Chapter 7, we present two metrics, structure preservation index and similarity accumulation index. First, we define these metrics and present the underlying intuition capturing the different aspects of the autoencoder's reconstruction capabilities. With the help of these metrics we define the notion of critical bottleneck dimensionality for the autoencoder. Finally, through the multilingual analysis on a parallel data we show that different languages have different critical bottleneck dimensionalities, which happens to be closely associated with the language grammatical complexities, measured in terms of n-gram perplexities.

Finally in Chapter 8, we draw the conclusions from the dissertation, discuss limitations and outline the future work.

4 Contributions

There are many facets of contributions in this dissertation. For the first time, we introduce the concept of MSIR formally. We also present the deep learning based cross-view models which provide the state-of-the-art performance for modelling mixed-script term equivalents for MSIR. The embedding based cross-view models: (*i*) cross-view autoencoder; and (*ii*) external-data compositional neural network (XCNN) provide state-of-the-art performance for many cross-view tasks such as cross-language ad-hoc IR, parallel sentence retrieval, cross-language plagiarism detection, source context features for machine translation and mixed-script IR. This dissertation also provides insightful information about the structural properties of the autoencoder architecture, which helps to analyse the training process in a more intuitive way. Here are more details on each of them.

4.1 Mixed-script information retrieval

Information retrieval in the mixed-script space, which can be termed as mixed-script IR, is challenging because queries written in either the native or the Roman scripts need to be matched to the documents written in both scripts. Transliteration, especially into Roman script, is used abundantly on the web not only for documents, but also for user queries that intend to search for these documents. Since there are no standard ways of spelling a word in certain non-native scripts, transliterated content almost always features extensive spelling variations; typically a native term can be transliterated into Roman script in very many ways. For example, the word *pahala* ("first" in Hindi and many other Indian languages) can be written in Roman script as *pahalaa, pehla, pahila, pehlaa, pehala, pehalaa, pahela, pahlaa* and so on.

This phenomenon poses a non-trivial term matching problem for search engines to match the native-script or Roman-transliterated query with the documents in multiple scripts taking into account the spelling variations. The problem of MSIR, although prevalent in web search for users of many languages around the world, has received very little attention till date. MSIR presents challenges that the current approaches for solving mono-script spelling variation and NE transliteration in IR are unable to address adequately, especially because most of the transliterated queries (and documents) belong to the *long tail* of online search activity, and hence do not have enough click-through evidence to rely on.

4.2 Cross-view models

We present a principled solution to handle the mixed-script term matching and spelling variation where the terms across the scripts are modelled jointly. Although cross-view autoencoder provides a good way to model mix-script equivalents, it has some limitations in modelling text. In contrast to the most of the existing models which rely only on the comparable/parallel data, our model (external-data compositional neural network – XCNN) takes the external relevance signals such as pseudo-relevance data to initialise the space monolingually and then, with the use of a small amount of parallel data, adjusts the parameters for different languages. There are a few approaches which go beyond the use of only parallel data. The framework also allows the use of click-through data, if available, instead of pseudo-relevance data. Our model, differently from other models, optimises an objective function that is directly related to an evaluation metric for retrieval tasks such as cosine similarity. These two properties prove crucial for XCNN to outperform existing techniques in the cross-language IR setting. We test XCNN on different tasks of CLIR and it attains the best performance in comparison to a number of strong baselines including machine translation based models.

4.3 Critical bottleneck dimensionality

Although deep learning techniques are in vogue, there still exist some important open questions. In most of the studies involving the use of these techniques for dimensionality reduction, the qualitative analysis of projections is never presented. Typically, the reliability of the autoencoder is estimated based on its reconstruction capability.

The dissertation proposed a novel framework for evaluating the quality of the dimensionality reduction task based on the merits of the application under consideration: the representation of text data in low dimensional spaces. Concretely, the framework is comprised of two metrics, structure preservation index (SPI) and similarity accumulation index (SAI), which capture two different aspects of the autoencoder's reconstruction capability. More specifically, these two metrics focus on assessing the structural distortion and the similarities among the reconstructed vectors, respectively. In this way, the framework gives better insight of the autoencoder performance allowing for conducting better error analysis and evaluation. With the help of these metrics, we also define the concept of critical bottleneck dimensionality which refers to the adequate size of the bottleneck layer of an autoencoder.

5 Conclusions and Future Work

This dissertation deals with cross-view projection techniques for cross-view information retrieval tasks. In the exploration, a very important and prevalent problem of mixed-script IR is formally defined and investigated. The deep learning based neural cross-view models proposed in this dissertation provide state-of-the-art performance for various cross-language and cross-script applications. The dissertation also explored the architectural properties of the autoencoders which has attained less attention and establishes the notion of critical bottleneck dimensionality. Some of the most important publications from the thesis work are listed in the References below.

References

Barrón-Cedeño, A., P. Gupta, and P. Rosso. 2013. Methods for cross-language plagiarism detection. *Knowl.-Based Syst.*, 50:211–217.

Franco-Salvador, M., P. Gupta, P. Rosso, and R. E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowl.-Based Syst.*, 111:87–99.

Gupta, P., K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 677–686, New York, NY, USA. ACM.

Gupta, P., R. E. Banchs, and P. Rosso. 2016a. Continuous space models for CLIR. *Information Processing & Management*, 53(2):359–370.

Gupta, P., R. E. Banchs, and P. Rosso. 2016b. Squeezing bottlenecks: Exploring the limits of autoencoder semantic representation capabilities. *Neurocomputing*, 175:1001–1008.

Gupta, P., M. R. Costa-Jussà, P. Rosso, and R. E. Banchs. 2016. A deep source-context feature for lexical selection in statistical machine translation. *Pattern Recognition Letters*, 75:24–29.