Information Sciences 000 (2016) 1–13

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Multi-granularity sequence labeling model for acronym expansion identification

Jie Liu[a,b,*], Caihua Liu[a,b], Yalou Huang[a,b]

[a] College of Computer and Control Engineering, Nankai University, China
[b] College of Software, Nankai University, China

ARTICLE INFO

Article history:
Received 3 September 2015
Revised 4 May 2016
Accepted 25 June 2016
Available online xxx

Keywords:
Conditional random fields
Neural network
Acronym
Sequence labeling
Multi-granularity

ABSTRACT

Identifying expansion forms for acronyms is beneficial to many natural language processing and information retrieval tasks. In this work, we study the problem of finding expansions in texts for given acronym queries by modeling the problem as a sequence labeling task. However, it is challenging for traditional sequence labeling models like Conditional Random Fields (CRF) due to the complexity of the input sentences and the substructure of the categories. In this paper, we propose a Latent-state Neural Conditional Random Fields model (LNCRF) to deal with the challenges. On one hand, we extend CRF by coupling it with nonlinear hidden layers to learn multi-granularity hierarchical representations of the input data under the framework of Conditional Random Fields. On the other hand, we introduce latent variables to capture the fine granular information from the intrinsic substructures within the structured output labels implicitly. The experimental results on real data show that our model achieves the best performance against the state-of-the-art baselines.

1. Introduction

Acronyms (e.g. CRF) are compressed forms of terms, and are used as substitutes of the fully expanded forms (e.g., conditional random fields). In many literature and documents, especially in the scientific and engineering fields, the amount of acronyms is increasing at an astounding rate. By using acronyms, people avoid repeating frequently used long phrases. For example, ‘ROM’ is often used to refer to ‘Read Only Memory’, ‘HIV’ is often used to take the place of the long phrase ‘Human Immunodeficiency Virus’, etc. Such abbreviations or acronyms can convey exactly the same information with less words, which simplifies our writing and reading. However, using acronyms obstructs the readers who do not have the domain-specific knowledge. Acronyms also present serious problems for Natural Language Processing (NLP) and Information Retrieval (IR) algorithms. Acronyms and abbreviations that are not common enough to be a part of daily conversation are typically not in the lexicons and can be considered as misspelled words, meaning they have negative aspects on NLP algorithms. Moreover, the existence of acronyms in text hinders the automatic creation of the very lexicons that are needed. As such, acronyms must be taken care of for related NLP tasks and IR tasks.

The previous acronym search systems are mainly based on two-step methods, i.e. acronym identification and expansion identification. Since the acronym identification is relatively easy to be solved using lexical methods, expansion finding is

[*] Corresponding author.
E-mail address: jliu@nankai.edu.cn (J. Liu).

http://dx.doi.org/10.1016/j.ins.2016.06.045
0020-0255/© 2016 Elsevier Inc. All rights reserved.

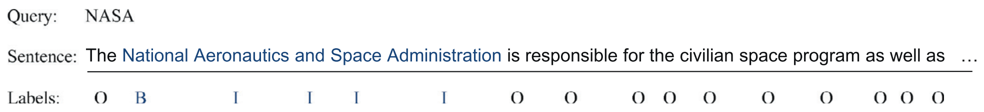
Query: NASA
Sentence: The National Aeronautics and Space Administration is responsible for the civilian space program as well as …
Labels: O B I I I I O O O O O O O O O O

Fig. 1. An example tagging of a training sentence for expansion finding. The label 'B' stands for 'Begin of an expansion', 'I' stands for 'Inside of an expansion', and 'O' stands for 'Others'.

the major bottleneck for many IR and NLP tasks. Previous methods mainly exploit pattern-matching method or supervised learning method. The pattern-matching methods often fail due to the variation of the ways to construct acronyms, it is very difficult to design sufficient and precise rules or patterns to get good precision and recall. Recently, the supervised methods have been widely adopted to overcome the shortcoming of rule-based methods. Machine learning methods that learn the patterns automatically from large corpus have demonstrated the advantages over other kinds of acronym-expansion finding methods. Furthermore, the supervised learning methods based on structured prediction model, i.e. Conditional Random Fields (CRF), have shown to be more appropriate for the expansion identification task [14]. All these methods are faced with two major challenges for this task. One is the variation of the expansion forms. And the other is the complex latent dynamic lies in the expansions.

As the first challenge, the expansion forms vary a lot, e.g. spelling, morphological, syntactic, semantic variations, term synonymy and homonymy. The performance of traditional linear-chain CRFs depends heavily on the quality of the input features which are costly and difficult for human feature engineering. The hand-crafted features are often noisy and redundant, which would constrain the performance of CRF.

More importantly, as the second challenge, there are complex fine-grained structures within the structured output which is a sequence of labels. Such substructures can be essentially regarded as granular structure information in granular computing [39]. Ignoring the fine granular structures may lead to important information loss. As far as expansion identification task concerned, the contexts of acronym expansions often have more complex underlying structures, and the widely used labeling scheme like 'BIO' is too coarse to fully encapsulate the syntactic and query matching behavior of word sequences. Taking the acronym query 'NASA' with the expansion 'National Aeronautics and Space Administration' as an example, as is shown in Fig. 1, both words 'and' and 'Space' are labeled as inside-expansion 'I', they match differently with the query and serve as different roles in the expansion constitution. Hence, the dependence between the neighboring word token pairs (eg. $<$ 'and', 'Space' $>$, and $<$ 'Space', 'Administration' $>$) should be different, even though their labels are identical. In practice, given the limited data, the relationship between specific words and their orthographic or syntactic contexts may be better modeled at a level finer than class labels. In other words, the underlying fine-grained structure is an important intermediate information between input features and labels.

In this paper, we propose a novel multi-granularity sequence labeling model, Latent-state Neural Conditional Random Fields (LNCRF), to deal with the two challenges described above in the problem of acronym expansion identification. Firstly, to alleviate the impact of the variations of the input sentences, we combine Conditional Random Fields (CRF) [9] with non-linear hidden layers which extend CRF to a deep learning model[11] to some extent. Being similar to other deep learning models, the hidden layers empower CRF to learn invariant higher levels features from input sequences automatically. From the point of view of granular computing, such multiple levels of abstractions of data corresponds to hierarchical granularity[36]. Moreover, we further introduce a set of latent-state variables to capture the finer granular structures within the structured output. Using the latent states, the coarse output space is partitioned into finer information granules. Learning the latent variables serves as a seamlessly integrated granulation process of the sequential tokens within sentences for acronym expansion identification. In summary, this new sequence labeling model for acronym expansion identification is able to capture the hidden sub-structures of each class and at the same time learn the non-linear relationships of complex input features and class labels. We evaluate our model on a real dataset collected from wikipedia [14]. Experimental results show that the proposed approach can achieve superior performance against the state-of-the-art baselines.

The rest of this paper is organized as follows. In Section 2, we introduce previous work that is related to this paper. In Section 3, we give a brief formal description of the task. In Section 4, we describe the proposed methodology for expansion finding task, including the Latent-state Neural Conditional Random Fields architecture and the learning algorithm. We then compare the proposed methods with existing representative methods on a real data collected from Wikipedia in Section 5. Finally, conclusions are given in Section 6.

2. Related work

To find acronym expansions, current research methods mainly extract pairs of $<$ acronym, expansion $>$ from texts, for example, *Conditional Random Fields (CRF)*. Most existing methods can be classified into two categories: pattern-matching technique and machine learning based methods. Pattern-matching techniques design rules and patterns to find the longest common substring. Representative works include AFP (Acronym Finding Program) [32] and TLA (Three Letter Acronym)[37]. Recently, machine learning based methods have been preferred as the pattern-matching methods require more human efforts on designing and tuning the rules and patterns, including example-based methods [16,19,35] and sequence-based methods [14,20].

Please cite this article as: J. Liu et al., Multi-granularity sequence labeling model for acronym expansion identification, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.06.045