Chen, W. et al. (Eds.) (2016). Proceedings of the 24th International Conference on Computers in Education.
India: Asia-Pacific Society for Computers in Education

A Competency Similarity Detection for Generating Career Path

Wasan Na Chai [a*], Taneth Ruangrajitpakorn [a b]
Marut Buranarach [a], Thepchai Supnithi [a]

[a]*Language and Semantic Technology Laboratory,*
National Electronics and Computer Technology Center, Pathum Thani, Thailand
[b]*Department of Computer Science, Faculty of Science and Technology,*
Thammasat University, Pathumthanee, Thailand
*wasan.na_chai@nectec.or.th

Abstract: In this paper, we propose a method to detect a similarity of competency to generate a career path. Career path is important for students and workers for their planning in career. In this work, data of competencies from Thailand Professional Qualification Institute are used for generating a relation among units of competency (UoC) as a crossable path for career transfer. Similarity Score using n-gram precision is exploited to find the commonness in UoC context to indicate the possibility in career relation. From an experiment, the proposed method gained 80% accuracy. As a result, career path is generated as a map for a person in career planning for both promotional path and crossable path.

Keywords: Career Path, Competency, Word Matching, Natural Language Processing, Text Similarity

1. Introduction

Previously, a research of career pathing [1] was conducted with the data from TPQI. The work applied string similarity matching to create a career path of similar wording competencies, and it was reported to perform fine in the task. However, we found that there are cases that some competencies are linked as transferrable since strings in those are commonly agree, but they are not semantically related. In details, some look-alike competencies are not technically related in skills or knowledge since they contain many same functional words due to the rich of function words used in Thai. Hence, in this work, we aim to improve an automatic career pathing using TPQI data by using word level similarity for more precision in concern of excessive grammatical word usage.

2. Methodology

This work aims to detect a similarity among competencies to create a career path in different career roles. Unlike the previous work, the detection focuses on word level to find similar words in competencies across career roles. To prevent inclusion of grammatical words in similarity calculation, function words are pruned out. An overview of the proposed method is drawn in Figure 1.

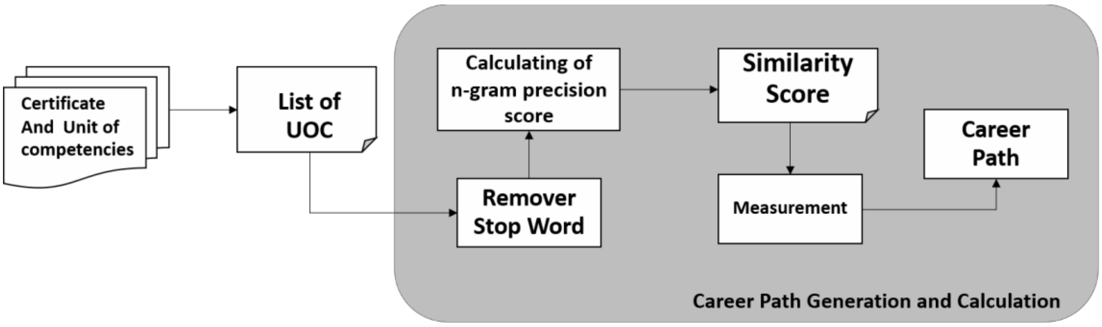
Certificate And Unit of competencies
List of UOC
Calculating of n-gram precision score
Remover Stop Word
Similarity Score
Measurement
Career Path
Career Path Generation and Calculation

Figure 1. An overview of the proposed method

List of certifications and its competency data are collected from TPQI website [2] and stored in our designed database. Structure of data is that professions are provided as a list while each profession (PROF) contains a list of certifications (CERT) numbered for certification level. Certification levels are level 1 - 7, which indicates an ordinal rank of certification. The higher-level number signifies superior certification. In each certification, a list of unit of competencies (UoC) is assigned. It refers to all competencies must be acquired for certification.

Since this work focuses on word level, Thai word boundary segmentation [3] is exploited to handle Thai text in UoC. Function words in UoC are removed because Thai language is rich with function word usage, and they tentatively cause a misleading high similarity score due to their appearance in textual usage. The removed function words [4] are the words with grammatical based part-of-speech including preposition, conjunction, unit classifier, and pronoun.

In this work, we aim to detect commonness of UoCs in CERT to link a relation between CERTs for transferability. In word level, each word in UoC in a CERT is compared to UoCs in another CERT. We apply a BLEU formula [5,6] to calculate word-base similarity. A targeted UoC is referred as candidate, and a compared UoC of another CERT is treated as reference. Applying BLEU measurement metric helps to measures the n-gram (n=1 to 4) precisions of the candidate comparing to the reference. The BLEU score is a product of the geometric mean of the cumulative n-gram precisions and the brevity penalty. BLEU has advantage as:

- focusing on words in the candidate in sequence that only exist in the reference
- balancing for the candidate containing low number of words from obtaining extremely high precision score
- smoothing precisions to prevent a zero score

As similarity measure metric, n-gram of words in UoC is calculated and compared with the reference UoC using (1) to gain n-gram precision (Pn) score.

$$Pn = \frac{\sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')} \qquad (1)$$

With (1), we can obtain a geometric mean of n-gram precision scores comparing between UoCs. In this work, N is set as 1-4 gram, and a ratio to indicate similar UoC is set as 0.80.

3. Experiment

3.1 Experiment Setting

To test the proposed method, collected TPQI data of 126 certifications (CERT) from six professions (PROF) are used as testing data. The chosen professions are *weaving industry, printing industry, vehicle service, transportation service, Thai spa service*, and *hair salon service*. The data contain 291 units of competency (UoC).

For approving the proposed method, we test it against two other methods. The first one is string similarity method [1][7] using in previous work. This method uses (2) to calculate similarity among UoCs.

$$NMCLCS_1(r_i, s_j) = \frac{len(MCLCS_1(r_i, s_j))^2}{len(r_i) \times len(s_j)} \qquad (2)$$

For both methods in this experiment, a ratio to indicate similar UoC is equally set as 0.80. In the proposed method, we set N = 4. The similarity results are approved with a gold standard career path generated by experts and are calculated into accuracy score.

3.2 Experiment Result and Discussion

By comparing the methods, we gained accuracy results shown in Table 1. From Table 1, we found that the proposed n-gram precision in word level with function word removed performed better on pathing between UoCs. We observe into detail of pathing results from the proposed method. The pathing result can be separated into three types: UoC with no matched path to other, UoC matched to another by one-to-one, and UoC matched to many other as many-to-many. From analysis, we obtain insight statistics shown in Table 2. We found that incorrect results are the case of pathing UoCs that are not related and the case of missing path to the related pair, respectively. To summarise types of incorrectness in the proposed system, there are two major issues. The first issue is words that represent specific skill, knowledge and technique are not focused. This issue is the major cause of the incorrect result. Since these technical words are merely a word among many words in UoC, they are equally counted regardless how important they are in the context. Rationally, a detection of these words will highly improve overall accuracy. The second issue is a synonym word using in UoC. Though this issue rarely occurs in this dataset, solving it can correct the missing result in the work.

Table 1: Comparison result of four methods by accuracy matrix

Method	**Accuracy**
String similarity	*54.98%*
n-gram precision in word level with function word removed	**80.07%**

Table 2: Results in details from the proposed method

Type	**Amount**	**Correct**		**Incorrect**			
				no matched	one-to-one	many-to-many	Sum
no matched	182	148 (81.32%)		X	32	2	**34**
one-to-one	71	55 (77.46%)		16	X	0	**16**
many-to-many	38	30 (78.95%)		6	2	X	**8**
Sum	**291**	**233 (80.07%)**		**22**	**34**	**2**	

4. Conclusion and Future Work

This paper presents a method to detect a crossable competency in career for generating paths among job positions. This work is the improved version of the previous work that used string base similarity. In the proposed method, word based similarity with n-gram precision is used to find a commonness in terms using in competency description and the score is used to signify a relation of competencies. From the experiment results, the proposed method obtained the higher accuracy score as around 80% and gained 25% higher accuracy than the previous method.

References

Ruangrajitpakorn T., Na Chai W., Buranarach M., Supnithi T., Kongkachandra R. (2015). An automatic Thai career path generation using similarity of roles and their competencies. In Proceeding of International Symposium on Multimedia and Communication Technology.Thailand.

Thailand Professional Qualification Institute Homepage. Retrieved April 20, 2016, from http://www.tpqi.go.th/en

SegIt : Thai Word Segmentation Tool. Retrieved May 22, 2016, from http://thaimt.org/lstnlp/wordseg.php

Ruangrajitpakorn T., Trakultaweekoon K., Supnithi T. (2009). A Syntactic Resource for Thai: CG Treebank. Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP (pp. 69–102). Singapore.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J.(July 2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 311-318). Philadelphia, USA.

Madnani N. (2011). iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems. Proceedings of the Fifth IEEE International Conference on Semantic Computing. Palo Alto, CA, USA.

A. Islam and D. Inkpen. (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. ACM Transactions on Knowledge Discovery from Data, Vol.2, No.2.