New Trends and Issues Proceedings on Advances in Pure and Applied Sciences

No 12 (2020) 01-07

www.propaas.eu

Determining abbreviations in Kariyer.net domain

Isilay Tuncer*, Kariyer.net Inc, Istanbul, Turkey
Kemal Can Kara, Kariyet.net Inc., Istanbul, Turkey
Askin Karakas, Kariyet.net Inc, Istanbul, Turkey

Suggested Citation:
Tuncer, I., Kara, K. C. & Karakas, A. (2020). Determining abbreviations in Kariyer.net domain. *New Trends and Issues Proceedings on Advances in Pure and Applied Sciences.* (12), 01–07.

Received date January10, 2020; revised date March 15, 2020; accepted date April 11, 2020.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, *Near East University*, Cyprus.

Abstract

In this paper, studies determining abbreviations and their meanings in job texts are explained. The data used in this study consist of job texts stored in the Kariyer.net database. The applied method consists of two separate steps: first, the words and phrases in all job text documents are vectorised with the Word2Vec model. The phrases and abbreviations that are compatible with each other in the proximity of these word vectors are then checked and matched. In the second step, sentences with abbreviations and their meanings in the dataset are defined by the rules determined by Regex. Then, the appropriate abbreviations are collected and added to the dictionary.

Keywords: Word embeddings, text mining, abbreviation detection.

* ADDRESS FOR CORRESPONDENCE: **IsilayTuncer,** Kariyer.net Inc, Istanbul, Turkey.
E-mail address: isilay.tuncer@kariyer.net

1. Introduction

Thousands of job searches, applications and resumes are reviewed by the Kariyer.net search engine every day. Job texts sought by jobseekers are searched as company name, position name, competence and skills. For the candidate to access the correct findings, the searched words and the title and content of the job text must be compatible. On the other hand, Kariyer.net calculates the conformity between the job seekers' resume and the job text while producing a percentage rate accordingly. In order to produce this ratio correctly and to be able to direct the employers and job seekers, it is very important to understand the information in the job text and resume correctly.

Understanding what words or phrases mean in the text is not an easy task in text analysis. Abbreviations play a major role in this task. An abbreviation is defined in the Grammar Terms Dictionary as 'the abbreviated form of the frequently used words, personal, place and organisation names for practical purposes such as saving space and providing convenience' [8]. It is possible to benefit from this convenience due to the fact that computer programmes can understand the meaning of text analysis. Accurate recognition of abbreviations and their definitions is crucial to understanding documents and extracting information from them [2].

The aim of this project is to identify the abbreviations and their meanings in the job texts in order to reach the correct results in searches on the search engine and to calculate the suitability of the job text and resumes.

In this study, which consists of three separate steps, all job texts first go through cleaning. Word2Vec model is used to convert the cleared words and word groups into a vector value. The closeness of the word vectors is examined and the abbreviation and the matching word group are made. In the second step, the sentences with the abbreviation and explanation with the rules determined by Regex are defined and their suitability is checked. Abbreviations and word groups that provide the desired compliance are added to the dictionary. As a result, the abbreviations collected from both steps are combined and the repeated words are removed from the dictionary; then a dictionary of abbreviations, with abbreviations and expansions, in the Kariyer.net domain is created.

2. Similar studies

It is important to determine and understand the abbreviations correctly in order to reach meaningful results from the analysis of the whole text.

Abbreviations are widely used in clinical texts, especially in the medicine and pharmaceutical industry [12]. For this reason, studies that determine abbreviations and expansions with regard to patient, hospital and drug data are frequently encountered in the literature. An example of natural language processing systems, which was developed using only clinical data, is the medical language determination and coding system MedLEE [5], which was developed by Carol [5]. At Columbia University, cTAKES [11] is used to obtain clinical concepts from radiological reports; the Knowledge Map concept identifier [3] can match clinical documents with concepts in the Unified Medical Language System, Pittsburgh SPIN information extraction system [9] and Harvard HITEx health information text extraction system [13].

Apart from the medical literature, Twitter data is often used to conduct sociological studies. There are not only abbreviations that have a real meaning in Twitter texts, there are also abbreviations that are abbreviated due to character restrictions, which are usually created by extracting vowels, or abbreviations that are created using special characters for words or conjunctions, for example, bakiniz (bkz) and tamam (tmm). The original form of the abbreviated word is needed to analyse similar semantic words and to understand the main topic under discussion in order to know which topic is discussed the most in user behaviour analysis or daily text [10].

Another field of study is based on the detection of spam SMS and emails. One reason for the difficulty of filtering SMS spam is that text messages usually consist of a few words of abbreviations [1]. With the detection of abbreviations, incoming SMS and emails can be separated as spam or not spam according to the words used.

3. System details

At Kariyer.net, thousands of applications are searched every day by jobseekers and thousands of resumes are scanned by employers. It is very important that the abbreviations and their meanings are understood in order to find the correct job text according to the searched word and to ensure the correct result between the job text and the CV. For example, it should be known that the system wants to be described with the abbreviation of 'Structured Query Language'. Finding abbreviations on the Kariyer.net domain is carried out in three steps that are explained in the following subsections.

3.1. Processing of data

The data used in this study consist of job texts stored in the Kariyer.net database. In order to train the Word2Vec model, the job texts published since January 2017 are based on, and are studied with, a total of 721,286 instances of job text data. The cleaning processes are described as follows:

- First of all, in order for the models to produce correct results, all job texts are cleaned from HTML codes, junk words and special characters. Using the Python Beautiful Soup library, HTML and XML residues are removed from the job texts. Special characters, such as #, *, /, are also cleared while processing, but there are also words, such as ASP.Net, C ++, OS/390, that contain special characters in the text. There are two ways to avoid losing the characters in these words. A search is made among the most frequently used skills and the most searched words in Kariyer.net, and if the word is included in this list, the special characters in the word are not deleted or changed.
- Turkish stop words, such as 'sey', 'bir' and 'de/da', are removed from the text in order to not mislead the model education. A pure text is obtained by clearing all punctuation marks, except for the points separating the sentences and special characters within the required skills and most searched words.
- The *n*-gram method is used to identify common words in the text. *N*-gram is the method used to search for data, make comparisons and learn the number of repetitions of the searched expression. For each sentence in all data, unigram, bigram and trigrams are used to extract the words and phrases that are used in the Word2Vec model. As a result of the transaction, the word list obtained from all job text data is shown in Figure 1.

```
['istatistiksel_analiz', 'inceleme', 'raporlama_becerisi', 'olan'],
['iyi_derecede', 'sözlü_yazılı', 'ingilizce', 'bilgiye_sahip'],
['sürekli_öğrenme', 'yetenek', 'isteği_olan'],
['veri_bilimci', 'çalışma_arkadaşı_arayışımız_bulunmaktadır'],
```

Figure 1. Word list created with *n*-gram

3.2. Word2Vec method

In the first step, the Word2Vec model, a word embedding algorithm, is used. Word2Vec is a model developed by Tomas Mikolov and his team in 2013, using the Gensim library. Gensim is the Python library, which is capable of document indexing with large corpus, similarity extraction and topic-based modelling, and its purpose is natural language processing [6].

Word2Vec is an unsupervised neural network model used to determine the semantic distance between words [7]. Word2vec detects similarities mathematically. Its purpose and usefulness are to group the vectors of similar words together in the vector field. Word 2vec creates vectors distributed

as numerical representations of features, such as the context of words. Given enough data, usage and contexts, Word2vec can make quite accurate predictions about the meaning of a word based on its past views [4]

Abbreviations and expressions can be found in different places within the sentence. Using word embeddings, the locations of these words are reached. For this reason, the Word2Vec model is used in this study.

The Word2Vec model is created with cleared texts and it offers proximity values for each word in the model text; an abbreviation is sought in these proximities. The words in the model are searched for and their suitability in the form of abbreviations and expansions is checked.

In the formed loop, proximity examinations are made for each of the words and phrases that are cleaned and formed with *n*-grams. The number of letters in the loop and the number of words of the expansion according to the first 10affinities are checked and the letter compatibility is checked. The situation is also checked for the reverse sample. The number of words of the expansion coming in the loop and the number of letters of the abbreviation are checked according to the first 10proximity,for example, when the closeness of 'nlp' is examined in the model.

natural_language_processing	0.6505253314971924
deep_learning	0.6494208574295044
machine_learning	0.6371470093727112
predictive_analytics	0.6174750328063965
neural_networks	0.6118414998054504
computer_vision	0.608015239238739
pattern_recognition	0.6064407229423523
information_retrieval	0.5872820615768433
neural_network	0.5870271325111389
tensorflow_keras	0.5860831141471863

Figure 2. The affinity of 'nlp' with the Word2Vec model

As can be seen in Figure 2, the abbreviation is considered correct and added to the dictionary because it is compatible with the 'nlp' letter number and 'natural language processing' word number and initials.

3.3. RegEx method

In the second step, the qualities containing parentheses '()' are determined within the job text. Both Turkish and English skills and most searched words datasets are used in the Kariyer.net database. A small sample of the data set is shown in Figure 3.

1	**NitelikTanim**
2	Veritabanı Sistemleri
3	Adabas
4	Btrieve
5	OSPF(Open Shortest Path First)
6	DataPerfect
7	DBMaker
8	FileMaker Pro
9	SNA (System Network Architecture)
10	GNU SQL

Figure 3. Example of skills and the most searched words data

While detecting parentheses, the Regular Expressions (RegEx) model is used. RegEx is a structure that allows a string of characters with the same syntax to be determined within the framework of the

specified rules. With the RegEx query shown in Figure 4, the properties containing parentheses shown in Figure 5 are determined.

```
re.finditer(r"\((.*?)\)")
```

Figure 4. RegEx query

NAT (Network Address Translation)
OSPF (Open Shortest Path First)
SNA (Systems Network Architecture)
TN320 (TelNet 3270)
VPN (Virtual Private Networking)
LAN (Local Area Network)
NAS (Network Attached Storage)
NCP (Network Control Program)
NDS (Novell Directory Services)
NFS (Network File System)
NIS (Network Information System)
NOC (Network Operations Center)
RAS (Remote Access Server)
SAN (Storage Area Network)
SMS (Systems Management Server)

Figure 5. Data obtained with the Regex query

If there is only one word in parenthesis, then that word is accepted as an abbreviation and the letter compatibility is checked with the phrases that come to the right or left of the letter. Since the expansion itself was found in parentheses, the operation was conducted in reverse order. Abbreviations and expansions that meet the requirements are added to the dictionary, for example, 'Air Pollution Control' or 'Kulak Burun Bogaz'.

Abbreviations obtained from both steps were combined and eliminated for those with the same expansions, and the number was reduced to one. As a result, a dictionary consisting of 2,955 abbreviations and extensions was obtained.

4. Conclusion

With this study, nearly 3000 abbreviations with different expansions were obtained from the job texts with the rules determined by the Word2Vec model and the Regex system. An example of the abbreviation and expansion dictionary that was created to integrate the Kariyer.net system, to calculate the eligibility of the job text and resume correctly and to reach the correct results with the searches made in the search engine is shown in Figure 6.

This project will be followed by further studies that will be aimed at finding the correct meaning of the abbreviations that have the same acronym with different expansions and have different meanings in context, for example, the abbreviation for 'Automated Teller Machine' or 'Asynchronous Transfer Mode'. The aim of the future study will be to analyse the entire text and to determine which context fits the text.

MHRS	Merkezi Hekim Randevu Sistemi
MHTC	Migrant Health Training Centers
MI	Management Information
MİB	Mücevher İhracatçıları Birliği
MIC	Medtronic Innovation Center
MIE	Microsoft Innovative Educator
MIFA	Material Information Flow Analysis
MİM	Müşteri İlgi Merkezi
MİP	Malzeme İhtiyaç Planlama
MIS	Management Information System
MİSEM	Meslek İçi Sürekli Eğitim Merkezi
MITG	Minimally Invasive Therapies Group
MİY	Müşteri Ilişkileri Yetkilisi
MİY	Müşteri İlişkileri Yöneticisi
MKE	Makine Kimya Endüstrisi

Figure 6. Sample of the dictionary

References

[1] Almeida, T.A., Hidalgo, J.M.G. & Yamakami, A. (2011). *Contributions to the study of SMS spam filtering: new collection and results* (pp. 259–262). Proceedings of the 2011 ACM symposium on document engineering.

[2] Byrd, R. (2001). *Hybrid text mining for finding abbreviations and their definitions*. Proceedings of the 2001 conference on empirical methods in natural language processing. Retrieved from https://www.aclweb.org/anthology/W01-0516

[3] Denny, J.C., Irani, P.R., Wehbe, F.H., Smithers J.D. & Spickard, A. (2003). 3rd The Knowledge Map project: development of a concept-based medical school curriculum database. *Annual Symposium proceedings/AMIA Symposium, 2003*, 195–199.Retrieved from https://pubmed.ncbi.nlm.nih.gov/14728161/

[4] Detection Abbreviations in Kariyer.net Domain. (2019). Retrieved from http://arge.kariyer.net/Makale/Kariyernet-Domainindeki-Kisaltmalarin-Bulunmasi

[5] Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J. & Johnson, S.B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association, 1*(2), 161–174. doi: 10.1136/jamia.1994.95236146

[6] Gensim 3.8.1. (2019).Retrieved from http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/doceng11.pdf [Online]. Retrieved from https://pypi.org/project/gensim/.

[7] Handler, A. (2014). *An empirical study of semantic similarity in Word Net and Word2Vec*. Retrieved fromhttps://scholarworks.uno.edu/td/1922/

[8] Korkmaz, Z. (1992). *Grammar terms dictionary*. Ankara, Turkey: Turkish Language Authority Press.

[9] Liu, K., Mitchell, K.J., Chapman, W.W.& Crowley, R.S. (2005). Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *Annual Symposium proceedings/AMIA Symposium, 2005*, 460–464. Retrieved from https://pubmed.ncbi.nlm.nih.gov/16779082/

[10] Malviya, S. & Nair, P.S. (2016). A new approach of semi-supervised clustering with abbreviation detection and domain prediction using online dictionaries. *International Journal of Engineering Science and Computing, 6*(7).

[11] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., & Chute, C.G. (2010). Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507–513.doi: 10.1136/jamia.2009.001560

[12] Wu, Y., Rosen bloom, S.T., Denny, J.C., Miller, R.A., Mani, S., Giuse, D.A. & Xu, H. (2011). Detecting Abbreviations in discharge summaries using machine learning methods. *AMIA Annual Symposium Proceedings, 2011*, 1541–1549. Retrieved from https://pubmed.ncbi.nlm.nih.gov/22195219/

[13] Zeng, Q.T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S.N. & Lazarus, R. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making,6*, 30 Retrieved from https:// bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-6-30