Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 385 – 394

Procedia Computer Science

Eva19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

A Comparison of Concept-base Model and Word Distributed Model as Word Association System

Akihiro Toyoshima[a], Noriyuki Okumura[b]

[a]*Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama Ikoma Nara, 630-0101, Japan*
[b]*Electrical and Computer Engineering, National Institute of Technology, Akashi College, 679-3 Nishioka Uozumi Akashi Hyogo, 674-8501, Japan*

Abstract

We construct Concept-base based on concept chain model and word vector spaces based on Word2Vec using EDR-electronic-dictionary and Japanese Wikipedia data. This paper describes verification experiments of these models regarding the word association system based on the association-frequency-table. In these experiments, we investigate the tendency using associative words of evaluation basis words obtained by these models. In Concept-base model, we observed a tendency that synonyms, superordinate words, and subordinate words are obtained as associative words. Furthermore we observed a tendency that words, which can be compounds or co-occurrence phrases after connecting headwords of the association-frequency-table, are used as associative words in the Word2Vec model. Moreover evaluation result showed the tendency that associative words mostly have category words in the Word2Vec model.

© 2016 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of KES International

Keywords: Concept-base; Associative words; Word2Vec; Concept-dictionary; Conversation

1. Introduction

With the development of computerized society and the technique of national language processing, a conversation between humans and computers is attracting attention as a problem. For example, various companies develop chatbot systems that converses with human through a network by the spread of Social Networking Service such as Twitter[1] and LINE[2]. These chatbot systems are conversation systems with human using national languages. For example, Softbank developed a robot called "Pepper", which communicates with human beings[3]. We predict the number of robots and systems which communicate with human will increase from now on.

* Corresponding author. Tel.: +81-743-72-5265 ; fax: +81-743-72-5269.
E-mail address: toyoshima.akihiro.su4@is.naist.jp

[1] http://twitter.com
[2] http://line.me/
[3] http://www.softbank.jp/robot/special/tech/

1877-0509 © 2016 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.080

We can do smoothly communicate with each other because we have the word associative knowledge which can associate other relation words from any words (hearinafter, referred to as "associative knowledge"). For example, when we heard "It will rain after this afternoon.", we can associate "umbrella" and "cold" based on "rain". Therefore we take the next utterance topics about "Do you have an umbrella?" and "Do you have a coat?" that related to the talking information of the partner. Computers needs this word associative knowledge such as Concept-base. We can make computers to communicate with human-beings using Concept-base.

In this paper, we constructed Concept-base and word vector spaces based on Word2Vec using EDR-electronic-dictionary[7] and Japanese Wikipedia data[4]. Moreover we verified these models that have human's word association using an association-frequency-table[11]. The association-frequency-table is a database that associative words defined as headwords. We verified these models using this database because this database made by large scale subject experiments. As a result, we observed a tendency Concept-base model contains synonymous, superordinate, and subordinate words as associative words and a Word2Vec model contains associative words which are connected any words and become compound or co-occurrence phrase. Moreover Word2Vec model has category words as associative words.

2. Related Works

Tamagawa et al.[2] constructed a large-scale general ontology based on Japanese Wikipedia information. They constructed the ontology based on the higher rank and lower rank relations between words and synonymous relations from Japanese Wikipedia data. For example, "human" and "animal" are extracted from "baby" using higher rank and lower rank relations between words. Moreover "infant" and "babe" are extracted from "baby" using synonymous relations between words. Although, it is difficult that we naturally extract human associative words using these relations. For instance, it is difficult that we extract "candy" and "toy" from "baby" using these relations.

Mikolov et al.[3,4] constructed the distribution expression of a word to study what kind of words appearance as opposed to the circumference of any words using a neural network. This method is called Word2Vec that we can calculate semantic addition and subtraction between words in this distribution expression of a word. For example, we subtract "man" from "king" and add "woman" in this distribution expression of a word. We can get "queen". This result shows Word2Vec can similarly calculate between words.

Word2Vec has some models to construct word vector spaces, Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skip-gram). CBOW is a method of sum of context circumference word weights as any words. Skip-gram is a method that estimate context circumference word appears. In this study, we verify characteristics of word vector spaces using Word2Vec and Concept-base as a human's word association.

Kasahara et al.[5] constructed Concept-base as word vector spaces. This word vector spaces use headwords of dictionary as independent base vectors. They verified Concept-base comparative usefulness evaluation with the distinction of similarly using the thesaurus[6]. Their subject is semantic similarly evaluation between words.

Our Concept-base is defined as word chain set and our goal is the realization of the associative system for natural conversation. For example, not only synonymous words that "mouth" and "nose" but "illness", "inflammation", and "medicine" also associate from "throat". Therefore in usages our Concept-base and the vector space model differ.

3. Concept-base

We explain about construction of Concept-base with electronic dictionaries. Concept-base is a knowledge base that as any headword and associative words to this headword[1]. In Concept-base, all associative words are defined as headwords. In ordinary, Concept-base is constructed with electronic dictionaries and electronic newspapers.

We extract headwords and independent words in each sentence that belongs to each headword. The headword is a dictionary headword and is defined as the concept A. Independent words are explanation sentence in the dictionary and are defined as attributes a_i of concept A. We give weights w_i to attributes a_i. Weights w_i show the evaluation of attributes a_i for the concept A. We define the concept A such as the equation (1).

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\} \tag{1}$$

[4] http://dumps.wikimedia.org/jawiki/20150402/

Download English Version:

https://daneshyari.com/en/article/4961838

Download Persian Version:

https://daneshyari.com/article/4961838

Daneshyari.com