Question Retrieval untuk Data Percakapan Pendek Menggunakan Informasi Subkata pada *Question Answering*

Helmi Satria Nugraha[1], Suyanto[2],

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]helmisatria@students.telkomuniversity.ac.id, [2]suyanto@telkomuniversity.ac.id

Abstrak

Banyaknya pertanyaan yang diajukan oleh para pengguna *(user)* terhadap suatu layanan membuat *user* lain tidak mendapat jawaban dengan cepat, hal ini dapat mengakibatkan tingkat kepuasan *user* terhadap layanan menurun. Untuk menyelesaikan permasalahan tersebut, penelitian ini membuat sebuah sistem *Question Answering* yang dapat memberikan jawaban yang relevan dengan lebih cepat, sesuai dengan pertanyaan yang diajukan. Metode yang digunakan adalah *Question Retrieval* dimana sistem akan memberikan jawaban yang paling relevan berdasarkan pertanyaan-pertanyaan tersimpan. Sulitnya memaknai dua pertanyaan dengan penggunaan kata yang berbeda namun memiliki makna sama, membuat penelitian ini menarik untuk diteliti. Selain itu dataset yang digunakan pada penelitian ini memiliki 49.66% kata pada kosakata yang baru muncul satu kali (langka), hal ini membuat besar kemungkinan adanya pertanyaan baru yang mengandung kata langka. Penelitian ini melakukan perbandingan antara penggunaan fitur satu kata penuh dengan subkata untuk mendapatkan hasil representasi kata *(word embedding)* terbaik. Hasil *word embedding* digunakan untuk mencari nilai *cosine similarity* antara pertanyaan baru *(query)* dengan pertanyaan yang tersimpan, sehingga pada akhirnya *user* akan menerima jawaban berdasarkan pertanyaan yang memiliki nilai *cosine* terbesar. Berdasarkan hasil perbandingan, disimpulkan bahwa fitur subkata tepat digunakan untuk melakukan *word embedding* pada dataset percakapan pendek yang digunakan.

Kata kunci: *question answering*, *question retrieval*, *word embedding*, subkata

Abstract

Many questions posed by users for a certain service can cause difficulties in answering them all. This results in the decrease of level in user satisfaction with the service. To solve this problem, a study called Question Answering system is created. It can provide relevant answers according to the questions asked. The method used by this system is the Question Retrieval, the system will provide answers based on the most relevant stored questions. The difficulty of interpreting two questions that is essentially the same but with different words makes this research an interesting to study. In addition, the dataset used in this study has 49.66% of the words in the new vocabulary appearing once in the (rare) dataset; this makes it possible for new questions to contain rare words. This study makes a comparison between the use of a complete one-word feature with sub-words to get the best word embedding in the question. The word embedding result is used to find the similar cosine value between new query questions and stored questions. In the end, the user will receive an answer based on a question that has the greatest cosine value. Based on the results of the comparison, it is concluded that the best sub-word feature is used for word embedding on the short conversation dataset used.

Keywords: question answering, question retrieval, word embedding, sub-word

1. Pendahuluan

1.1 Latar Belakang

Perkembangan cepat pada pengguna media sosial di Indonesia seperti Line, Facebook, Instagram dan Whatsapp memfasilitasi pengguna sehingga dapat saling berbagi dan berkomunikasi. Perkembangan ini berdampak sehingga saat ini berkomunikasi melalui media sosial telah menjadi tren dan seakan sudah menjadi kebutuhan utama bagi setiap orang [15]. Hal ini membuat orang-orang mulai meninggalkan layanan pesan singkat dan telepon sebagai alat komunikasi utama. Terjadinya hal tersebut, membuat penyedia layanan seperti institusi atau perusahaan yang menyediakan layanan tanya jawab, mengalihkan orang-orang yang ingin bertanya dari menggunakan pesan singkat atau telepon, menjadi menggunakan media sosial yang menyediakan fitur *chat*. Banyaknya