ABSTRAK

SISTEM IDENTIFIKASI HUBUNGAN SEBAB AKIBAT PADA PEMROSESAN TEKS BAHASA INDONESIA DENGAN PENDEKATAN TEMPLATE SEMANTIK

Oleh
Susetyo Bagas Bhaskoro
NIM: 33212024
(Program Studi Doktor Teknik Elektro dan Informatika)

Tujuan pengolahan bahasa alami adalah melakukan transformasi informasi dari bentuk kompleks menjadi lebih ringkas, mudah dipahami dan dapat pula dijadikan referensi analisis pada beberapa domain tertentu salah satunya adalah domain kesehatan. Informasi yang biasanya dimanfaatkan pada domain kesehatan adalah informasi tentang kausalitas atau sebab – akibat yaitu pengetahuan yang dibangun berdasarkan fakta-fakta khusus yang memiliki makna <sebab> dan disimpulkan menjadi fakta-fakta umum yang memiliki makna <akibat>, dan antara kedua makna tersebut memiliki hubungan yang saling menjelaskan. Jika diformulasikan hubungan sebab – akibat menjadi sebuah fungsi $f(x,y)$, untuk x adalah makna <sebab>, dan y adalah makna <akibat>, maka terlihat seperti berikut $f(x,y) = x \rightarrow y$ atau $f(x,y) = x\ cause\ y$ atau $f(x,y) = if\ x\ then\ y$. Permasalahannya adalah tidak semua makna <sebab> dan makna <akibat> dijelaskan pada kalimat sederhana atau *single sentences*, tetapi terdapat juga kemungkinan makna <sebab> dan makna <akibat> dijelaskan pada kalimat kompleks atau *multiple sentences* yang posisinya dapat berada di awal kalimat atau di akhir dari sebuah kalimat. Selain itu, makna <sebab> dan makna <akibat> juga tidak selalu diabstraksikan secara eksplisit melainkan implisit.

Penelitian ini memiliki tujuan menghasilkan metode template semantik, yaitu metode untuk melakukan identifikasi makna <sebab> dan makna <akibat> pada kalimat sederhana dan juga pada kalimat kompleks. Metode template semantik yang dikembangkan saat ini lebih difokuskan pada penggunaan template untuk melakukan identifikasi hubungan sebab – akibat yang tidak tergantung pada pembahasan topik yang khusus, melainkan bersifat terbuka untuk beragam pembahasan teks. Usulan yang terdapat pada template semantik yaitu mempertimbangkan <seleksi fitur>, <posisi kata>, <makna kata>, <frekuensi kemunculan kata>, <dinamis root> dan penggunaan <aturan template semantik> itu sendiri. *Dataset* yang digunakan ketika melakukan pelatihan dan pengujian berasal dari kumpulan contoh kalimat hubungan sebab – akibat pada publikasi penelitian lainnya dan berasal dari artikel media online sebanyak 1,017 kalimat. Pengujian kinerja sistem identifikasi dibedakan menjadi dua kategori evaluasi yaitu ekstrinsik dan pengujian intrinsik. Hasil evaluasi kinerja dari metode template semantik pada pengujian karakteristik teks *single sentences* yaitu 0,874; 0,747 dan 0,803, sedangkan pada pengujian karakteristik teks *multiple sentences* yaitu 0,877; 0,815 dan 0,845 untuk *recall*, *precision* dan *f-measure* sedangkan rata-rata error sebesar 0,280. Hasil evaluasi untuk kualitas informasi *accuracy* sebesar 0,719, sedangkan *precision* sebesar 0,781 dan hasil untuk *completeness* adalah 0,953.

Metode template semantik yang dihasilkan telah berhasil diimplementasikan pada domain kesehatan yaitu sistem surveilans kesehatan masyarakat. Pada saat implementasi terdapat beberapa usulan yang dihasilkan yaitu transformasi bahasa alami ke anotasi elemen medis (LPpAJSFPePnGOD) untuk identifikasi pola hubungan sebab – akibat, pola anotasi paragraf untuk klasifikasi posisi kalimat pada paragraf artikel medis dan membangun pola keterhubungan *semantic relationship* untuk anotasi semantik.

Kata kunci: template semantik, *single sentences*, *multiple sentences*, anotasi elemen medis, anotasi paragraf, anotasi semantik, sebab - akibat.

ABSTRACT

CAUSAL RELATIONSHIP IDENTIFICATION SYSTEM ON INDONESIAN TEXT PROCESSING USING SEMANTIC TEMPLATE APPROACH

By
Susetyo Bagas Bhaskoro
NIM: 33212024
(Doctoral Program in Electrical Engineering and Informatics)

The objective of natural language processing is to transform complex information into brief information, comprehensible and able to be used as an analysis reference for certain domain, i.e. health domain. Information that frequently utilized on health domain is an information about causality or cause-effect which is a knowledge that is built on particular facts that have `<cause>` *meaning and is concluded as general facts that have* `<effect>` *meaning, and between those two meanings have a relationship that explains each other. If causal relationship is being formulated, it turns into function* $f(x,y)$*, where* x *is* `<cause>` *meaning, and* y *is* `<effect>` *meaning, so it will be* $f(x,y) = x \rightarrow y$ *or* $f(x,y) = x$ *cause* y *or* $f(x,y) =$ *if* x *then* y*. The issue is not all of* `<cause>` *and* `<effect>` *meanings are explained in simple sentence ot single sentences, but there is also possibility for* `<cause>` *and* `<effect>` *meanings to be explained in complex sentence or multiple sentences that the position itself is at the beginning of sentence or at the end of sentence. Furthermore,* `<cause>` *and* `<effect>` *meanings are not always be abstracted explicitly but also implicitly.*

Research's aim is to generate semantic template method, which is method to identify `<cause>` *and* `<effect>` *meanings on simple sentence and complex sentence. Current developed of semantic template method is focusing on template using for identify causal relationship that has no dependency on certain topic, but variety of textual topic. Proposal contained in semantic template is considering* `<feature selection>`, `<word position>`, `<word meaning>`, `<frequency of word occurence>`, `<root dynamic>` *and use of* `<template rule>`*. This research using dataset from collection of causal relationship sentences found in some publications and online articles as many as* `1.017` *sentences for training and testing. Identification system performance testing is divided into two evaluation cateogries, extrinsic and intrinsic testing. Performance evaluation results of semantic template method on single sentences are* `0,874; 0,747` *and* `0,803`*, meanwhile results on multiple sentences are* `0,877; 0,815` *and* `0,845` *for each recall, precision and f-measure, meanwhile error rate is* `0.280`*. Evaluation result for information accuracy quality is* `0.719`*, meanwhile for precision is* `0.781` *and for completeness is* `0.953`*.*

Generated semantic template method that has been successfully implemented on health domain is public health surveillance system. During implementation, there are several proposals have been generated, i.e. natural language transformation into medical element

annotation (LPpAJSFPePnGOD) to identify causal relationship pattern, paragraph annotation pattern to classify sentence position on medical article paragraph and build a semantic relationship pattern for semantic annotation.

Keywords: semantic template, single sentences, multiple sentences, medical element annotation, paragraph annotation, semantic annotation, causality.