Chapter 3: A SURVEY OF TEXT SUMMARIZATION TECHNIQUES

1. How do Extractive Summarizers Work?

Extractive Summarizers adalah sistem yang digunakan untuk menghasilkan ringkasan teks yang singkat dan padat dengan menyampaikan informasi utama dari teks input. Sistem ini biasanya menggunakan teks input berupa satu dokumen atau kumpulan dokumen terkait, dan mengambil beberapa kalimat penting dari teks input tersebut untuk dijadikan ringkasan.

Terdapat tiga proses utama, yaitu menciptakan representasi intermediat dari teks input, memberikan skor pada setiap kalimat berdasarkan representasi tersebut, dan memilih beberapa kalimat terbaik sebagai ringkasan.

Representasi intermediat merupakan penyederhanaan teks input menjadi representasi yang lebih sederhana dan dapat diinterpretasikan sebagai topik yang dibahas dalam teks. Dan setiap kalimat akan diberi skor berdasarkan pentingnya dalam teks input. Sistem extractive memilih beberapa kalimat terbaik untuk dijadikan ringkasan

Terdapat beberapa yang mempengaruhi penilaian pentingnya suatu kalimat dalam teks, seperti konteks pengguna atau query, lingkungan tempat dokumen input berada, dan genre dokumen.

2. Topic Representation Approaches

Beberapa metode yang paling umum digunakan

2.1. Topic Words

Metode pertama adalah penggunaan kata-kata topik untuk mengidentifikasi kata-kata deskriptif dalam sebuah dokumen yang akan diringkas. Kata-kata deskriptif dalam pendekatannya mengecualikan kata-kata yang paling sering muncul "tanda topik" dalam dokumen, seperti determiner, preposisi, atau kata-kata khusus domain, serta kata-kata yang hanya muncul beberapa kali.

Pentingnya suatu kalimat dihitung berdasarkan jumlah tanda topik yang terkandung di dalamnya atau sebagai proporsi tanda topik dalam kalimat. Kedua fungsi penilaian kalimat ini didasarkan pada representasi topik yang sama, tetapi nilai yang mereka berikan pada kalimat dapat sangat berbeda.

2.2 Frequency-driven Approaches

Merupakan pendekatan kata topik dan pendekatan probabilitas kata. Pendekatan ini dapat digunakan untuk memilih kalimat atau kata-kata yang penting dalam dokumen dan membuat ringkasan yang efektif.

Pendekatan kata topik menggunakan representasi biner (1 atau 0) untuk menentukan kata mana yang terkait dengan topik, sementara pendekatan kata menggunakan bobot untuk menentukan kepentingan kata dalam dokumen.

SumBasic, suatu sistem yang menggunakan probabilitas kata untuk menentukan kepentingan setiap kalimat dalam dokumen. Bobot yang diberikan pada setiap kalimat dihitung berdasarkan probabilitas rata-rata kata dalam kalimat tersebut, dan memilih kalimat skor terbaik.

Pendekatan Centroid yang merupakan ringkasan dokumen menggunakan representasi topik dan menetapkan ambang batas untuk menentukan kata yang penting dalam dokumen. Pendekatan lain yang disajikan adalah pendekatan rantai leksikal, yang menentukan kepentingan kalimat dengan memanfaatkan hubungan semantik antara kata.

2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) adalah teknik yang digunakan untuk dalam scoring kalimat dan memilih konten penting. Dengan menggunakan teknik ini, kita dapat merangkum berita tunggal atau multi-dokumen tanpa menggunakan resource seperti WordNet. Representasi topik digambarkan dengan matriks n x m, dimana setiap baris merupakan kata dari input (n kata) dan setiap kolom adalah kalimat dalam input (m kalimat).

Teknik singular value decomposition (SVD) dari aljabar linear diterapkan pada matriks tersebut untuk merepresentasikannya sebagai hasil perkalian tiga matriks: A = UΣVT.

Dalam proses LSA, matriks U merepresentasikan topik yang berisi kombinasi kata-kata dari input, matriks Σ adalah matriks diagonal, dan matriks VT merepresentasikan kalimat dalam bentuk topik dari U. Proses reduksi dimensionalitas dilakukan dengan menghapus topik dengan bobot rendah.

2.4. Bayesian Topic Models

Bayesian Topic Models adalah sebuah metode yang digunakan untuk merepresentasikan topik dalam ringkasan teks. Metode ini menggunakan distribusi probabilitas kata yang muncul dalam input untuk bahasa Inggris umum (G), untuk seluruh cluster yang akan diringkas (C), dan untuk setiap dokumen individu dalam cluster tersebut (Di).

BTM menggunakan prosedur skor kalimat yang berbeda yaitu Kullback-Lieber (KL) divergence yang dianggap sebagai cara baik untuk menilai dan memilih kalimat dalam rangkuman karena dapat mengukur seberapa mirip antara rangkuman dan input. Prosedur pemilihan kalimat dilakukan secara iteratif menggunakan pendekatan greedy, dimana setiap iterasi, kalimat terbaik yang akan dipilih adalah yang memiliki nilai KL divergence terkecil antara probabilitas kata dalam cluster dan rangkuman yang telah terbentuk.

2.5 Sentence Clustering and Domain-dependent Topics

Merupakan salah satu pendekatan yang digunakan adalah pengelompokan kalimat, di mana kalimat-kalimat serupa dikelompokkan menjadi topik-topik yang penting dalam rangkuman. Namun, pendekatan ini memiliki kelemahan karena setiap kalimat harus ditempatkan dalam satu kelompok saja.

Untuk rangkuman yang lebih spesifik terhadap domain tertentu, digunakan pendekatan model Hidden Markov yang menangkap "story flow" dalam domain dan dapat dilatih untuk memilih dan menyusun konten dalam rangkuman. Pendekatan yang lebih sederhana dapat diterapkan ketika sampel tersedia dan terstruktur dengan baik, seperti di Wikipedia, di mana pengelompokan judul yang

mirip digunakan untuk mengidentifikasi topik-topik yang dibahas dalam setiap jenis artikel.

Dalam membuat rangkuman, penting untuk memilih pendekatan yang sesuai dengan kebutuhan dan karakteristik dokumen atau berita yang akan diringkas.

3. Influence of Context

pengaruh konteks dalam proses summarization atau penyederhanaan teks. Summarizer dapat menggunakan bahan tambahan untuk menentukan topik yang paling penting dalam dokumen yang akan disederhanakan.

3.1 Web Summarization

Dalam web page summarization, konteks yang dapat digunakan adalah teks pada halaman web yang memiliki link ke halaman yang ingin disederhanakan. Sedangkan dalam blog summarization, konteks yang dapat digunakan adalah diskusi yang terjadi setelah blog post. Pada summarization of scholarly papers, konteks yang dapat digunakan adalah paper lain yang mengutip paper yang ingin disederhanakan, khususnya kalimat kutipan.

Terdapat pula pendekatan query-focused summarization yang mempertimbangkan minat pengguna sebagai konteks tambahan. Dalam web summarization, konteks dapat digunakan untuk menentukan kalimat-kalimat penting yang akan disederhanakan. Sedangkan dalam summarization of blog posts, kalimat-kalimat penting ditentukan berdasarkan frekuensi kata dalam komentar.

3.2. Summarization of Scientific Articles

Salah satu metode SSA adalah impact summarization, yaitu mengekstrak kalimat-kalimat dari suatu artikel dapat diukur dengan membandingkan kata kunci yang digunakan dalam artikel tersebut dengan kata kunci pada artikel lain yang mengutip artikel tersebut. Nilai penting suatu kalimat dinilai berdasarkan kemiripannya dengan model bahasa yang dibangun dari artikel yang mengutip artikel tersebut. Skor akhir suatu kalimat adalah gabungan antara pengaruh penting dari model bahasa dan kepentingan dari kata-kata dalam. Skor kesimpulan sebuah

kalimat diukur oleh kesamaan antara kalimat tersebut dan model bahasa menggunakan divergensi KL. Skor akhir dari sebuah kalimat merupakan kombinasi linear dari pentingnya dampak dari divergensi KL dan pentingnya intrinsik dari probabilitas kata dalam artikel yang diinput.

3.3. Query-focused Summarization

Query-focused Summarization merupakan teknik untuk menentukan pentingnya kalimat dalam konteks input dan relevan dengan pertanyaan pengguna. Terdapat dua kelas pendekatan dalam teknik ini, yaitu adaptasi teknik summarization generik dan pengembangan metode baru untuk kalimat yang relevan. Pendekatan pertama menggunakan teknik topic signature words yang diperluas untuk dengan mempertimbangkan probabilitas kata dalam ringkasan. Pendekatan kedua menggunakan model bahasa dari teks biografi dan non-biografi untuk mengidentifikasi kalimat biografi dalam dokumen input. Aplikasi dari teknik ini yang berguna adalah memproduksi snippet untuk mesin pencari.

3.4. Email Summarization

Email summarization adalah ringkasan dari email dengan mempertimbangkan karakteristik unik dari genre bahasa email dengan memperhatikan sifat interaktif dialog. Metode menggunakan teknik ekstraktif dengan mengambil satu kalimat dari setiap level diskusi. Salah satu teknik terbaru menggunakan analisis berbasis grafik dari kutipan dalam email untuk menghitung grafik yang mewakili setiap email yang disebutkan secara langsung oleh email lainnya dalam bentuk kalimat.

4. Indicator Representations and Machine Learning for Summarization

Ada dua pendekatan dalam mengembangkan sistem ringkasan teks. Pendekatan pertama adalah menggunakan representasi indikator. Pendekatan ini menciptakan representasi teks yang dapat digunakan untuk mengurutkan kalimat-kalimat berdasarkan tingkat penting. Salah satu contoh adalah metode Graph.

Pendekatan kedua adalah menggunakan machine learning untuk mengekstraksi kalimat-kalimat yang penting. Pendekatan ini menggunakan berbagai macam indikator dan menggunakan machine learning untuk menyelesaikan tugas tersebut.

4.1. Graph Methods for Sentence Importance

Merupakan representasi grafik dalam membuat rangkuman teks yang telah diinputkan. Setiap kalimat dalam teks direpresentasikan sebagai sebuah vertex atau simpul, dan setiap hubungan antara kalimat diberi bobot sesuai dengan kemiripan antara dua kalimat tersebut. Kemiripan antara kalimat biasanya dihitung dengan menggunakan kosinus kemiripan dengan bobot TF*IDF untuk kata.

Bobot pada setiap hubungan antar kalimat tersebut kemudian dinormalisasi menjadi sebuah distribusi probabilitas, sehingga grafik tersebut menjadi sebuah rantai Markov. Dengan menggunakan algoritma proses stokastik, dapat dihitung probabilitas setiap simpul pada grafik, sehingga simpul-simpul yang memiliki probabilitas lebih tinggi dianggap lebih penting. pendekatan ini dapat ditingkatkan dengan mempertimbangkan informasi semantik dan sintaktik dalam membangun grafik teks.

Selain itu, bobot dalam satu dokumen dapat dibedakan dengan bobot pada beberapa dokumen. Hal ini dapat membantu membedakan topik dalam satu dokumen dan topik yang sama di antara beberapa dokumen. Pendekatan ini juga dapat digabungkan dengan teknik pembelajaran mesin untuk mengkombinasikan beberapa kalimat penting.

4.2. Machine Learning for Summarization

Edmundson di awal telah mengusulkan bahwa berbagai indikator penting dapat digabungkan dalam membuat ringkasan, dan korpus input serta ringkasan yang ditulis oleh manusia dapat digunakan untuk menentukan bobot setiap indikator. Dalam metode yang disupervisi, pengambilan keputusan mengenai kalimat yang penting direpresentasikan sebagai masalah klasifikasi biner, dengan setiap kalimat direpresentasikan sebagai daftar potensi indikator penting. Klasifikasi dilatih menggunakan korpus yang diannotasi oleh manusia untuk membedakan kalimat yang harus dimasukkan ke dalam ringkasan, dengan bobot

setiap indikator yang berbeda menentukan probabilitas kalimat dimasukkan ke dalam ringkasan. Beberapa fitur yang umum digunakan dalam indikator penting antara lain posisi kalimat dalam dokumen, panjang kalimat, kesamaan kalimat dengan judul atau subjudul dokumen, serta kehadiran entitas bernama atau frasa kunci dari daftar tertentu.

Meskipun metode machine learning telah berhasil dalam membuat ringkasan dokumen dalam berbagai domain, ada beberapa masalah yang harus diatasi, salah satunya adalah kurangnya data yang diannotasi oleh manusia yang dapat digunakan untuk melatih klasifikasi. Sebagai alternatif, beberapa peneliti telah menggunakan abstrak yang ditulis oleh manusia sebagai ganti data yang diannotasi. Selain itu, terdapat juga pendekatan semi-supervised yang dapat digunakan untuk melatih klasifikasi. Terdapat juga beberapa modifikasi dari pendekatan machine learning standar yang cocok untuk pembuatan ringkasan dokumen, seperti memformulasikan pembuatan ringkasan sebagai masalah klasifikasi biner dan memilih kalimat dengan skor tertinggi sebagai bagian dari ringkasan yang terbaik.

5. Selecting Summary Sentences

Banyak pendekatan untuk ringkasan memilih konten secara berurutan dari kalimat ke kalimat: kalimat yang paling informatif pertama kali dimasukkan, dan jika memungkinkan, kalimat yang paling informatif berikutnya dimasukkan dalam ringkasan dan seterusnya. Beberapa proses pemeriksaan kesamaan antara kalimat yang dipilih biasanya juga digunakan untuk menghindari penambahan kalimat yang berulang.

5.1. Greedy Approaches: Maximal Marginal Relevance

pendekatan ringkasan Maximal Marginal Relevance (MMR) yang menggunakan metode greedy dalam pemilihan kalimat demi kalimat. Dalam setiap tahap pemilihan, algoritma greedy dipaksa untuk memilih kalimat yang paling relevan dengan pertanyaan pengguna dan minim redundan dengan kalimat yang sudah termasuk dalam ringkasan. MMR mengukur relevansi dan kebaruan secara terpisah dan kemudian menggunakan kombinasi linear dari kedua nilai tersebut untuk menghasilkan skor pentingnya sebuah kalimat dalam proses pemilihan.

Pendekatan MMR awalnya diusulkan untuk ringkasan berfokus pada pertanyaan dalam konteks temuan informasi, tetapi dapat dengan mudah disesuaikan untuk ringkasan umum. Meskipun begitu, pendekatan greedy ini mungkin tidak efektif untuk pemilihan konten yang optimal untuk seluruh ringkasan. Oleh karena itu, beberapa kasus lebih disukai menggunakan beberapa kalimat yang lebih pendek namun secara kolektif tidak mengekspresikan informasi yang tidak perlu.

5.2. Global Summary Selection

Teknik optimisasi global dapat digunakan untuk menyelesaikan tugas ringkasan yang baru diformulasikan, yaitu memilih ringkasan terbaik secara keseluruhan. Algoritma pemrograman dinamis dapat digunakan untuk menemukan solusi pendekatan. Meskipun demikian, penentuan solusi yang tepat memerlukan waktu yang lama menggunakan teknik pencarian. Dalam teknik optimisasi global, informativitas masih didefinisikan dan diukur menggunakan fitur yang sudah dikenal di literatur seleksi kalimat. Berdasarkan evaluasi yang dilakukan pada data berita dan rapat, teknik optimisasi global telah terbukti lebih unggul daripada algoritma pemilihan kalimat berdasarkan urutan. Pada penelitian terperinci tentang algoritma inferensi global, terbukti bahwa solusi yang tepat dapat ditemukan menggunakan pemrograman linear bulat. Namun, solusi pendekatan yang lebih cepat juga dapat digunakan dengan skala waktu yang lebih linear dan praktis.

6 Kesimpulan