BAB 2. TINJAUAN PUSTAKA

2.1. Penelitian yang Mendahului

Penulis mencermati berbagai penelitian terkait dengan analisis similaritas, seperti diurai pada beberapa paragraf berikut.

Bao, et al. (2007) membandingkan berbagai algoritma analisis similaritas. Dalam penelitian tersebut, dicermati sistem Ferret yang menggunakan similaritas leksikal (kata per kata), kemudian dicermati pula metode yang menggunakan similaritas semantik (berdasarkan makna kata dalam kalimat). Teks yang diteliti berasal dari kalimat-kalimat bahasa Inggris yang terdapat dalam majalah *Financial Times*. Hasil penelitian tersebut menunjukkan bahwa similaritas semantik menggunakan sinonim lebih penting dibanding frase atau deretan kata ketika mencari teks yang mirip. Frase kata benda memberi kontribusi penting dalam identifikasi similaritas teks, namun kontribusinya tidak lebih besar dibanding sinonim.

Winarsono, et al. (2009) meneliti penerapan metode *syntactic-semantic similarity* (*SynSemSim*) untuk mencermati kemiripan kalimat singkat. Metode tersebut mencermati struktur kalimat (sintaksis) dan makna kata dalam kalimat (semantik). Para peneliti ini menyimpulkan bahwa metode *SynSemSim* dapat dengan baik digunakan pada struktur kalimat yang mirip, namun kurang baik digunakan pada struktur kalimat majemuk atau kalimat yang mengandung banyak *stop word* (kata tambahan seperti “it is”, “yet”). Sayangnya, para peneliti dari Indonesia ini menggunakan *WordNet*, yaitu jejaring kata bahasa Inggris, dalam penelitiannya sehingga belum dapat digunakan pada kalimat berbahasa Indonesia.

Sun, et al. (2010) melakukan pengamatan terkait similaritas teks pada kumpulan artikel biomedis. Para peneliti memeriksa lebih dari 70 ribu dokumen. Tiap dokumen dicermati kemudian dibuat himpunan data untuk *full text*, subjudul, dan paragraf. Tiap himpunan diperiksa dan dihitung similaritasnya. Para peneliti ini mendapat kesimpulan bahwa similaritas abstrak yang tinggi mencerminkan similaritas *full text* yang tinggi. Similaritas abstrak dan similaritas *full text* mempunyai korelasi moderat. Di antara subbab dalam sebuah tulisan, subbab “Metode Penelitian” mempunyai tingkat pengulangan yang paling tinggi. Namun, dalam pemeriksaan manual terhadap artikel dan duplikatnya,

subbab “Hasil Penelitian” merupakan bagian yang sering berulang. Pengulangan subbab “Pendahuluan” dan “Metodologi” lebih sering dilakukan oleh penulis yang sama. Tingkat similaritas lebih tinggi didapat pada perbandingan antara dua paper yang di-*review*, dan similaritas jauh lebih rendah terdapat pada perbandingan antara satu paper yang di-*review* dan *paper* yang tidak di-*review*. Para peneliti ini menyimpulkan bahwa penentuan similaritas abstrak cukup efektif untuk mencari duplikasi sitasi, sedangkan analisis *full text* diperlukan untuk menemukan semua kemungkinan duplikasi sitasi.

Boyack, et al. (2011) meneliti penerapan algoritma similaritas pada proses pengelompokan dokumen. Sembilan metode diteliti untuk melihat keakuratannya dalam mengelompokkan dua juta artikel biomedis. Pengelompokan artikel bermanfaat antara lain untuk manajemen koleksi, mempermudah penelusuran berkas, dan menganalisis data. Para peneliti ini mencermati artikel pada *MEDLINE* yang di-*submit* pada kurun 2004 – 2008. Boyack dkk. menggunakan metode statistik dan algoritma semantik dalam penelitiannya. Contoh metode yang digunakan adalah frekuensi kemunculan kata (statistik) dan LSA (*latent semantic analysis*). Sumber data yang digunakan adalah katagori subjek, kata-kata pada judul, dan abstrak. Disimpulkan bahwa metode *related article* yang ada pada *PubMed* menghasilkan pengelompokan (kluster) yang paling terkonsentrasi di antara kesembilan metode yang diamati.

Thamrin dan Wantoro (2012) meneliti penerapan jarak Levenshtein sebagai landasan dalam menilai kemiripan jawaban siswa dengan kunci jawaban. Tingkat kemiripan dihitung berbanding terbalik (resiprokal) terhadap jarak Levenshtein. Tingkat kemiripan hasil perhitungan kemudian dibandingkan dengan cara guru sekolah dasar dan menengah menilai jawaban siswa. Kedua peneliti memodifikasi perangkat lunak Moodle dan membuat tipe soal baru. Dengan tipe soal baru tersebut, jawaban soal pendek dapat diberi skor secara fleksibel secara otomatis oleh komputer. Terdapat kesamaan dalam pola pemberian skor oleh guru maupun oleh komputer. Namun, kecenderungan penilaian oleh guru dan komputer akan mempunyai perbedaan signifikan jika jawaban yang diberikan siswa membentuk kata yang dikenal dalam kamus. Kedua peneliti menyarankan penggunaan algoritma similaritas semantik untuk meningkatkan akurasi penentuan skor secara otomatis.

2.2. Peta Jalan Penelitian

Gambar 1 pada halaman 7 memperlihatkan peta jalan penelitian yang menggambarkan penelitian terdahulu yang telah dilakukan baik oleh pengusul maupun oleh peneliti lain. Penelitian terdahulu dapat dikategorikan dalam empat objek penelitian, yaitu:

1. Pengembangan algoritma umum,
2. Pengembangan algoritma untuk penerapan spesifik,
3. Kajian penerapan algoritma, dan
4. Kajian penerapan pada bahasa Indonesia.

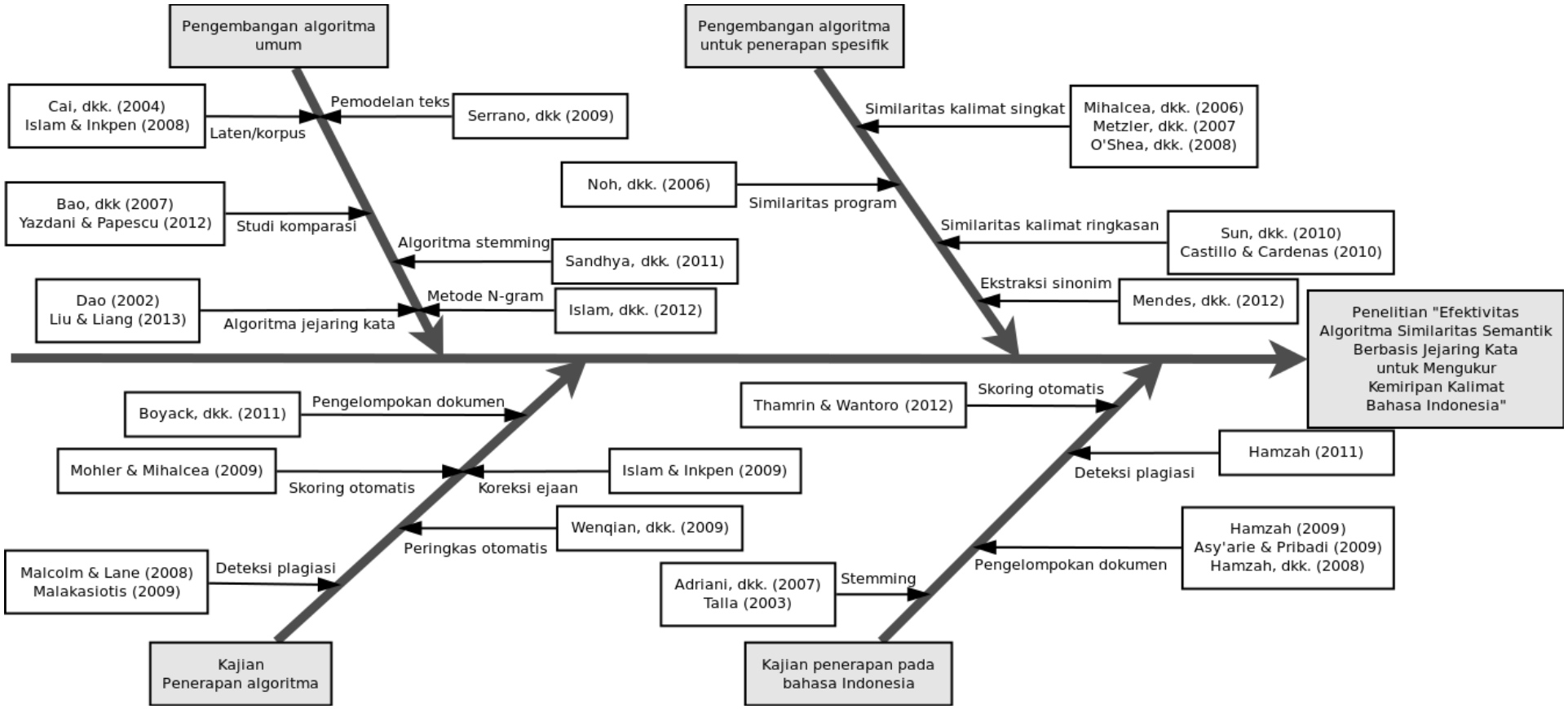
Pengembangan algoritma umum dilakukan dalam berbagai bentuk. Cai dkk. (2004) dan Islam & Inkpen (2008) meneliti algoritma similaritas berdasarkan data dalam korpus. Sedangkan Dao (2002) dan Liu & Liang (2013) mencermati algoritma berbasis jejaring kata. Serrano dkk. (2009) membuat model teks sedangkan Sandhya dkk. (2011) meneliti tentang algoritma *stemming* untuk mendapatkan makna kata secara lebih akurat. Selain itu, Islam dkk. (2012) mengembangkan metode N-Gram termasuk yang berasal dari mesin pencari Google.

Berbagai penelitian telah dilakukan untuk mengembangkan algoritma untuk penerapan spesifik. Pengembangan metode untuk mengukur kemiripan kalimat pendek dilakukan oleh Metzler dkk. (2007), O'shea dkk. (2008) dan Mihalcea dkk. (2006). Pengembangan metode untuk mengukur kemiripan kode program komputer dilakukan oleh Noh dkk. (2006). Kemiripan kalimat ringkasan diteliti oleh Sun dkk. (2010) dan Castillo & Cardenas (2010). Sedangkan upaya menemukan sinonim secara otomatis diteliti oleh Mendes dkk. (2012).

Kajian penerapan algoritma untuk kebutuhan nyata dilakukan oleh banyak orang. Malcolm & Lane (2008) dan Malakasiotis (2009) mencoba mengerapkan algoritma similaritas untuk mendeteksi plagiasi. Wenqian dkk. (2009) mencoba membuat mesin peringkas otomatis. Islam & Inkpen (2009) meneliti lebih lanjut mesin pengkoreksi otomatis sedangkan Mohler & Mihalcea (2009) mencoba menerapkan algoritma

similaritas untuk memberi skor otomatis pada sistem evaluasi belajar. Boyack dkk. (2011) telah pula berupaya menerapkan pada proses pengelompokan dokumen atau artikel.

Kebanyakan penelitian dilakukan terhadap dokumen dan teks berbahasa Inggris. Kajian penerapan algoritma similaritas pada bahasa Indonesia belum banyak dilakukan. Talla (2003) dan Adriani dkk. (2007) telah berupaya mengembangkan algoritma *stemming* untuk memisahkan kata dasar dari imbuhannya. Sementara itu, Hamzah dkk. (2008), Asy'arie & Pribadi (2009) dan Hamzah (2009) telah berupaya menerapkan algoritma untuk pengelompokan dokumen berbahasa Indonesia. Penerapan untuk deteksi plagiasi telah pula dicoba oleh Hamzah (2011). Belum lama ini, Thamrin & Wantoro (2012) berupaya menerapkan pada proses skoring otomatis. Ketiadaan jejaring kata menjadi salah satu kendala dalam upaya menerapkan pengukuran similaritas untuk mengukur kemiripan teks bahasa Indonesia. Oleh karena itu dalam penelitian ini akan diupayakan konstruksi jejaring kata sekaligus dilakukan pengujian *Efektivitas Algoritma Similaritas Semantik Berbasis Jejaring Kata dalam Mengukur Kemiripan Kalimat Bahasa Indonesia*.



Gambar 1. Peta jalan penelitian yang mengawali penelitian yang sedang diusulkan