

Pengembangan Teks Pertanyaan Menggunakan Teknik Persamaan Semantik Dalam Capaian Maklumat Al-Quran

Mohamad Hafizi Bin Jainal

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

hafizi_86@hotmail.com

ABSTRAK

Ayat suci Al-Quran adalah sumber bacaan teks umat Islam. Al-Quran dianggap sebagai teks standard yang sangat baik dan penting untuk membangunkan, membuat model dan menilai komponen Pemprosesan Bahasa Asli Arab (NLP). Ketersediaan format Al-Quran di dalam bentuk digital kini memudahkan ramai penyelidik untuk membuat kerja-kerja analisis pada teks Al-Quran menjadi senang dan cepat, terutamanya kepada mereka yang tidak biasa dengan perkataan bahasa Arab. Kerja-kerja analisis seperti mencari makna perkataan yang sesuai adalah satu cara untuk membantu para pembaca mudah memahami ayat Al-Quran dalam konteks perkataan yang dicari. Untuk mendapatkan makna yang sesuai atau hampir sama dengan teks pertanyaan dimasukkan ada beberapa teknik dan pendekatan yang ada dalam bidang Capaian Maklumat ini. Penyahtaksaan Makna Perkataan (PMP) dikenalpasti sebagai masalah terbuka yang dikaitkan dengan pemprosesan bahasa tabii dalam pengkomputeran linguistik. Masalah penyahtaksaan makna perkataan boleh diselesaikan dengan membuat satu pilihan terbaik daripada beberapa makna perkataan, yang jika tidak akan menyebabkan ketaksaan dan kekeliruan. Objektif kajian ini adalah untuk mengembangkan teks pertanyaan asal menggunakan teknik semantik dan seterusnya dapat menyediakan satu sistem Capaian Maklumat Al-Quran yang boleh menambahbaikkan teks pertanyaan menjadi lebih tepat berdasarkan konteks. Beberapa teknik pengukuran persamaan semantik digunakan iaitu Wu Palmer (WUP), Lin (LIN), dan Jiang-Conrath (JCN) untuk mendapatkan hasil skor persamaan semantik seterusnya pengembangan teks pertanyaan asal boleh dilakukan. Disamping itu, kajian ini juga akan menggunakan pangkalan data dari WordNet bagi mendapatkan maksud yang sama bagi perkataan di dalam teks pertanyaan. Pengubahsuaian teks pertanyaan dilakukan dengan menggunakan kaedah persamaan semantik antara teks pertanyaan asal dengan senarai maksud ayat dari WordNet. Set data yang diguna dalam kajian ini adalah terjemahan Al-Quran dalam bahasa Inggeris. Keputusan eksperimen menunjukkan sedikit peningkatan nilai panggilan balik, ketepatan dan F-measure dengan pelaksanaan teknik persamaan semantik yang dicadangkan dalam pengembangan teks pertanyaan bagi capaian maklumat Al-Quran.

1. PENGENALAN

Ayat suci Al-Quran adalah sumber bacaan teks umat Islam. Al-Quran dianggap sebagai teks standard yang sangat baik dan penting untuk membangun, membuat model dan menilai komponen Pemprosesan Bahasa Asli Arab (NLP).

Aplikasi Al-Quran yang mengandungi surah lengkap dalam bentuk digital boleh didapati di mana-mana sumber di Internet contoh seperti di laman sesawang dari Islamicity.com dan Tafsir.com, dan juga terdapat lebih 100 laman sesawang yang memberi akses kepada Al-Quran digital.

Aplikasi Al-Quran digital yang ada di luar sana juga banyak yang menyediakan fungsi bacaan dan juga fungsi mencari perkataan berbanding dengan keseluruhan teks perkataan yang ada di dalam pangkalan data aplikasi Al-Quran tersebut. Penggunaan fungsi carian di dalam aplikasi kebiasaannya akan menggunakan satu perkataan sebagai carian dan ia akan mencari perkataan yang sama seperti yang dimasukkan ke dalam ruang carian oleh pengguna.

Kebanyakan aplikasi carian berfungsi mencari perkataan di dalam pangkalan data menggunakan kata carian yang dimasukkan oleh pengguna. Namun, kata kunci teks pertanyaan yang dimasukkan

oleh pengguna berkemungkinan tidak sama dengan kata kunci yang dijadikan sebagai perkataan indeks dalam pangkalan data. Ini menyebabkan keputusan carian teks pertanyaan pengguna (dokumen) tidak menepati kehendak pengguna.

Salah satu cara untuk menyelesaikan permasalahan ketidaktepatan pemilihan kata kunci pertanyaan carian penggunaan adalah dengan mengubahsuai teks pertanyaan yang dimasukkan oleh pengguna. Penyahtaksaan makna perkataan ialah proses menghapuskan kekaburan yang ada pada perkataan dengan mengenal pasti erti kata sebenar perkataan yang diberikan. Sekiranya sesuatu perkataan dapat ditentukan maknanya berdasarkan konteks, makna sesuatu ayat akan lebih jelas.

Dalam penggunaan permasalahan Penyahtaksaan Makna Perkataan (PMP) untuk pengubahsuaian pertanyaan pengguna, PMP boleh membantu dalam menjelaskan makna pertanyaan carian pengguna. Contohnya, jika pengguna memasukkan teks pertanyaan perkataan 'Ashar prayer', perkataan 'prayer' dikaitkan dengan perbuatan dan bukannya bacaan doa, enjin carian akan dapat memberikan pulangan balik dokumen yang lebih tepat; iaitu yang berkaitan dengan solat dan bukan doa.

Oleh itu, PMP boleh menjadi satu komponen dalam beberapa aplikasi Capaian Maklumat (CM). Oleh kerana akan sentiasa ada kes di mana pengguna tidak menggunakan perkataan yang betul untuk menggambarkan carian yang mereka ingin dapatkan.

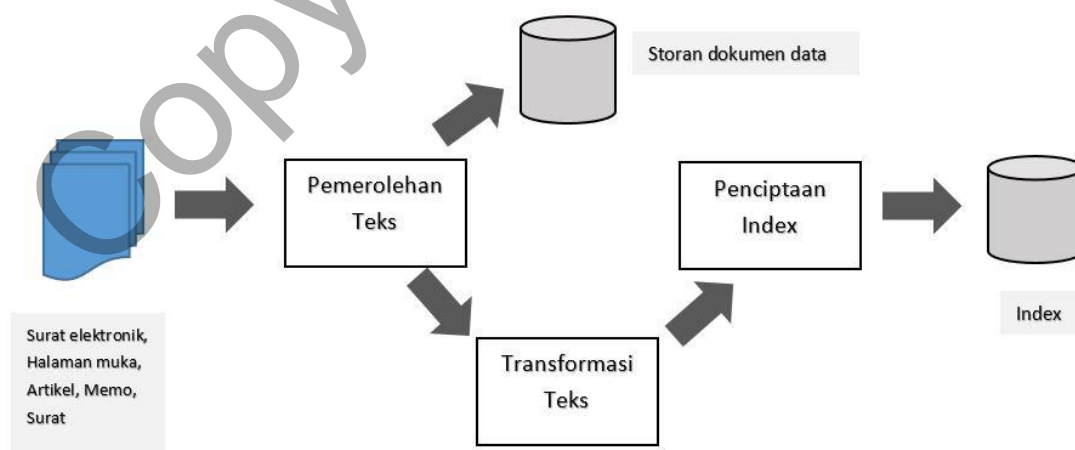
Oleh kerana ada peningkatan ke atas keputusan panggilan balik dan ketepatan akibat terkesan dari teks pertanyaan yang diubahsuai, maka usaha untuk menambahbaik cara pengubahsuaian teks pertanyaan dilihat sebagai sesuatu yang boleh memberi manfaat pada penyelidikan di dalam sistem capaian A1-Quran.

2. Penyahtaksaan Makna Perkataan dalam Capaian Maklumat

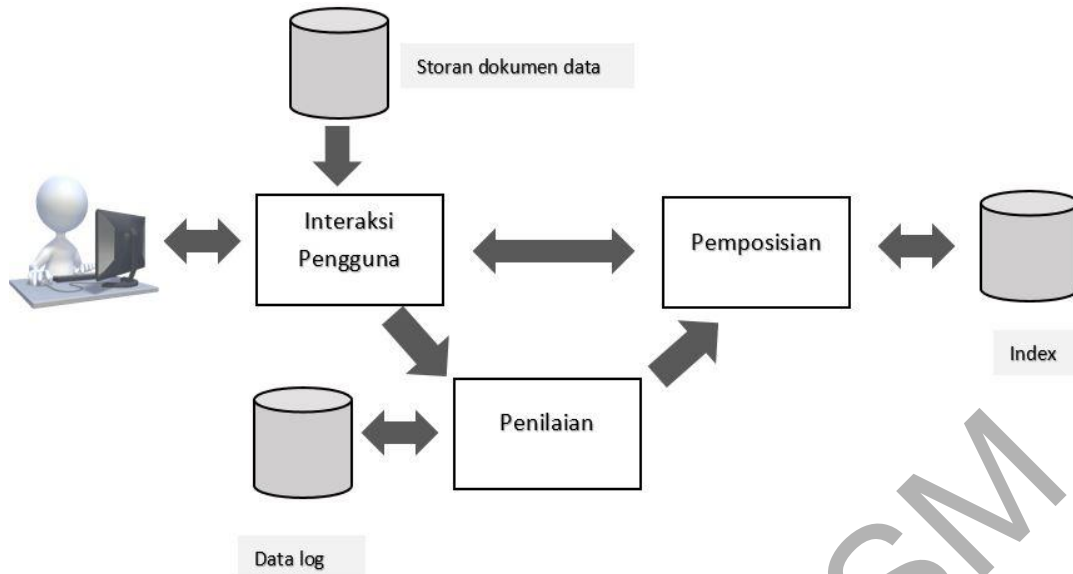
Penyahtaksaan Makna Perkataan (PMP) ialah masalah kekaburan makna perkataan keatas capaian maklumat. PMP memainkan peranan yang penting dalam beberapa persekitaran dalam pembelajaran teks khususnya, contoh seperti penterjemahan mesin (MT), capaian maklumat (CM), pengekstrakan maklumat dan perkamusan. Ketepatan CM bertambah baik melalui aplikasi PMP semenjak analisa kedalaman semantik untuk proses-proses lain adalah menunjukkan impak yang positif terhadap prestasi Capaian Maklumat, seperti juga masalah NLP yang lain.

Capaian Maklumat (CM) selalunya digunakan oleh seseorang yang terlibat dalam beberapa aktiviti pencarian bahan ilmiah: perpustakaan, peguam dan sama juga seperti penyelidik profesional.

Terdapat beberapa komponen-komponen penting dalam membangunkan enjin carian. Rajah 1 dan 2 adalah senibina enjin carian yang meliputi dua proses iaitu Proses Pengindeksan dan Proses Teks Pertanyaan.



Rajah 1 Proses Pengindeksan



Rajah 2 Proses Teks Pertanyaan

Komponen enjin carian menyokong dua fungsi utama, dimana ia dipanggil Proses Pengindeksan dan Proses Teks Pertanyaan. Proses pengindeksan membina struktur yang membolehkan pencarian, dan proses Teks Pertanyaan menggunakan struktur tersebut untuk mendapatkan senarai kedudukan dokumen yang dicari oleh seseorang menggunakan teks pertanyaan. Rajah 1 menunjukkan blok binaan bagi proses pengindeksan pada tahap tinggi. Tiga komponen yang penting adalah pemerolehan teks, transformasi teks dan penciptaan index. Dalam Rajah 2 menunjukkan blok binaan bagi proses teks pertanyaan. Fungsi utama adalah interaksi pengguna, pemposisian dan penilaian.

Bagi menyelesaikan masalah PMP ini, penyelidikan ini menggunakan beberapa pendekatan. Pendekatan-pendekatan yang digunakan adalah:

- Menggunakan synset WordNet bagi mendapatkan senarai makna perkataan yang sama maksud dengan teks pertanyaan.
- Menggunakan 4 jenis pengukuran persamaan semantik iaitu WUP, LIN, JCN dan RES bagi mendapatkan nilai persamaan semantik bagi menyinkirkan perkataan yang tiada persamaan.

Seterusnya dengan mendapatkan nilai persamaan semantik ini, teks pertanyaan asal dapat diubahsuai dengan perkataan-perkataan yang dicadangkan oleh synset *WordNet*. Dengan memperolehi teks pertanyaan yang baru akan di ukur nilai panggilan balik dan ketepatan dan *F-measure*. Nilai panggilan balik dan ketepatan dan pengukuran F sebelum dan selepas pengubahsuain teks pertanyaan akan dicatit.

3. Perbandingan Nilai Keberkesanan Perubahan Teks Pertanyaan

Metod yang digunakan di dalam kajian ini mengandungi beberapa fasa. Fasa pertama adalah mendapatkan nilai prestasi kecekapan bagi teks pertanyaan asal terhadap set data menggunakan aplikasi Indri. Aplikasi Indri boleh di muat turun dengan menggunakan pautan <https://lemurproject.org>.

Fasa ini akan menggunakan tujuh teks pertanyaan untuk mengukur kecekapan dengan mendapatkan nilai panggilan balik dan ketepatan dan *F-measure*.

Fasa kedua pula menggunakan kaedah pemprosesan teks. Fasa ini akan menapis teks pertanyaan daripada tanda-tanda berhenti, simbol atau kata-kata am yang tiada maksud. Proses-proses yang terlibat dalam menapis kandungan teks pertanyaan menjadi satu bentuk pertanyaan yang bebas dari perkataan yang tiada maksud adalah seperti berikut;

- 1) Penyingkiran simbol atau tanda baca
- 2) Penyingkiran kata am,
- 3) Penyingkiran kata imbuhan

Fasa ketiga melibatkan kerja-kerja penyahtaksan makna perkataan menggunakan *WordNet* synset dengan menggunakan pendekatan Lesk. Senarai makna yang diperolehi dari *WordNet* akan dipilih mengikut kesesuaian maksud keatas perkataan sasaran. Beberapa senarai makna yang dipilih akan di jadikan uji kaji seterusnya untuk menyaring perkataan menggunakan aplikasi *Word Similarity* dari Hideki Shima (2015). Aplikasi ini boleh didapati di pautan <http://ws4jdemo.appspot.com>.

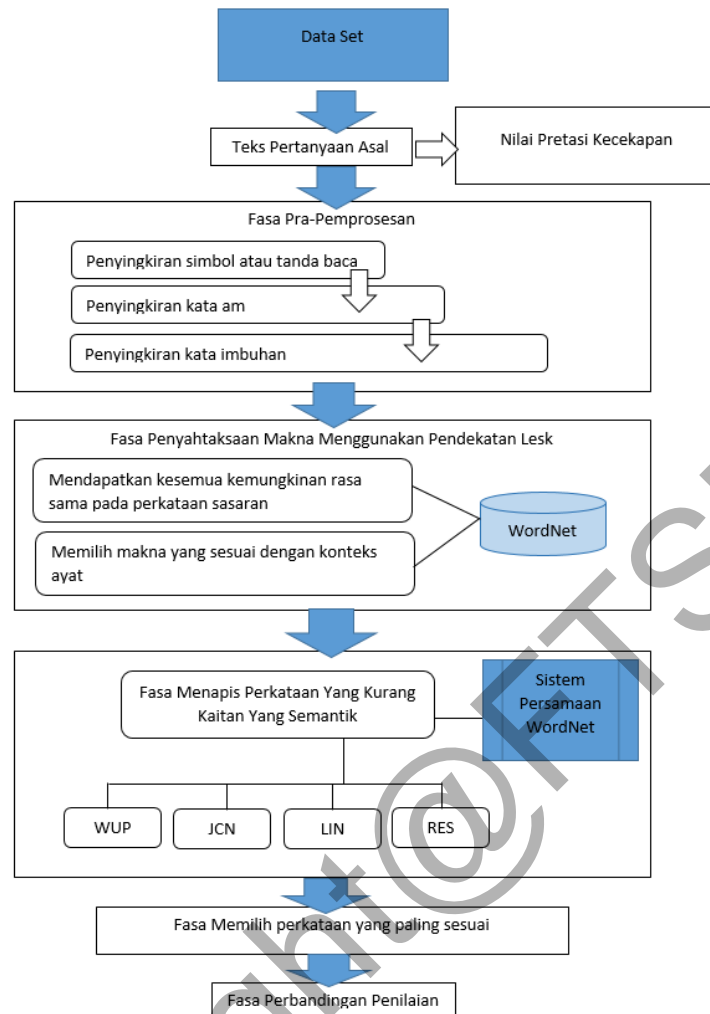
Aplikasi ini membekalkan beberapa modul persamaan perkataan yang menyediakan pendekatan-pendekatan seperti berikut;

- 1) Algoritma Wu Palmer (WUP)
- 2) Algoritma Lin (LIN)
- 3) Algoritma Jiang Conrath (JCN)
- 4) Algoritma Resnik Algorithm (RES)

Proses bagi fasa keempat ini akan mendapatkan nilai persamaan semantik antara perkataan diperolehi dari *WordNet* dengan teks pertanyaan asal yang telah melalui proses Teks Pemprosesan. Algoritma-algoritma ini akan dilaksanakan keatas setiap teks pertanyaan asal untuk mendapatkan nilai persamaan semantik terhadap perkataan yang diperolehi dari *WordNet*.

Skor yang diperolehi dari setiap algoritma ini akan dibandingkan dengan setiap algoritma di mana perkataan yang memperoleh nilai adalah mempunyai kaitan. Seterusnya perkataan yang mendapat nilai kaitan dan juga mendapat nilai kaitan daripada algoritma yang lain akan dikekalkan dan dipadankan kepada teks pertanyaan asal. Kemudian teks pertanyaan tersebut akan menjadi teks pertanyaan yang baru.

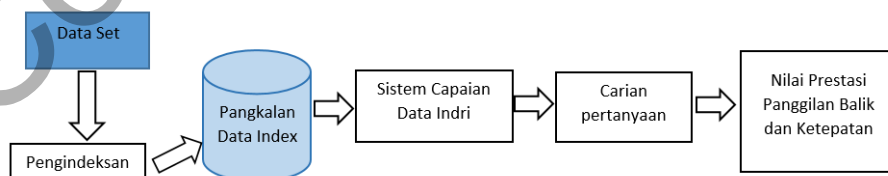
Fasa kelima iaitu fasa yang terakhir sekali adalah fasa perbandingan nilai prestasi kecekapan antara dua teks pertanyaan asal dan teks pertanyaan yang diubahsuai yang diperolehi dari bacaan aplikasi Indri. Rajah 1 menunjukkan reka bentuk am terhadap fasa-fasa yang teribat.



Rajah 1 Reka bentuk am terhadap fasa terlibat

3.1 Fasa Pemilihan Teks Pertanyaan Asal Dan Mendapatkan Nilai Kecekapan

Fasa pertama adalah proses pemilihan teks pertanyaan dan mendapatkan nilai kecekapan. Untuk mendapatkan bacaan prestasi kecekapan set data yang digunakan adalah dari 6 surah dari Al-Quran. Rajah 2 menunjukkan seni bina proses capaian nilai prestasi kecekapan Indri. Nilai prestasi kecekapan ini akan di catat sebagai nilai awal bagi teks pertanyaan yang asal.



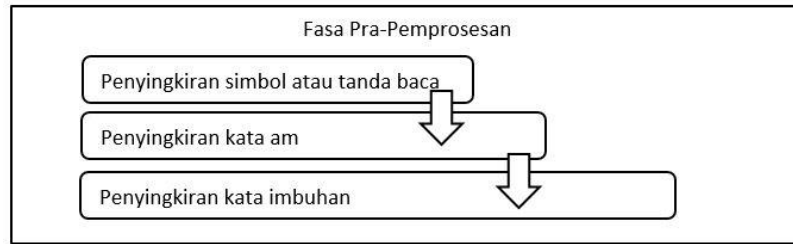
Rajah 2 Proses awal mendapatkan nilai prestasi kecekapan teks pertanyaan asal menggunakan aplikasi

Indri

3.2 Fasa Pra-Pemrosesan Teks

Fasa Pra-Pemrosesan Teks adalah fasa menyingkirkan perkataan atau simbol-simbol yang tiada makna dalam konteks ayat. Fasa ini adalah memastikan perkataan yang ada dalam teks pertanyaan adalah bebas dari segala imbuhan, kata henti, kata am dan simbol.

Langkah-langkah ini dipakai ke atas dokumen yang diambil dari teks Al-Quran. Rajah 3 menunjukkan rangka Pra-Pemrosesan teks.



Rajah 3 Proses Pra-Pemrosesan Teks

i. Penyingkiran simbol atau tanda baca

Secara asasnya, proses tokenizer ini adalah proses penyingkiran simbol atau tanda baca yang membahagikan turutan teks kepada ayat dan kemudian ayat kepada token. Oleh itu, sebelum mula pemrosesan, teks hendaklah di bahagikan kepada ayat-ayat dan perkataan dalam bahasa Inggeris, perkataan yang terikat oleh ruang kosong dan selalunya didahului dan diikuti oleh tanda kurungan, tanda pembuka kata atau tanda seru.

ii. Pembuangan Kata Am

Dalam perspektif pemrosesan teks, kata am selalunya adalah perkataan yang kosong dan cuma membazir ruang dokumen. Oleh itu penterjemahan dokumen yang telah di proses dengan mengeluarkan kata umum dari dokumen adalah penting bagi di implementasi dalam bidang Capaian Maklumat.

iii. Proses Stemmer

Proses stemmer perkataan dianggap sebagai salah satu aspek yang penting bagi pra-pemrosesan, dimana ia mempengaruhi kecekapan pengukuran persamaan antara teks pendek atau ayat. Proses stemmer digunakan untuk menapis perkataan menjadi kata akar mereka. Dalam kata lain, stemmer adalah proses menjana kata dasar, dimana ia dengan mudah boleh membandingkan dengan kata-kata lain melalui penghapusan imbuhan awal dan akhir dalam perkataan.

3.3 Fasa Penyahtaksan Makna Menggunakan Pendekatan Lesk

Fasa ini menggunakan synset *WordNet* yang menggunakan pendekatan Lesk bagi mendapatkan maksud hampir sama dengan teks pertanyaan sasaran. Rajah 4 menunjukkan hasil capaian synset *WordNet* bagi teks pertanyaan 'sign dooms quran day'.

```

>>> print(lesk(sent, 'sign'))
Synset('sign.v.07')
>>> from nltk.corpus import wordnet as wn
>>> for ss in wn.synsets('sign'):
...     print(ss, ss.definition())
...
Synset('sign.n.01') a perceptible indication of something not immediately
apparent (as a visible clue that something has happened)
Synset('sign.n.02') a public display of a message
Synset('signal.n.01') any nonverbal action or gesture that encodes a message
Synset('signboard.n.01') structure displaying a board on which advertisements
can be posted
Synset('sign_of_the_zodiac.n.01') (astrology) one of 12 equal areas into which
the zodiac is divided
Synset('sign.n.06') (medicine) any objective evidence of the presence of a
disorder or disease
Synset('polarity.n.02') having an indicated pole (as the distinction between
positive and negative electric charges)
Synset('augury.n.01') an event that is experienced as indicating important
things to come
Synset('sign.n.09') a gesture that is part of a sign language
Synset('sign.n.10') a fundamental linguistic unit linking a signifier to that
which is signified
Synset('sign.n.11') a character indicating a relation between quantities
Synset('sign.v.01') mark with one's signature; write one's name (on)
Synset('sign.v.02') approve and express assent, responsibility, or obligation
Synset('sign.v.03') be engaged by a written agreement
Synset('sign.v.04') engage by written agreement
Synset('sign.v.05') communicate silently and non-verbally by signals or signs
Synset('sign.v.06') place signs, as along a road
Synset('sign.v.07') communicate in sign language
Synset('bless.v.03') make the sign of the cross over someone in order to call
on God for protection; consecrate
Synset('gestural.s.01') used of the language of the deaf
>>>

>>> print(lesk(sent, 'dooms'))
Synset('sentence.v.01')
>>> from nltk.corpus import wordnet as wn
>>> for ss in wn.synsets('dooms'):
...     print(ss, ss.definition())
...
Synset('doom.n.01') an unpleasant or disastrous destiny
Synset('destine.v.01') decree or designate beforehand
Synset('sentence.v.01') pronounce a sentence on (somebody) in a court of law
Synset('doom.v.03') make certain of the failure or destruction of
>>>

-----

>>> print(lesk(sent, 'quran'))
Synset('koran.n.01')
>>> from nltk.corpus import wordnet as wn
>>> for ss in wn.synsets('quran'):
...     print(ss, ss.definition())
...
Synset('koran.n.01') the sacred writings of Islam revealed by God to the
prophet Muhammad during his life at Mecca and Medina
>>>

-----

>>>
>>> print(lesk(sent, 'day'))
Synset('sidereal_day.n.01')
>>> from nltk.corpus import wordnet as wn
>>> for ss in wn.synsets('day'):
...     print(ss, ss.definition())
...
Synset('day.n.01') time for Earth to make a complete rotation on its axis
Synset('day.n.02') some point or period in time
Synset('day.n.03') a day assigned to a particular purpose or observance
Synset('day.n.04') the time after sunrise and before sunset while it is light

```

Rajah 4 Senarai maksud dari synsnet *WordNet*

3.4 Fasa Menapis Perkataan Yang Kurang Nilai Persamaan Semantik

Fasa seterusnya adalah fasa menapis perkataan yang kurang persamaan semantik dengan teks pertanyaan asal yang disusun secara matriks. Bahagian Lajur adalah untuk teks pertanyaan asal dan

bahagian Baris adalah bagi senarai maksud diperolehi dari WordNet. Rajah 5 adalah hasil eksperimen untuk pendekatan WUP, JCN, LIN dan RES.

WUP						JCN										
	sign /NN	dooms /NNS	day /NN	quran /.		sign /NN	dooms /NNS	day /NN	quran /.		sign /NN	dooms /NNS	day /NN	quran /.		
perceptible/JJ	-	-	-	-	-	perceptible/JJ	-	-	-	-	-	perceptible/JJ	-	-	-	-
apparent/JJ	-	-	-	-	-	apparent/JJ	-	-	-	-	-	apparent/JJ	-	-	-	-
visible/JJ	-	-	-	-	-	visible/JJ	-	-	-	-	-	visible/JJ	-	-	-	-
clue/NN	0.9333	0.4286	0.5000	-	-	clue/NN	9.4912	0.0559	0.0744	-	-	clue/NN	8.7700	0.7794	0.7794	-
∅:	-	-	-	-	-	∅:	-	-	-	-	-	∅:	-	-	-	-
failure/NN	0.8000	0.8000	0.7619	-	-	failure/NN	0.0942	0.0917	0.0771	-	-	failure/NN	4.0796	4.0796	3.1688	-
destruction/NN	0.7500	0.7500	0.6667	-	-	destruction/NN	0.0848	0.0828	0.0746	-	-	destruction/NN	4.0796	4.0796	3.1688	-
∅:	-	-	-	-	-	∅:	-	-	-	-	-	∅:	-	-	-	-
point/NN	0.8889	0.5000	0.7273	-	-	point/NN	0.1277	0.0689	0.1546	-	-	point/NN	9.1267	1.7798	3.1688	-
period/NN	0.6667	0.4444	0.9231	-	-	period/NN	0.1168	0.0777	0.3196	-	-	period/NN	6.3586	1.7798	5.9789	-
in/IN	-	-	-	-	-	in/IN	-	-	-	-	-	in/IN	-	-	-	-
time/NN	0.8750	0.7500	0.9333	-	-	time/NN	0.2352	0.1126	0.6587	-	-	time/NN	7.7955	4.0796	5.9850	-
∅:sacred/JJ	-	-	-	-	-	∅:sacred/JJ	-	-	-	-	-	∅:sacred/JJ	-	-	-	-
writing/NN	0.8000	0.6250	0.5000	-	-	writing/NN	0.2037	0.0725	0.1067	-	-	writing/NN	7.3837	2.2541	1.3696	-
reveal/VBP	-	-	-	-	-	reveal/VBP	-	-	-	-	-	reveal/VBP	-	-	-	-
prophet/NN	0.5556	0.2667	0.7619	-	-	prophet/NN	0.0689	0.0531	0.0695	-	-	prophet/NN	1.3696	0.0000	1.9033	-
muhammad/JJ	-	-	-	-	-	muhammad/JJ	-	-	-	-	-	muhammad/JJ	-	-	-	-
life/NN	0.7059	0.5714	0.8571	-	-	life/NN	0.1724	0.0677	0.1257	-	-	life/NN	6.2686	1.7798	5.6455	-
Mecca/NNP	0.6667	0.2500	0.4211	-	-	Mecca/NNP	0.0000	0.0000	0.0000	-	-	Mecca/NNP	3.0088	0.0000	1.1692	-
Medina/NNP	0.5882	0.2222	0.3810	-	-	Medina/NNP	0.0000	0.0000	0.0000	-	-	Medina/NNP	3.0088	0.0000	1.1692	-

LIN						RES										
	sign /NN	dooms /NNS	day /NN	quran /.		sign /NN	dooms /NNS	day /NN	quran /.		sign /NN	dooms /NNS	day /NN	quran /.		
perceptible/JJ	-	-	-	-	-	perceptible/JJ	-	-	-	-	-	perceptible/JJ	-	-	-	-
apparent/JJ	-	-	-	-	-	apparent/JJ	-	-	-	-	-	apparent/JJ	-	-	-	-
visible/JJ	-	-	-	-	-	visible/JJ	-	-	-	-	-	visible/JJ	-	-	-	-
clue/NN	0.9940	0.0802	0.1039	-	-	clue/NN	8.7700	0.7794	0.7794	-	-	clue/NN	8.7700	0.7794	0.7794	-
∅:	-	-	-	-	-	∅:	-	-	-	-	-	∅:	-	-	-	-
failure/NN	0.4345	0.4280	0.1072	-	-	failure/NN	4.0796	4.0796	3.1688	-	-	failure/NN	4.0796	4.0796	3.1688	-
destruction/NN	0.4090	0.4032	0.1042	-	-	destruction/NN	4.0796	4.0796	3.1688	-	-	destruction/NN	4.0796	4.0796	3.1688	-
∅:	-	-	-	-	-	∅:	-	-	-	-	-	∅:	-	-	-	-
point/NN	0.5640	0.1969	0.4770	-	-	point/NN	9.1267	1.7798	3.1688	-	-	point/NN	9.1267	1.7798	3.1688	-
period/NN	0.1540	0.1080	0.7063	-	-	period/NN	6.3586	1.7798	5.9789	-	-	period/NN	6.3586	1.7798	5.9789	-
in/IN	-	-	-	-	-	in/IN	-	-	-	-	-	in/IN	-	-	-	-
time/NN	0.7857	0.4788	0.8875	-	-	time/NN	7.7955	4.0796	5.9850	-	-	time/NN	7.7955	4.0796	5.9850	-
∅:sacred/JJ	-	-	-	-	-	∅:sacred/JJ	-	-	-	-	-	∅:sacred/JJ	-	-	-	-
writing/NN	0.5559	0.2250	0.1426	-	-	writing/NN	7.3837	2.2541	1.3696	-	-	writing/NN	7.3837	2.2541	1.3696	-
reveal/VBP	-	-	-	-	-	reveal/VBP	-	-	-	-	-	reveal/VBP	-	-	-	-
prophet/NN	0.1478	0.0000	0.0000	-	-	prophet/NN	1.3696	0.0000	1.9033	-	-	prophet/NN	1.3696	0.0000	1.9033	-
muhammad/JJ	-	-	-	-	-	muhammad/JJ	-	-	-	-	-	muhammad/JJ	-	-	-	-
life/NN	0.6837	0.1943	0.4861	-	-	life/NN	6.2686	1.7798	5.6455	-	-	life/NN	6.2686	1.7798	5.6455	-
Mecca/NNP	0.0000	0.0000	0.0000	-	-	Mecca/NNP	3.0088	0.0000	1.1692	-	-	Mecca/NNP	3.0088	0.0000	1.1692	-
Medina/NNP	0.0000	0.0000	0.0000	-	-	Medina/NNP	3.0088	0.0000	1.1692	-	-	Medina/NNP	3.0088	0.0000	1.1692	-

Rajah 5 Bacaan persamaan semantik bagi WUP, JCN, LIN dan RES

Dengan memilih perkataan yang memperoleh skor akan dikekalkan sebagai input seterusnya, input yang dikekalkan akan digunakan untuk proses seterusnya.

3.5 Fasa Perbandingan

Fasa ini melibatkan perbandingan nilai prestasi kecekapan antara dua teks pertanyaan asal dan teks pertanyaan yang diubahsuai. Rajah 6 adalah keputusan awal di fasa pertama adalah hasil dari pengukuran prestasi kecekapan bagi teks pertanyaan asal iaitu 'What the signs of the Dooms Day that are stated in the *Quran*'.

Setelah melalui proses pengembangan, teks pertanyaan bagi 'What the signs of the Dooms Day that are stated in the *Quran*' yang telah diubah suai menjadi seperti 'sign dooms day clue failure destruction point period time writing life'. Untuk membandingkan dokumen yang dapat dipanggil balik dengan menggunakan dua teks pertanyaan asal dan yang diubah suai, aplikasi Indri sekali lagi digunakan untuk mendapatkan dokumen yang relevan. Rajah 6 dan Rajah 7 menunjukkan nilai prestasi kecekapan sebelum dan selepas.

Indri -- Retrieval

File Help

Enter your query: what the signs of the dooms day that are stated in the quran

Number of documents: 6

Database(s): D:\Master\emur_test\Index

Search Clear Hide Scores

done.

Score	Document	Title
-5.837	SURAH1(TOTAL7)-The Opening	
-5.971	SURAH5(TOTAL120)-The Food	
-6.069	SURAH3(TOTAL200)-The Family Of Imran	
-6.091	SURAH2(TOTAL286)-The Cow	
-6.114	SURAH6(TOTAL165)-The Cattle	
-6.274	SURAH4(TOTAL176)-Women	

Rajah 6 Keputusan nilai prestasi relevan dokumen Indri bagi teks pertanyaan asal

Indri -- Retrieval

File Help

Enter your query: sign dooms day clue failure destruction point period time writing life

Number of documents: 6

Database(s): D:\Master\emur_test\Index

Search Clear Hide Scores

done.

Score	Document	Title
-9.492	SURAH1(TOTAL7)-The Opening	
-10.025	SURAH2(TOTAL286)-The Cow	
-10.115	SURAH6(TOTAL165)-The Cattle	
-10.36	SURAH3(TOTAL200)-The Family Of Imran	
-10.426	SURAH5(TOTAL120)-The Food	
-10.465	SURAH4(TOTAL176)-Women	

Rajah 7 Keputusan nilai prestasi relevan dokumen Indri bagi teks pertanyaan diubahsuai

Jika dilihat dari skor prestasi, nilai skor adalah semakin menurun, ia mungkin yang terbaik untuk menganggap bahawa nilai-nilai ini tidak dapat dibandingkan seluruh pertanyaan. Khususnya, ia mungkin akan melihat bahawa apabila kajian ini menambah perkataan ke dalam teks pertanyaan, skor dokumen purata cenderung untuk jatuh, walaupun sistem itu mungkin menjadi lebih baik dalam mencari dokumen yang baik.

3.6 Fasa Penilaian

Fasa Penilaian adalah fasa menilai dua keputusan nilai prestasi kecekapan teks pertanyaan asal dan teks pertanyaan diubah suai. Teks pertanyaan asal iaitu 'What the signs of the Dooms Day that are stated in the Quran' akan dibandingkan nilai prestasi kecekapan dengan teks pertanyaan yang diubah suai iaitu 'sign dooms day clue failure destruction point period time writing life'.

Dengan menggunakan sistem Indri untuk mendapatkan nilai skor dokumen pada setiap dokumen yang relevan, kajian ini juga menggunakan sistem Indri untuk mengukur panggilan balik dan ketepatan dan juga nilai F-measure bagi setiap dokumen yang ada. Untuk mengukur panggilan balik dan ketepatan bagi setiap dokumen formula di bawah digunakan;

$$\text{Ketepatan, } P = \frac{|TP|}{|TP| + |FP|}$$

$$\text{Panggilan Balik, } R = \frac{|TP|}{|TP| + |FN|}$$

$$\text{Pengukuran } F = \frac{2 \times |TP|}{(|TP| + |FP|) + (|TP| + |FN|)} = 2 \frac{PR}{P \times R}$$

Pada jadual 1 menunjukkan keputusan perbandingan panggilan balik dan ketepatan dan *F-measure*.

Jadual 1 Keputusan perbandingan panggilan balik dan ketepatan dan *F-measure*.

	Teks Pertanyaan 1	Teks Pertanyaan 2	Teks Pertanyaan 3	Teks Pertanyaan 4	Teks Pertanyaan 5	Teks Pertanyaan 6	Teks Pertanyaan 7
Panggilan Balik Pertanyaan Asal	0.583	0.583	0.58	0.58	0.58	0.58	1
Pertanyaan diubahsuai	0.5	1	0.6	0.6	0.6	0.58	0.5
Ketepatan Pertanyaan Asal	1	1	1	1	1	1	0.5
Pertanyaan diubahsuai	0.5916	1	0.71	0.71	0.71	1	1
<i>F-measure</i> Pertanyaan Asal	0.7365	0.7365	0.7341	0.7341	0.7341	0.7341	0.6666
Pertanyaan diubahsuai	0.5419	0.7365	0.6503	0.6503	0.6503	0.7341	0.6666
<i>Peratus Pengukuran F (%)</i>	26.422	0	11.415	11.415	11.415	0	0

Pengembangan teks pertanyaan menggunakan persamaan semantik dan penggunaan synset *WordNet* yang menggunakan kaedah Lesk dapat menunjukkan peningkatan terhadap nilai panggilan balik dan ketepatan walaupun ada juga menampakkan penurunan nilai pada teks pertanyaan tertentu.

4. KESIMPULAN

Fokus utama dalam penyelidikan perbandingan ini ialah menyelesaikan masalah Penyahtaksaan Makna Perkataan (PMP). Penyelidikan ini tertumpu pada domain tertutup yang menggunakan surah-surah Al-Quran sebagai koleksi dokumen dan menggunakan domain terbuka sebagai sumber rujukan bagi koleksi perkataan yang diperolehi dari synset *WordNet*. Set data yang digunakan di dalam penyelidikan ini adalah dari surah-surah Al-Quran yang telah di alih bahasa ke bahasa Inggeris.

Laporan ini memberi fokus kepada prosedur teknikal bagi menampakkan cara untuk menyelesaikan masalah PMP ini. Aplikasi luar yang digunakan adalah Indri dan Word Similarity For Java (<http://ws4jdemo.appspot.com/>).

Kajian ini dijalankan agar dapat menghasilkan keputusan skor nilai yang berbeza keatas pengembangan teks pertanyaan dari sudut prestasi kecekapan panggilan balik dan kejituan menggunakan set data keatas perkataan dari ayat suci Al-Quran.

RUJUKAN

Hideki Shima. <http://ws4jdemo.appspot.com>. [10 November 2016].

Lemur Project. <https://lemurproject.org>. [10 November 2016].

Roslina, O . 2011. Information Retrieval: Design And Source. First Edition, Selangor: IIUM Press.

Copyright@FTSM