JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 5
No. 2
Agustus
2019

Features of Distributional Method for Indonesian Word Clustering

Herry Sujaini[#1]

[#]*Department of Informatics, University of Tanjungpura*
Jl. Prof.Hadari Nawawi, Pontianak, 78124, Indonesia
[1]hs@untan.ac.id

Abstract— **We described the results of a study to determine the best features for algorithm EWSB (Extended Word Similarity Based). EWSB is a word clustering algorithm that can be used for all languages with a common feature. We provided four alternative features that can be used for word similarity computation and experimented toward the Indonesian Language to determine the best feature format for the language. We found that the best feature used in the algorithm to Indonesian EWSB is t w w' format (3-gram) with 0 (zero) word relation. Moreover, we found that using 3-gram is better than 4-gram for all the proposed features. Average recall of 3-gram is 83.50%, while the average 4-gram recall is 57.25%.**

***keywords*— n-gram, word clustering, word similarity, EWSB.**

I. Introduction

Word similarity can be computed by measuring the semantic distance in a thesaurus like WordNet or MeSH (thesaurus methods), by using distributional similarity in a corpus, or by using information-theoretic methods [1]. Thesaurus methods have a weakness, mainly because we don't have such thesauruses for every language. Even if we do, they have problems with recall, including many words are missing, most phrases are missing, some connections between senses are missing, and thesauri work less well for verbs and adjectives. In additional, thesaurus methods only work if rich hyponymy knowledge is present in the thesaurus. We focus on distributional rather than semantic similarity because of the low resource of Indonesian language, including the semantic resource.

The intuition of distributional methods is that the meaning of a word is related to the distribution of words and punctuation marks around it. In distributional methods, we can represent a word as a feature vector. For example, suppose we had one binary feature f_i representing each of the N words in the lexicon v_i. Two entities can be said to be similar if they have similar characteristics or features; if some entities are grouped, they will be processed on the degree of similarity of each entity to one another. Because of the features possessed by an entity usually very much, usually those features selected or given weight in accordance with the purpose of the grouping.

If we define a universe, or a set containing "father, mother, and son", grouping with a bigger weight in the recommended age group would result in separating "father and mother" with "son". While grouping with a bigger weight on gender feature that separates the group will produce a "father and son" with "mother". Selected features on a method determine the outcome of a process that uses such a method.

Contextual word similarity can be determined by looking at the distribution of these words in a sentence. The intuition of distributional methods is that the meaning of a word is related to the distribution of words around it. For example, suppose there are three Indonesian sentences in the corpus as follows :

***Jokowi** segera berkonsentrasi menghadapi pilkada DKI Jakarta,*
*Pesaing **Jokowi** juga berasal dari Amerika Utara,*
*"Waduh , no comment. Bukan wilayah saya," kata **Jokowi**.*

From these sentences, the features for the word "*Jokowi*" can be determined, for example, "appears at the beginning of the sentence before the word *segera*", "appears immediately after the word *pesaing*", "appears after the word *kata* and located at the end of the sentence" and others. If there are other words that also have such a feature, it can be said that the word is similar in context with the word "*Jokowi*". In general, the features can be defined as "a word w that appears around the word v_i" . For computational purposes, the features of a word in the sentence needs to be defined more specifically.

We describe the results of a study to determine the contextual word similarity features to words clustering in Indonesian is appropriate. Issues raised in this study is a feature of what is best for determining the similarity of two words in Indonesian through the distributional approach. Thus, the purpose of this research is to find the best feature of these problems.

The semantic similarity of words is a longstanding topic in computational linguistics because it is theoretically intriguing and has many applications in the field. Ker and Zhang [2] used man-made thesauri in their work to help to align words. Many researchers have conducted studies based on the distributional hypothesis [3], which states that words that occur in the same contexts tend to have similar meanings. A number of semantic similarity measures have been proposed based on this hypothesis [4-9].

A number of semantic clustering algorithms have been reported, such as those in [8, 10-18]. Some work has thus focused on a re-ranking strategy, Geffet and Dagan [12,19] improved the output of a distributional similarity system for an entailment task using a web-based feature inclusion check, and comment that their filtering produces better outputs than cutting off the similarity pairs with the lowest ranking.

II. Methodology

Jeff et al. [17] developed an algorithm based on the Lin [8] and named it word-similarity-based (WSB) clustering algorithm. Based on the "WSB algorithm", Sujaini [18] developed the algorithm and named it EWSB (Extended Word Similarity Based) clustering algorithm. WSB algorithm proposed by Jeff et al. [17] using the feature Tw (r,w_2), where (w_1,r,w_2) is taken from the n-gram that starts with w_1 and ends with w_2. In EWSB algorithm, Sujaini et al. [18] used the feature Tw (t,r,w_2), where (t,w_1,r,w_2) is taken from the n-gram with the position w_1,r, and w_2 are varies.

We tested the position variations w_1,r, and w_2 in Indonesian to obtain the best configuration of w_1,r, and w_2. In this experiment, we tested 4 (four) variations each using 3-gram and 4-gram. Word similarity of the equation (3) is modified into [18]:

$$S_1(w_1, w_2) = \frac{\sum_{(t,r,w)\in T,(w_1)\cap T(w_2)}[I(t,w_1,r,w).I(t,w_2,r,w)]}{\sum_{(t,r,w)\in T,(w_1)} I(t,w_1,r,w)+\sum_{(t,r,w)\in T,(w_2)} I(t,w_2,r,w)} \quad (1)$$

We used equation (1) for the t w w' dan t w r w' formats, while for other formats, equation (5) is modified into:

$$S_1(w_1, w_2) = \frac{\sum_{(t,r,w)\in T,(w_1)\cap T(w_2)}[I(t,w,r,w_1).I(t,w,r,w_2)]}{\sum_{(t,r,w)\in T,(w_1)} I(t,w,r,w_1)+\sum_{(t,r,w)\in T,(w_2)} I(t,w,r,w_2)} \quad (2)$$

for w' w and t w' r w,

$$S_1(w_1, w_2) = \frac{\sum_{(t,r,w)\in T,(w_1)\cap T(w_2)}[I(w_1,r,w,t).I(w_2,r,w,t)]}{\sum_{(t,r,w)\in T,(w_1)} I(w_1,r,w,t)+\sum_{(t,r,w)\in T,(w_2)} I(w_2,r,w,t)} \quad (3)$$

for w w' t and w r w' t, and

$$S_1(w_1, w_2) = \frac{\sum_{(t,r,w)\in T,(w_1)\cap T(w_2)}[I(w,r,w_1,t).I(w,r,w_2,t)]}{\sum_{(t,r,w)\in T,(w_1)} I(w,r,w_1,t)+\sum_{(t,r,w)\in T,(w_2)} I(w,r,w_2,t)} \quad (4)$$

for w' w t and w' r w t.

Variable t in equation (1), (2), (3) and (4) is a word in the word window that can be positioned left or right of the word window, while the relation (r) is between w and w' which can consist of 0 (zero) or more words.

In this work, we perform a comparison of clustering algorithms EWSB with variation in n-gram features. We conducted this experiment to determine the most appropriate features for Indonesian. In this experiment, we used 171K sentences Indonesian corpus, as shown in Figure. 1 which has the characteristics : 3,406,412 tokens, tokens of each sentence mean of 19.9, and 114,758 unique tokens. The number of words distributed between 1 and 97 words with an average of 20 words per sentence. The 10 tokens with the highest count in the corpus are :

1. , (188,043),
2. *yang* (102,882),
3. *dan* (84,293),
4. *di* (44,594),
5. *dengan* (36,783),
6. *itu* (33,123),
7. *untuk* (29,444),
8. *dari* (28,687),
9. *dalam* (27,442), and
10. *tidak* (26,65).

tetapi , berkoinsiden dengan ditetapkannya tiga orang anggota polisi sebagai tersangka perampokan dan pembunuhan seorang sopir taksi , polri memberikan " hadiah lebaran " yang menghebohkan : dr. azahari tewas dalam penggerebekan di sebuah rumah di kota wisata batu di malang
kapolri memastikan identitas mayat sang pakar bom dengan data sidik jari yang , katanya , " memiliki tingkat akurasi lebih tinggi dibanding tes dna "
meskipun kontroversial , kata pakar dirgantara jusman s.d. dan purnawirawan tni hasnan habib , pembelian pesawat sukhoi jangan dibatalkan
sebab , jika itu dilakukan , akan menampar muka presiden dan rusia , kata jusman (republika)
beberapa pakar teknologi pun telah memperkirakan akan adanya revolusi sosial karena berkonvergensinya teknologi informasi dengan ilmu-ilmu lain seperti biologi , kimia , fisika , dan matematika
tidak heran jika nanti akan muncul model e-business dimana fungsi seorang staf telah dapat digantikan oleh robot hasil kloning , atau fungsi seorang manajer telah digantikan oleh robot pintar berbasis kecerdasan buatan , atau produk-produk fisik telah dapat digantikan oleh produk-produk imitasi yang diciptakan secara instan , dan lain sebagainya
teknologi informasi jugalah yang telah mematikan dan menimbun batas-batas geografis dan waktu sehingga setiap individu dapat berinteraksi dengan individu lain dalam hitungan detik

Figure 1. Example of Indonesian corpus

We conducted an experiment on 100 pairs of words that are considered similar to determine the best features of EWSB algorithm for Indonesian manually. 100 pairs of test samples taken from the word unigram sorted from the largest value and sampled varies based on the types of word classes. The inputs for this system are 200 words without their pairs information; the system output is a clustering result, that output compared against the reference word pairs. To the test words, we conducted experiments using features that varied by changing the

position of t, w, and w'. In this experiment, we used 3-gram and 4-gram which four variations each of
t w w',
t w' w,
w w' t, and
w' w t
for 3-gram, and
t w r w',
t w' r w,
w r w' t, and
w' r w t
to 4-gram.

Totally, we conducted 8 (eight) times experiments with the different features for the same test words.
We used Newick format to describing the agglomerative word clustering process and customized an approach to get the history of clustering. Newick format (Newick notation) is a way to represent graph-theoretical trees by using parentheses and commas [20]. Agglomerative algorithms which have been adjusted to obtain the results of the Newick format is as follows :

1. Initialize each unique word (token) as a cluster
2. Calculate the similarity between two clusters
3. Sort ranking between all pairs of clusters based on similarity, then combine the two top clusters
4. Add clusters are combined in Newick format
5. Stop until it reaches a single cluster if not, return to step 2.

III. Results and Discussion

Results of hierarchical clustering illustrated with a dendrogram, where the dendrogram is a curve that describes the cluster grouping. At this stage, Newick format generated in the previous stage be used as input to obtain a visualization cluster dendrogram. After that, we compared the results of each feature with reference to the word pairs and computed its precision and recall. Example of the system output to variations t w w' as Newick format. Before we did the clustering process, we computed the word similarity between the words that define the input words. Word similarity score (top 20) is shown in Table I.

Experiment result for t w w' format shows that of 200 words have a pair, 196 words (98 words pair) clustered correctly according to the word pair in the initial clustering. As shown in Figure 2, four words that fail merged with its pair are “meskipun”, “walaupun”, “mulai”, and “selesai”. The word "walaupun" not directly affiliated with "meskipun", but first joined to the cluster ("tapi" and "tetapi"), and then joined with the word "meskipun" . The word "mulai" joined to the cluster (“tidak”, “tak”, “sudah”, “telah”, “ingin” and “mau”) while the word "selesai" joined to the cluster (“tertawa”, “menangis”, “diperiksa”, and “ditahan”). Thus the feature with t w w' form produced a precision value of 98/98 = 100% and a recall of 98/100 = 98% . Precision value shows the percentage of correct pairs to the number of pairs found, while recall shows the percentage of correct pairs to the number of reference pair. Precision value of 100 % means that all pairs are found to be true, while the recall value of 98% means that there is a 2% reference pair that is not found in the output.

TABLE I
Word Similarity Score of t,w,w'

Word 1	Word 2	Word Similarity Score
primer	*sekunder*	0.17842
kanan	*kiri*	0.17115
ratus	*puluh*	0.17009
1	*2*	0.16805
dua	*tiga*	0.14473
gadis	*wanita*	0.13076
berdua	*bertiga*	0.13075
rabu	*senin*	0.12687
gadis	*kakek*	0.12345
2007	*2006*	0.11974
sini	*sana*	0.11831
kedua-duanya	*ketiga-tiganya*	0.11383
kakek	*nenek*	0.11295
depan	*belakang*	0.10697
gadis	*nenek*	0.10660
selatan	*utara*	0.10451
mengerikan	*menakutkan*	0.10095
menguat	*melemah*	0.09843
wah	*aduh*	0.09717
maret	*januari*	0.08724

Recapitulation of the eight formats used are shown in Table II, from these results, it appears that the use of 3-gram is better than 4-gram. Average recall of 3-gram is 83.50%, while the average 4-gram recall is 57.25%; the difference between the values is 26.25%. Average precision 3-gram is 95.63%, while the average precision 4-gram is 77.92%; the difference between the values is 17.71%.

TABLE II
Precision and Recall for Varies Format

Feature Format	Input	Output	True	Precision (%)	Recall (%)
t w w'	100	98	98	100.00	98.00
t w' w	100	81	75	92.59	75.00
w w' t	100	78	71	91.03	71.00
w' w t	100	91	90	98.90	90.00
t w r w'	100	79	61	77.22	61.00
t w' r w	100	64	48	75.00	48.00
w r w' t	100	73	51	69.86	51.00
w' r w t	100	77	69	89.61	69.00

Among the four (4) 3-gram formats, which has the best results is the format t w w'. Means for Indonesian, the word similarity algorithm features EWSB is one word after word marker (t) before the word, or in other words, T(w) is defined as the one word before and the and word after word w. Jeff et al. (2011) proposed relation (r) is between w and w'. That format similar to twrw' at our format. Our research indicated that the format has a lower accuracy compared with t w w' format. This is due to English being used by Jeff et al. (2011) have different grammars with Indonesian. This study also concluded that the 3-gram format better than the 4-gram format, because the number of features found in the corpus with 4-gram format is much less than the 3-gram format. This is evident from the average for the 4-gram recall of 57.25 % compared with the average for the 3-gram recall of 83.5 % .

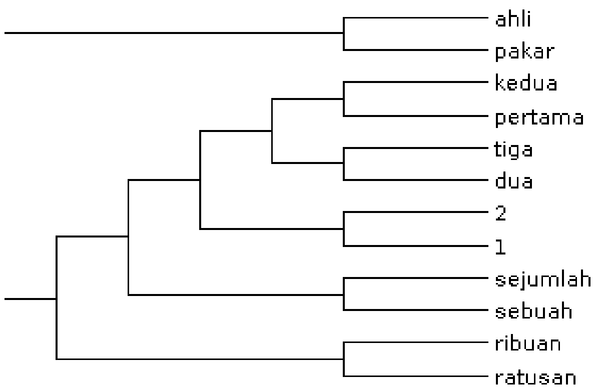

Figure 2. Dendogram of t w w' format for "pakar"

It is interesting to analyze further why t w w' feature better than other features. We observe from word pair ("ahli dan "pakar") computational results have been found using t w w' feature is shown in Figure 2, but that word pair has not been found using t w' w feature is shown in Figure 3.

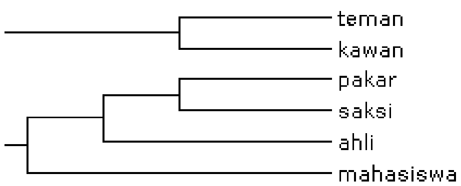

Figure 3. Dendogram of t w' w format for "pakar"

There are 94 features of the word "pakar", 423 features of the word "ahli", and 209 features of the word "saksi" at t w w' format. For example, the features for the word "pakar" are : T(pakar) = {(para,yang) ; (nasehat,independen) ; (banyak,yang); (beberapa,origami) ; … ; (beberapa,teknologi) }. We calculated mutual information for each feature by using equation (3), for example, 3-gram for feature : (para,yang) has 3 words sequence of "para pakar yang", 5.695 words sequence of "para * yang", 28 words sequence of "para pakar *", dan 468 words sequence of "* pakar yang". The value of I(para,pakar,yang) is: log (3 x 5695) / (28 x 468) = 0.26528. In the same way, I(nasehat,pakar,independen) = 2.77259, I (banyak,pakar,yang) = 1.52343, I(beberapa,pakar,origami) = 6.61114, and so on. Mutual Information for each input word (200 words) calculated as applicable to the word "pakar". To compute word similarity between two words, we computed first the intersection between $T(w_1)$ and $T(w_2)$. For example, T(pakar) ∩ T(pakar) with each of its mutual information value are shown in Table III. In comparison, T(pakar) ∩ T(saksi) only has one member as shown in Table IV. We obtained the word similarity by using equation (4), sim (pakar,ahli) = 0.04197, and sim (pakar,saksi) = 0.00335.

TABLE III
MUTUAL INFORMATION T(PAKAR) ∩ T(AHLI) FOR T W W' FORMAT

T(x) = T(pakar) ∩ T(ahli)	I(pakar,T(x))	I(ahli,T(x))
beberapa_teknologi	5.00170666335195	4.59624155524379
beberapa_lainnya	3.90309437468384	3.49762926657568
para_lingkungan	3.36922921665077	2.02860489236406
para_bahasa	3.70570145327198	3.46368941765339
para_telah	1.50847687593576	0.86099973220900
seorang_di	2.49663297010048	0.42766272828794
oleh_ilmu	7.58426481838906	5.18636954559069
para_biologi	3.92884500458619	3.6868329689676
sejumlah_(	2.94312232328169	3.23080439573347
banyak_pemasaran	6.03428454429091	0.41916697992996
para_dari	1.06664412365672	0.41916697992996
seorang_dalam	2.93195104135833	2.06695360387172

TABLE IV
MUTUAL INFORMATION T(PAKAR) ∩ T(SAKSI) FOR T W' W FORMAT

T(x) = T(pakar) ∩ T(saksi)	I(pakar,T(x))	I(saksi,T(x))
dan_</s>	0.4955777673088	0.4955777673088

T(pakar) ∩ T(ahli) for t w' w format with each of its mutual information value are shown in Table V. In comparison, T(pakar) ∩ T(saksi) only has one member as shown in Table IV. We obtained the word similarity by using equation (6), sim (pakar,ahli) = 0.04889, and sim (pakar,saksi) = 0.05139.

TABLE V
MUTUAL INFORMATION T(PAKAR) ∩ T(AHLI) FOR T W' W FORMAT

T(x) = T(pakar) ∩ T(ahli)	I(pakar,T(x))	I(ahli,T(x))
,_para	4.27544976855720	4.36701696208269
,_banyak	4.30071836273208	3.00599119513768
dan_para	4.94089214121860	3.84227985255049
salah_seorang	1.88732642140508	2.29279152951325
dari_para	5.23320351654185	4.87652857260312
dan_banyak	5.84888407027806	3.36397742049006

TABLE VI
MUTUAL INFORMATION T(PAKAR) ∩ T(SAKSI) FOR T W' W FORMAT

T(x) = T(pakar) ∩ T(saksi)	I(pakar,T(x))	I(saksi,T(x))
kata_seorang	2.95001381174319	1.44593641496692
maupun_para	5.29956658594847	5.29956658594847
tidak_ada	2.73383220777187	2.73383220777187
,_namun	2.31890267208080	1.57696532735142
kata_para	4.13516655674236	2.63108915996608
,_seorang	4.04656457467531	3.30462722994594
oleh_semua	3.47612693403462	2.08983257291473

By comparing the results of word similarity : sim (pakar,ahli) and sim (pakar,saksi), We concluded that the use of the t w w' format obtain results sim (pakar,ahli) is greater than the sim (pakar,saksi), whereas the t w' w format obtain results sim (pakar,ahli) is smaller than the sim (pakar,saksi). This is caused by features T (pakar) that intersect with T (ahli) is much more than an intersection of T (pakar) and T (saksi) if using the t w w' format. While using the t w' w format, features T (pakar) that intersect with T (ahli) is relatively the same as the intersection of T (pakar) and T (saksi).

TABLE VII
EXAMPLES OF INDONESIAN FEATURES W=PAKAR

Format	w = pakar
t w w'	*banyak_yang* *para_yang* *sejumlah_,* *mengundang_atau* *kelompok_,* *sekaligus_ilmu* *para_tersebut* *beberapa_origami* *dua_asal* *dan_kontra* *para_bencana* *menurut_ilmu* *beberapa_ekonomi* *beberapa_teknologi* *dari_sex* *para_lain* *atau_)* *banyak_</s>* *sejumlah_(* *para_manajemen* *para_pengobatan* *para_kriptografi* *pertimbangan_(* *para_memperkirakan* *para_</s>* *manurut_,* *kata_dirgantara* *banyak_pemasaran* *para_bahasa*
t w r w'	*para_pakar_punya_banyak* *pada_pakar_telematika_acing* *para_pakar_lain_menyatakan* *dan_pakar_islam_di* *para_pakar_memperkirakan_bahwa* *pertimbangan_pakar_(_expert* *banyak_pakar_yang_menghentikan* *dengan_pakar_,_pencatatan* *seorang_pakar_dalam_sejarah* *para_pakar_pengobatan_alternatif* *para_pakar_botani_mengatakan* *para_pakar_yang_dapat* *dua_pakar_asal_jerman* *menurut_pakar_yang_mengetahui* *dan_pakar_kontra_-* *atau_pakar_,_sesuai* *oleh_pakar_ilmu_hewan* *beberapa_pakar_teknologi_pun* *para_pakar_telah_berhasil* *para_pakar_bencana_alam*

TABLE VIII
EXAMPLES OF INDONESIAN FEATURES W=AHLI

Format	w = ahli
t w w'	*para_mesin* *tenaga_yang* *,_biologi* *para_juga* *seorang_silat* *staf_menteri* *kepada_untuk* *kepada_waris* *banyak_pemasaran* *seorang_paleontologi* *,_kimia* *,_fisika* *dengan_riil* *dan_sejarah* *seorang_biokimia* *para_sering* *ada_waris* *lisensi_perawatan* *oleh_kimia* *perserikatan_mesin* *yang_dalam* *sebagai_pedang* *bagi_dari* *seorang_dalam* *oleh_ilmu* *para_mengatakan* *para_,* *adalah_waris* *kalangan_bahasa*
t w r w'	*para_ahli_menyatakan_bahwa* *dengan_ahli_riil_estate* *staf_ahli_menteri_koordinator* *dan_ahli_sejarah_</s>* *para_ahli_berpendapat_bahwa* *banyak_ahli_pemasaran_yang* *,_ahli_gizi_,* *lisensi_ahli_perawatan_pesawat*

	dari_ahli_bologi_molekul *seorang_ahli_strategi_pasar* *,_ahli_biologi_dinas* *dijadikan_ahli_waris_kakek* *seorang_ahli_etika_michael* *seorang_ahli_dalam_melakukan* *dialah_ahli_warisku_</s>* *para_ahli_mesin_melobi* *sesungguhnya_ahli_dalam_hal* *bukan_ahli_tiam-hiat-hoat_,* *bagi_ahli_silat_umumnya* *adalah_ahli_zoologi_prancis*

TABEL IX
EXAMPLES OF INDONESIAN FEATURES W=SAKSI

Format	w = saksi
t w w'	*beberapa_mata* *dan_mata* *menjadi_</s>* *seorang_mata* *pemeriksa_kawan* *semua_kenal* *para_</s>* *pemeriksaan_pollycarpus* *empat_yang* *para_mata* *pemeriksaan_achirina* *sebagai_dalam* *kedua_mencabut* *juga_sejarah* *keterangan_rahmat* *antara_,* *menemukan_baru* *keterangan_raden* *memeriksa_dan* *untuk_kawan* *sebagai_kunci* *sedangkan_daan* *keterangan_muchtar* *seorang_,* *menjadi_kekejamanmu* *keterangan_indrianto* *keterangan_kawan* *pokoknya_menerangkan* *bahwa_mencabut*
t w r w'	*beberapa_saksi_mata_</s>* *a._saksi_adalah_pemeriksa* *sebagai_saksi_untuk_tersangka* *namun_saksi_baru_tersebut* *menjadi_saksi_mata_dan* *untuk_saksi_kawan_tidak* *menjadi_saksi_ketika_itu* *,_saksi_kembali_mengatakan* *beberapa_saksi_mata_dan* *sebagai_saksi_kunci_kasus* *pemeriksaan_saksi_verbalisan_ni* *,_saksi_kawan_,* *empat_saksi_yang_akan* *kata_saksi_mata_,* *,_saksi_suradi_membenarkan* *dan_saksi_mata_palestina* *para_saksi_yang_berada* *satu_saksi_yang_minta* *pemeriksaan_saksi_ahli_ruby* *pemeriksaan_saksi_dr_tarmizi*

Intersection of T(pakar) and T(ahli) more than intersection of T(pakar) dan T(saksi) for t w w' format because the words of w' are more unique like “teknologi”, “lingkungan”, “bahasa”, “biologi” and “pemasaran” is more related to the "pakar" and "ahli" in comparison to "saksi". While the t w' w format, w' words are more general such as “para”, “seorang”, “banyak”, “ada”, and “namun” that could be associated with the word "pakar", "ahli" or "saksi". Some examples of Indonesian features generated from the corpus are shown in Table VII-IX.

IV. CONCLUSION

We provided four alternative features that can be used for word similarity computation and experimented against the Indonesian Language to determine the best feature format for the Indonesian language. From the results of experiments, the best feature is used in the EWSB algorithm for Indonesian is t w w' format (3-gram) with the relation 0 (zero) word. The number of features found in the corpus with 4-gram format (57.25%) is much less than the 3-gram format (83.50%). This is why a 3-gram format better than the 4-gram format.

The best feature for other languages may be different, of course, it is necessary to do another experiment to determine the features that are suitable for use in a specific language to use the features of the proposed EWSB algorithm.

REFERENCES

[1] D. Jurafsky, dan H.Martin, “Speech and language processing”, Parson International Edition, New Jersey, 2009.

[2] S. Ker and J. Zhang, “A Class-based Approach to Word Alignment”, in Computational Linguistics, Vol. 23, No. 2, pp 313-343, 1997.

[3] Z. Harris, “Distributional structure”, Word, pages 146–142, 1954.

[4] D. Hindle, “Noun classification from predicate-argument structures”, In Proceedings of ACL-90, pages 268–275, 1990.

[5] G. Grefenstette, “Explorations In Automatic Thesaurus Discovery”, Kluwer Academic Publishers, 1994.

[6] I. Dagan, F. Pereira, and L. Lee. “Similarity-Based Estimation of Word Cooccurrence Probabilities”, In Proceedings of ACL 94, 1994.

[7] I. Dagan, S. Marcus, and S. Markovitch. “Contextual Word Similarity and Estimation From Sparse Data”, Computer, Speech and Language, 9:123–152, 1995.

[8] D. Lin, “Automatic Retrieval and Clustering of Similar Words”, Proceedings of the 17th international conference on computational linguistics. Vol. 2. Canada, 1998.

[9] I. Dagan, L. Lee, and F. Pereira, “Similarity-based models of word cooccurrence probabilities. Machine Learning”, 34(1-3):43–69, 1999.

[10] J. Bellegarda, J.W. Butzberger, Y.L. Chow, B.C. Noah, D. Naik, “A Novel Word Clustering Algorithm Based on Latent Semantic Analysis”, in Proceedings of ACSSAP 1996, Atlanta, USA, 1996.

[11] L. Lee, “Measures of Distributional Similarity”, In Proceeding of the 37th Annual Meeting of the ACL, pages 25–32, 1999.

[12] M. Geffet and I. Dagan, “Feature Vector Quality and Distributional Similarity”, Proceedings Of the 20th International Conference on Computational Linguistics, 2004.

[13] J. Weeds and D. Weir, “Co-oocurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity”, in Computational Linguistics, 31(4):439–476, 2005.

[14] P. Muller, N. Hathout, and B. Gaume, “Synonym Extraction Using a Semantic Distance on a Dictionary”, in Proceedings of the Workshop on TextGraphs on Graph-based Algorithms for Natural Language Processing, 2006.

[15] R. Sinha and R. Mihalcea, “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”, in Proceedings of the IEEE International Conference on Semantic Computing, CA, USA, 2007.

[16] K. Ichioka and F. Fukmoto, “Graph based Clustering for Semantic Classification of Onomatopoetic Words”, in Proceedings of the 3rd Text graphs Workshop on Graph-based Algorithms for Natural Language Processing, Manchester, UK, 2008.

[17] M.A. Jeff, S. Matsoukas, S.R. Schwartz, “Improving Low-Resource Statistical Machine Translation with a Novel Semantic Word Clustering Algorithm”, Proceedings of the MT Summit XIII, Xiamen, China, 2011.

[18] H. Sujaini, Kuspriyanto, A.A. Arman, and A. Purwarianti, “Extended Word Similarity Based Clustering on Unsupervised PoS Induction to Improve English-Indonesian Statistical Machine Translation”, 16th ORIENTAL COCOSDA/CASLRE-2013, Gurgaon, India, 2013.

[19] M. Geffet and I. Dagan. “The Distributional Inclusion Hypothesises and Lexical Entailment”, In Proceedings Of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 107–114, 2005.

[20] F. Joseph. “Inferring Phylogenies”, Sinauer Associates, Inc., Sunderland, Mass, 2004.