Comparing the value of Latent Semantic Analysis on two English-to-Indonesian lexical mapping tasks

David Moeljadi

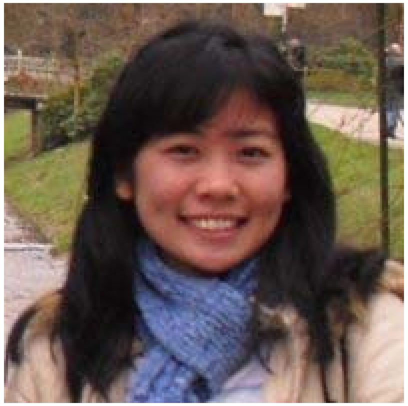
Nanyang Technological University
October 16, 2014

Outline

- The Authors
- The Experiments (general idea and results)
- The Details
  - Concept and word
  - Bilingual word mapping
  - Bilingual concept mapping
- Results and Discussion

The Authors

Eliza Margaretha



Wissenschaftliche Angestellte
(Research Staff)
at Institut für Deutsche Sprache

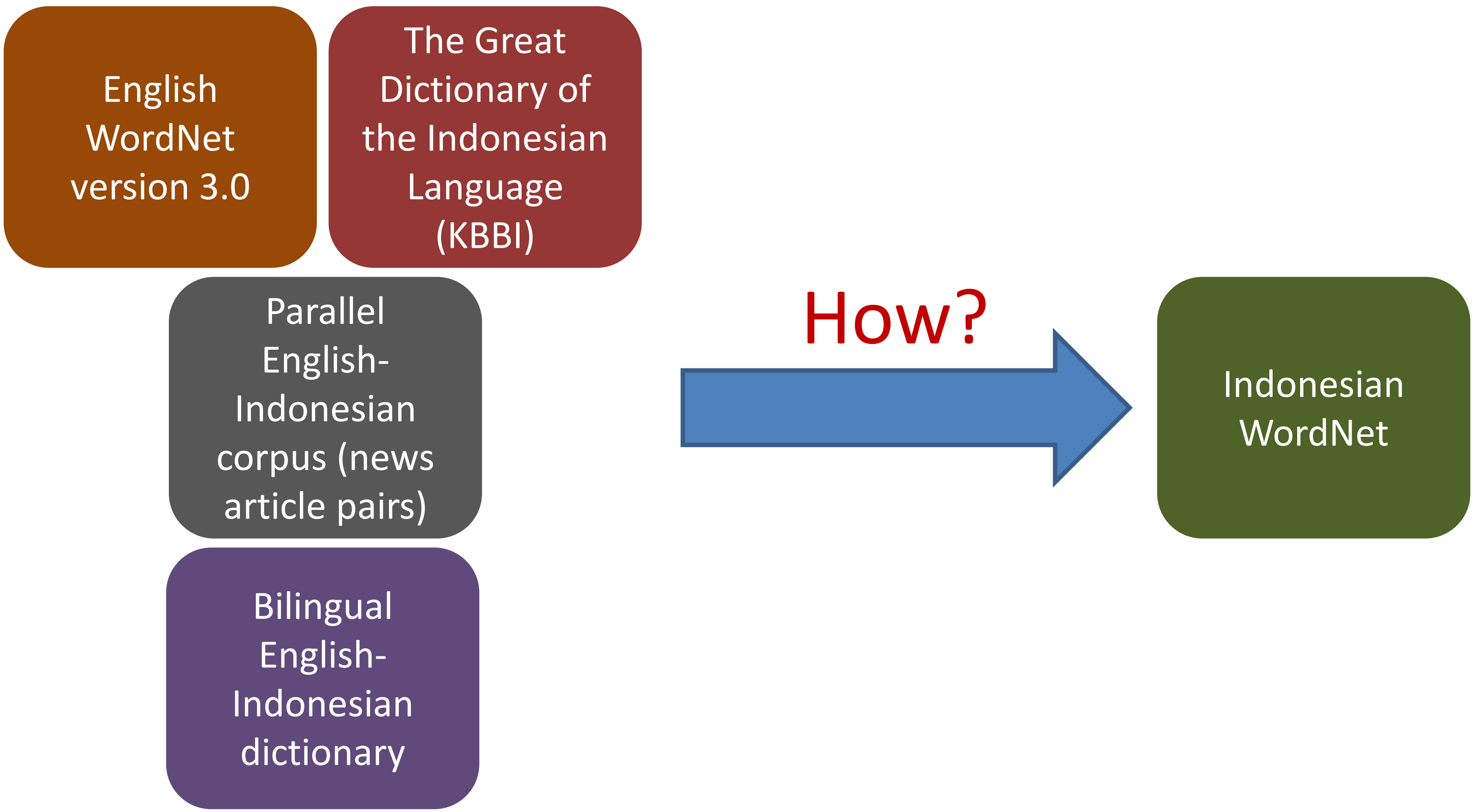
Ruli Manurung

Coordinator of Computer Science Dept.
at [Faculty of Computer Science](),
University of Indonesia

Eliza Margaretha's undergraduate theses supervised by Ruli Manurung

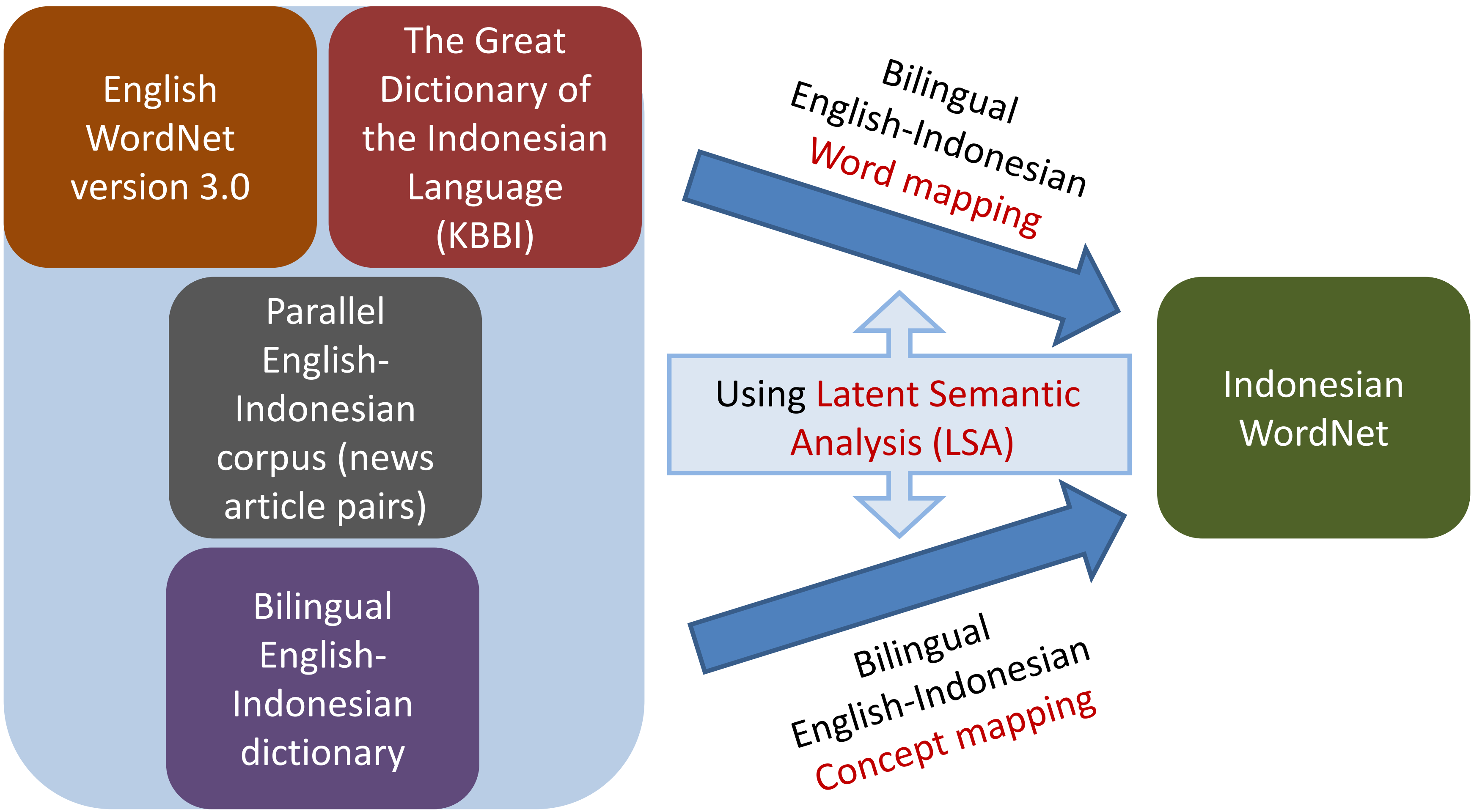
The Experiments

- General Idea -

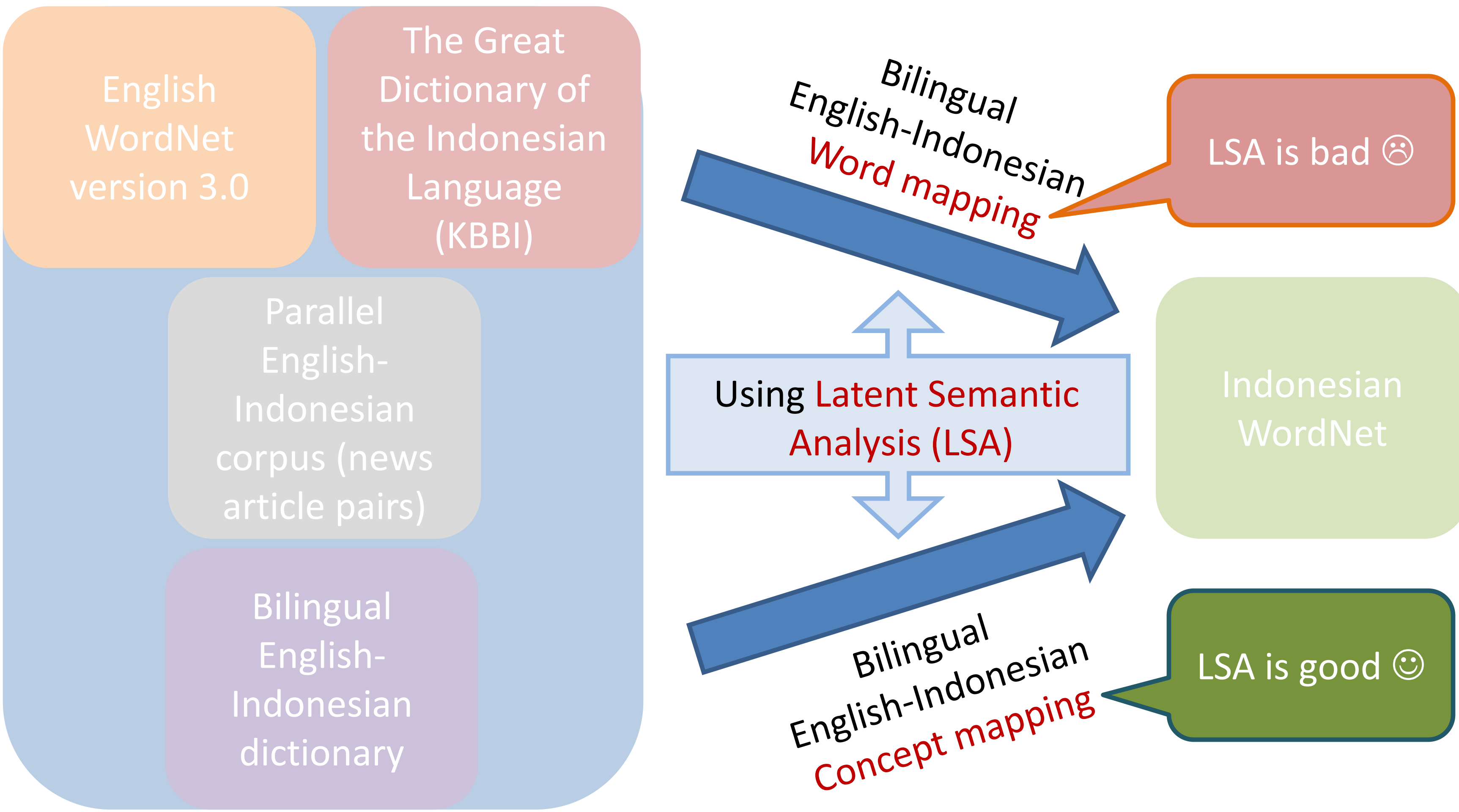


The Experiments

- General Idea -



The Experiments
- Results -



Concept and Word

Language	Concept	Word
Indonesian	00464894-n	golf
English	0842027 09213565-n	bank

00464894-n ✪ 'a game played on a large open course with 9 or 18 holes';

Bahasa Indonesia	*golf*

08420278-n (20) bank, depository financial institution, banking concern, banking company

a financial institution that accepts deposits and channels the money into lending activities

binatang

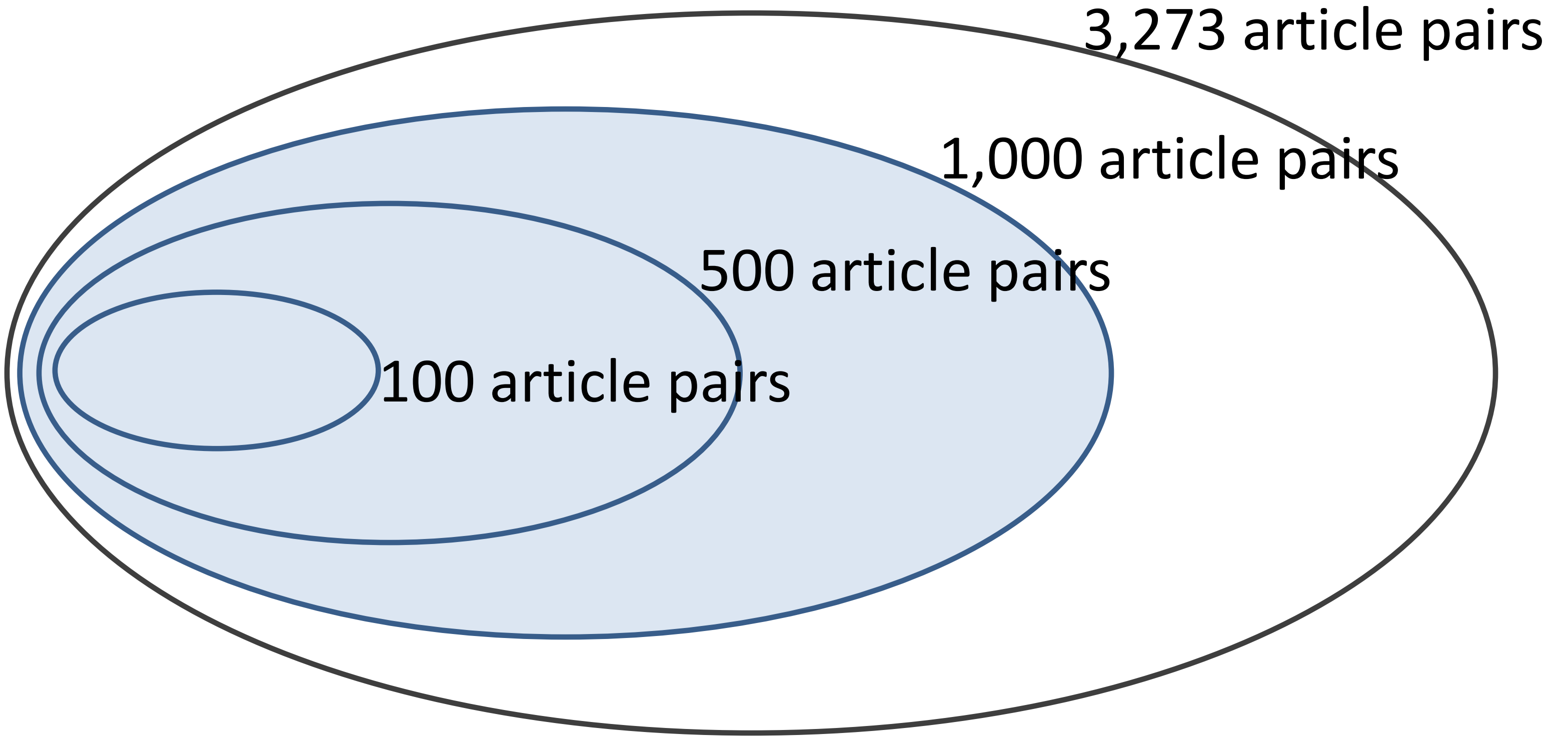
water)

00015388-n ✪ 'a living organism characterized by voluntary movement';

Indonesian	*animal , hewan , sato , manusia , margasatwa , fauna , binatang*

Bilingual Word Mapping
- The Corpus -

1. Define a collection of parallel article pairs



Bilingual Word Mapping
- Latent Semantic Analysis -

2. Set up a bilingual word-document matrix for LSA

ENG	Article 1E	Article 2E	…	Article 100E
dog	5	0	…	0
the	10	15	…	50
car	4	0	…	7
…	…	…	…	…

IND	Article 1I	Article 2I	…	Article 100I
anjing	5	0	…	0
itu	12	10	…	30
mobil	3	0	…	10
…	…	…	…	…

Each column is a pair of parallel articles

Bilingual Word Mapping
- Latent Semantic Analysis -

2. Set up a bilingual word-document matrix for LSA

$$M_E \begin{bmatrix} 5 & 0 & \dots & 0 \\ 10 & 15 & \dots & 50 \\ 4 & 0 & \dots & 7 \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

$$M_I \begin{bmatrix} 5 & 0 & \dots & 0 \\ 12 & 10 & \dots & 30 \\ 3 & 0 & \dots & 10 \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

Bilingual Word Mapping
- Latent Semantic Analysis -

2. Set up a bilingual word-document matrix for LSA

For each of these rows, compute the similarity

to each of these rows

$$
M = \begin{bmatrix}
\boxed{5 \quad 0 \quad \dots \quad 0} \\
10 \quad 15 \quad \dots \quad 50 \\
4 \quad 0 \quad \dots \quad 7 \\
\dots \quad \dots \quad \dots \quad \dots \\
\\
\boxed{5 \quad 0 \quad \dots \quad 0} \\
\boxed{12 \quad 10 \quad \dots \quad 30} \\
\boxed{3 \quad 0 \quad \dots \quad 10} \\
\boxed{\dots \quad \dots \quad \dots \quad \dots}
\end{bmatrix}
$$

Bilingual Word Mapping
- Latent Semantic Analysis -

2. Set up a bilingual word-document matrix for LSA

However, there are irrelevant information and noise need to be removed

$$M \begin{bmatrix} 5 & 0 & \cdots & 0 \\ 10 & 15 & \cdots & 50 \\ 4 & 0 & \cdots & 7 \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ 5 & 0 & \cdots & 0 \\ 12 & 10 & \cdots & 30 \\ 3 & 0 & \cdots & 10 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

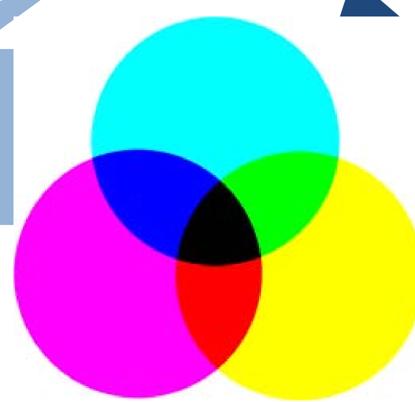
Bilingual Word Mapping
- Latent Semantic Analysis -

3. LSA: Compute SVD (Singular Value Decomposition)

$$M \quad = \quad U \quad S \quad V^T$$

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Matrix of left singular vectors



e
es

Matrix of right singular vectors

Bilingual Word Mapping
- Latent Semantic Analysis -

3. LSA: Compute SVD (Singular Value Decomposition)

singular value decomposition tutorial

ウェブ　動画　画像　ニュース　もっと見る　検索ツール

Highly recommended if you want to know more! (especially for beginners)

[PDF] **Singular Value Decomposition Tutorial PDF**

www.ling.ohio-state.edu/.../**Singular_Value_Decompos**... ▾ このページを訳す

K Baker 著 - 引用元 48 - 関連記事

Most **tutorials** on complex topics are apparently written by very smart people whose goal is to use as little space as possible and ... It's about the mechanics of **singular value decomposition**, especially as it relates to some techniques in natural ...

14/10/01 にこのページにアクセスしました。

Bilingual Word Mapping
- Latent Semantic Analysis -

4. Compute the optimal reduced rank approximation (reducing dimensions of the matrix)

- unearth implicit patterns of semantic concepts
- the vectors representing English and Indonesian words that are closely related should have high similarity

	10%	25%	50%	100% (no reduction)
100 art.pairs	10	25	50	100
500 art.pairs	50	125	250	500
1000 art.pairs	100	250	500	1,000

Bilingual Word Mapping

- Latent Semantic Analysis -

4. Words are represented by row vectors in *U*, word similarity can be measured by computing row similarity in *US*.

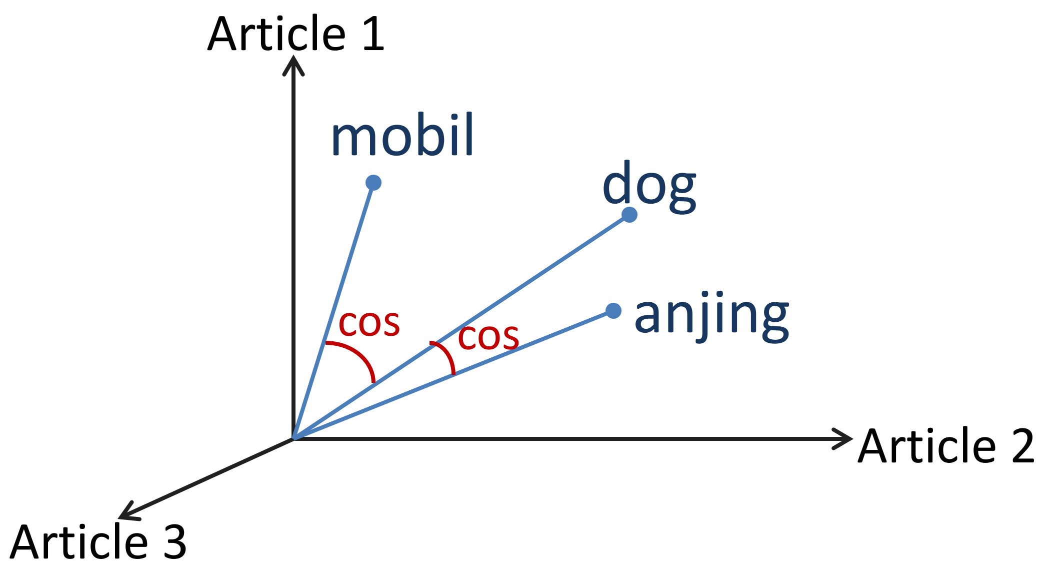
$$M = U \quad S \quad V^T$$

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Bilingual Word Mapping

- Latent Semantic Analysis -

5. For a randomly chosen set of vectors representing English words, compute the *n* nearest vectors representing the *n* most similar Indonesian words using the cosine of the angle between two vectors



Bilingual Word Mapping
- Some Experiments -

6. Remove the stopwords from the matrix
   English: the, a, of, in, by, for, …
   Indonesian: itu, sebuah, dari, di, oleh, untuk, …
   and do SVD again.

7. Apply two weighting schemes:
   - TF-IDF
   - Log-entropy
   and do SVD again.

Bilingual Word Mapping
- Some Experiments -

7. Apply TF-IDF

- term frequency-inverse document frequency
- TF: to measure how frequently a word occurs in a document

$$\frac{\text{Number of word } w \text{ in a document}}{\text{Total number of words in a document}}$$

- IDF: to measure how important a word is in a corpus

$$\log \frac{\text{Total number of documents}}{\text{Number of documents with word } w \text{ in it}}$$

- can be used for stopwords filtering

Bilingual Word Mapping
- Some Experiments -

7. Apply TF-IDF (example)

	Article 1	Article 2	…	Article 100
dog	5	0	…	0
the	10	15	…	50
car	4	0	…	7
…	…	…	…	…
Total	100	150	…	125

TF x IDF

$$\frac{\text{Number of word } w \text{ in a document}}{\text{Total number of words in a document}} \times \log \frac{\text{Total number of documents}}{\text{Number of documents with word } w \text{ in it}}$$

Bilingual Word Mapping
- Some Experiments -

7. Apply TF-IDF (example)

	Article 1	Article 2	…	Article 100
dog	5	0	…	0
the	10	15	…	50
car	4	0	…	7
…	…	…	…	…
Total	100	150	…	125

TF IDF of *dog*

$$\frac{5}{100} \times \log \frac{100}{1} = 0.05 \times \log 100 = 0.05 \times 2 = \boxed{0.1}$$

Bilingual Word Mapping
- Some Experiments -

7. Apply TF-IDF (example)

	Article 1	Article 2	…	Article 100
dog	5	0	…	0
the	10	15	…	50
car	4	0	…	7
…	…	…	…	…
Total	100	150	…	125

TF-IDF of *the* in article 1 $\frac{10}{100} \times \log\frac{100}{100} = 0.1 \times \log 1 = 0.1 \times 0 = \boxed{0}$

TF-IDF of *car* in article 1 $\frac{4}{100} \times \log\frac{100}{2} = 0.04 \times \log 50 = 0.04 \times 1.7 = \boxed{0.07}$

TF-IDF of *car* in article 100 $\frac{7}{125} \times \log\frac{100}{2} = 0.06 \times \log 50 = 0.06 \times 1.7 = \boxed{0.09}$

Bilingual Word Mapping
- Some Experiments -

7. Apply TF-IDF and do SVD (example)

	Article 1	Article 2	…	Article 100
dog	0.10	0.00	…	0.00
~~the~~	~~0.00~~	~~0.00~~	…	~~0.00~~
car	0.07	0.00	…	0.09
…	…	…	…	…

Stopwords filtering

Bilingual Word Mapping

- Some Experiments -

7. Apply TF-IDF and do SVD (example)

$$M = \begin{bmatrix} 0.10 & 0.00 & \dots & 0.00 \\ 0.07 & 0.00 & \dots & 0.09 \\ \dots & \dots & \dots & \dots \\ & & & \end{bmatrix}$$

$$\underset{\begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}}{M} = \underset{\begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}}{U} \underset{\begin{bmatrix} \dots & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 \end{bmatrix}}{S} \underset{\begin{bmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}}{V^T}$$

Bilingual Word Mapping

- Some Experiments -

7. Apply Log-entropy and do SVD

log = $\log(\mathrm{tf}_{ij} + 1)$

entropy = $1 - \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$, where $p_{ij} = \frac{\mathrm{tf}_{ij}}{\mathrm{gf}_i}$

gf_i is the total number of times a word appears in a corpus,
n is the number of documents in a corpus

After getting a new matrix from log-entropy, do SVD
(same as in TF-IDF)

Bilingual Word Mapping

- Some Experiments -

8. Do mapping selection

Take the top 1, 10, 50, and 100 mappings based on similarity

film	0.814
filmnya	0.698
sutradara	0.684
garapan	0.581
perfil	54
pena	44
kontroversial	0.526
koboi	0.482
irasional	0.482
frase	0.482

(a)

GOOD

pembebanan	0.973
kijang	0.973
halmahera	0.973
alumina	0.973
terjadw	.973
viskosit	.973
tabel	0.973
royalti	0.973
reklamasi	0.973
penyimpan	0.973

(b)

BAD

The Most 10 Similar Indonesian Words for the English Words (a) Film and (b) Billion using 1,000 article pairs with 500-rank approximation and no weighting

- *billion* is not domain specific
- *billion* can sometimes be translated numerically instead of lexically
- lack of data: the collection is too small

Bilingual Word Mapping

- Some Experiments -

9. Compute the precision and recall values for all experiments

$$P = \frac{\Sigma \text{ correct mappings (check with bilingual dictionary)}}{\Sigma \text{ total mappings found}}$$

$$R = \frac{\Sigma \text{ correct mappings (check with bilingual dictionary)}}{\Sigma \text{ total mappings in bilingual dictionary}}$$

Bilingual Word Mapping
- The Results -

1. As the collection size increases, the precision and recall values also increase

Collection Size	FREQ		LSA	
	P	R	P	R
P_{100}	0.0668	0.1840	0.0346	0.1053
P_{500}	0.1301	0.2761	0.0974	0.2368
P_{1000}	**0.1467**	**0.2857**	0.1172	0.2603

2. The higher the rank approximation percentage, the better the mapping results

Rank Approximation	P	R
10%	0.0680	0.1727
25%	0.0845	0.2070
50%	0.0967	0.2226
100%	**0.1009**	**0.2285**

Bilingual Word Mapping
- The Results -

3. On account of the small size of the collection, stopwords may carry some semantic information

Stopwords	FREQ		LSA	
	P	R	P	R
Contained	0.1108	**0.2465**	0.0840	0.2051
Removed	**0.1138**	0.2440	0.0822	0.1964

4. Weighting can improve the mappings (esp. Log-entropy)

Weighting Usage	FREQ		LSA	
	P	R	P	R
No Weighting	0.1009	0.2285	0.0757	0.1948
Log-Entropy	**0.1347**	**0.2753**	0.1041	0.2274
TF-IDF	0.1013	0.2319	0.0694	0.1802

Bilingual Word Mapping
- The Results -

5. As the number of translation pairs selected increases, the precision value decreases and the possibility to find more pairs matching the pairs in bilingual dictionary (the recall value) increases

Mapping Selection	FREQ		LSA	
	P	R	P	R
Top 1	**0.3758**	0.1588	0.2380	0.0987
Top 10	0.0567	0.2263	0.0434	0.1733
Top 50	0.0163	0.2911	0.0133	0.2338
Top 100	0.0094	**0.3183**	0.0081	0.2732

Conclusion: FREQ baseline (basic vector space model) is better than LSA

Bilingual Concept Mapping

- Semantic Vectors for Concepts -

1. Construct a set of textual context representing a concept *c* by including (1) the sublemma words, (2) the gloss words, and (3) the example sentence words, which appear in the corpus.

WordNet Synset ID: 100319939, **Words**: chase, following, pursual, pursuit, **Gloss**: the act of pursuing in an effort to over-take or capture, **Example**: the culprit started to run and the cop took off in pursuit, **Textual context set**: {{following, chase}, {the, effort, of, to, or, capture, in, act, pursuing, an}, {the, off, took, to, run, in, culprit, started, and}}

Bilingual Concept Mapping
- Semantic Vectors for Concepts -

1. Construct a set of textual context representing a concept *c* by including (1) the sublemma words, (2) the definition words, and (3) the example sentence words, which appear in the corpus.

KBBI ID: k39607 - **Similarity**: 0.804, **Sublemma**: menge-jar, **Definition**: berlari untuk menyusul menangkap dsb mem-buru, **Example**: ia berusaha mengejar dan menangkap saya, **Textual context set**: {{mengejar}, {memburu, berlari, me-nangkap, untuk, menyusul},{berusaha, dan, ia, mengejar, saya, menangkap}}

Bilingual Concept Mapping
- Semantic Vectors for Concepts -

2. Compute the semantic vector of a concept, that is a weighted average of the semantic vectors of the words in the set

Textual context set: {{following, chase},
{the, effort, of, to, or, capture, in, act, pursuing, an},
{the, off, took, to, run, in, culprit, started, and}}

Sublemma 60%
Gloss 30%
Example 10%

Textual context set: {{mengejar},
{memburu, berlari, menangkap, untuk, menyusul},
{berusaha, dan, ia, mengejar, saya, menangkap}}

Sublemma 60%
Definition 30%
Example 10%

Bilingual Concept Mapping
- Latent Semantic Analysis -

3. Use 1,000 article pairs and set up a bilingual concept-document matrix for LSA

ENG	Article 1E	…	Article 1000E
100319939	…	…	…
201277784	…	…	…
…	…	…	…

IND	Article 1I	…	Article 1000I
k39607	…	…	…
k02421	…	…	…
…	…	…	…

Bilingual Concept Mapping
- Latent Semantic Analysis -

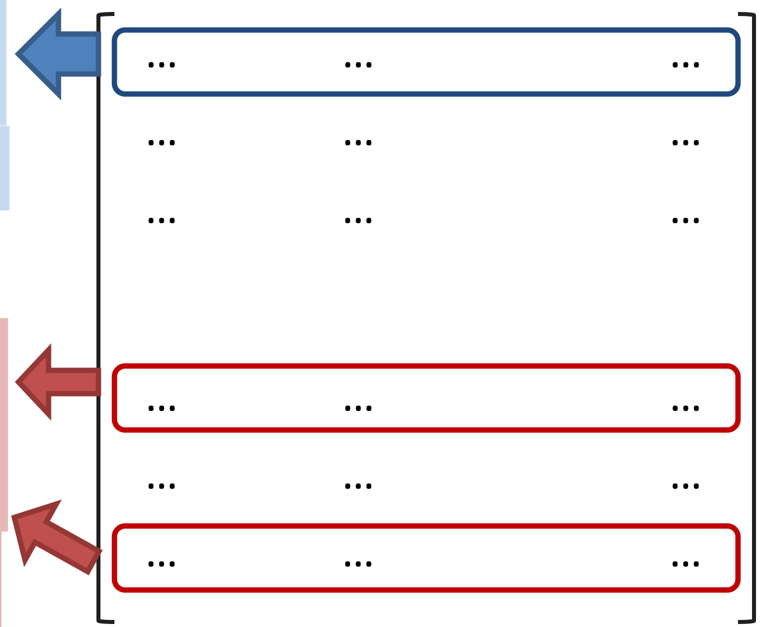
3. Set up a bilingual concept-document matrix for LSA

Given a WordNet synset,
look up in bilingual dictionary
the Indonesian words

e.g. for synset *communication*

select the most appropriate
KBBI sense
from a subset of senses

compare it with *komunikasi*
and *perhubungan* only



Bilingual Concept Mapping

- Latent Semantic Analysis -

4. LSA: Compute SVD (Singular Value Decomposition)

$$M = U \quad S \quad V^T$$

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \quad \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \begin{bmatrix} \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Matrix of left singular vectors

Matrix containing the singular values of M

Matrix of right singular vectors

Bilingual Concept Mapping
- Latent Semantic Analysis -

5. Compute the optimal reduced rank approximation (reducing dimensions of the matrix)

	10%	25%	50%
1,000 art. pairs	100	250	500

6. Compute the level of agreement between the LSA-based mappings with human annotations (ongoing experiment to manually map WordNet synsets to KBBI senses)

Bilingual Concept Mapping

- Check the results -

7. As a baseline, select three random suggested Indonesian word senses as a mapping for an English word sense

8. As another baseline, compare English concepts to their suggestion based on a full rank word-document matrix

9. Choose top 3 Indonesian concepts with the highest similarity values as the mapping results

Bilingual Concept Mapping
- Results -

10. Compute the Fleiss kappa values

Judges	Synsets	Fleiss Kappa Values					
		Judges only	Judges + RNDM3	Judges + FREQ Top 3	Judges + LSA 10% Top3	Judges + LSA 25% Top3	Judges + LSA 50% Top3
≥ 2	144	0.4269	0.1318	0.1667	0.1544	0.1606	0.1620
≥ 3	24	0.4651	0.2197	0.2282	0.2334	0.2239	0.2185
≥ 4	8	0.5765	0.3103	0.2282	0.3615	0.3329	0.3329
≥ 5	4	0.4639	0.2900	0.2297	0.3359	0.3359	0.3359
Average		0.4831	0.2380	0.2132	0.2713	0.2633	0.2623

Results of Conc

Rank Approximation	P	R
10%	0.0680	0.1727
25%	0.0845	0.2070
50%	0.0967	0.2226
100%	**0.1009**	**0.2285**

- LSA 10% is better than the randor frequency baseline (FREQ)
- LSA 10% is better than LSA 25% and LSA 50% (cf. the word mapping results)

Bilingual Concept Mapping

- Mapping results -

WordNet Synset ID: 100319939, **Words**: chase, following, pursual, pursuit, **Gloss**: the act of pursuing in an effort to overtake or capture, **Example**: the culprit started to run and the cop took off in pursuit, **Textual context set**: {{following, chase}, {the, effort, of, to, or, capture, in, act, pursuing, an}, {the, off, took, to, run, in, culprit, started, and}}

KBBI ID: k39607 GOOD 4, **Sublemma**: mengejar, **Definition**: be l menangkap dsb memburu, **Example**: ia dan menangkap saya, **Textual context set**: {{mengejar}, {memburu, berlari, menangkap, untuk, menyusul},{berusaha, dan, ia, mengejar, saya, menangkap}}

(a)

The textual context sets both are fairly large -> provide sufficient context for LSA to choose the correct KBBI sense

WordNet synset ID: 201277784, **Words**: crease, furrow, wrinkle
Gloss: make wrinkled or creased, **Example**: furrow one's brow,
Textual context set: {{}, {or, make}, {s, one}}

KBBI ID: k02421 - Sim BAD **Sublemma**: alur, **Definition**: jalinan peristiwa a untuk mencapai efek tertentu pautannya dapa oleh hubungan temporal atau waktu dan oleh hubungan kausal atau sebab-akibat, **Example**: (none), **Textual context set**: {{alur}, {oleh, dan, atau, jalinan, peristiwa, diwujudkan, efek, dapat, karya, hubungan, waktu, mencapai, untuk, tertentu}, {}}

(b)

The textual context set for the synset is very small -> no sufficient context for LSA to choose the correct KBBI sense

Discussion

- Initial intuition:
  LSA is good for both word and concept mappings
- Results:
  1. LSA blurs the co-occurrence information/details
     -> bad for word mapping
  2. LSA is useful for revealing implicit semantic patterns
     -> good for concept mapping
- Reasons:
  - The rank reduction in LSA perhaps blurs some details
  - A problem of polysemous words for LSA
- Suggestion:
  Make a finer granularity of alignment (e.g. at a sentential level) for word mapping

☺ Special thanks to Giulia and Yukun ☺