channels the money into lending activities" and c_2 is associated with the gloss *"sloping land (especially the slope beside the body of water)"*.

- A term $t \in T$ is an orthographic form which represents a term, namely a word or a phrase, in a particular language. In the case of Princeton WordNet, the language is English. For example, term t_1 and t_2 are defined, where t_1 represents orthographic form *bank* and t_2 represents orthographic form *bank note*.
- A term may convey several distinct concepts. The function $\chi: T \to P(C)$ returns all concepts which can be conveyed by a particular term. Thus, $\chi(t)$, where $t \in T$, returns $C_t \subset C$, the set of all concepts which can be which can be conveyed by t. According to the examples above, $\chi(t_1) = \{c_1, c_2\}$.
- Conversely, a concept may be conveyed by several terms. The function $\omega: C \to P(T)$ returns all terms that can convey a particular concept. Thus, $\omega(c)$, where $c \in C$, returns $T_c \subset T$, the set of all terms that can convey c. According to the examples above, $\omega(c_1) = \omega(c_2) = \{t_1\}$. Essentially, T_c is a *synset* where each term of T_c is regarded as a synonym of the other terms.

A WordNet of any language can be generalised using this 4-tuple. If a WordNet for language x is represented as N^x, then an English WordNet is represented as $N^e = (C^e, T^e, \chi^e, \omega^e)$ and an Indonesian WordNet is represented as $N^i = \left(C^i, T^i, \chi^i, \omega^i\right)$. Moreover, the notation t_i^x is used to denote term i in T^x and c_j^x to denote concept j in C^x.

Although KBBI was never conceived as an Indonesian WordNet, it can be seen as a very simple Indonesian WordNet N^e with an unusual behaviour with respect to the 4-tuple definition. In KBBI, each concept c^i is solely associated with a specific term t^i so that t^i does not own any synonym. In other words, the *synset* for c^i has not been established yet. Accordingly, the work in this thesis attempts to provide the *synsets* for Indonesian concepts.

If concepts are assumed to be language independent, then N^e and N^i should share the same set of universal concepts. In practise, however, WordNets in different languages may have different conceptual representations with respect to the difference of the degree of granularity among the languages. Therefore, most likely, there are some distinctions between C^e and C^i. Explicitly, the relation $E: C^e \times C^i$ is defined to represent the mapping of equivalent concept in C^e and C^i.

3.1.1 Task of Bilingual Term Mapping

Although the primary objective of this work is to map English and Indonesian concepts, the task of bilingual term mapping was conceived to verify whether LSA can map English terms to their Indonesian analogues. Given $t_x^e \in T^e$, the task of bilingual term mapping is to find the set of all its plausible translation terms in T^i, regardless of the concepts being conveyed by t_x^e. Hence, it yields the union of the sets of all terms in T^i that convey $C_{t_x^e}$, the set of all concepts conveyed by t_x^e.

More formally, the set is $\{t_y^i : t_y^i \in \omega^i(c^i)$ where $(c^e, c^i) \in E$ and $c^e \in \chi^e(t_x^e)\}$. For example, given t_{time}^e, i.e. the English orthographic form $time$, $\chi^e(w_{time}^e)$ returns more than fifteen distinct concepts in Princeton WordNet, among others $\{c_1^e, c_2^e, c_3^e\}$ (see Table 3.1).

Table 3.1 Sample of Concepts and Terms in N^e and N^i

Concept	**Term**	**Gloss**	**Example**
c_1^e	t_{time}^e	an instance or single occasion for some event	"this **time** he succeeded"
c_2^e	t_{time}^e	a suitable moment	"it is **time** to go"
c_3^e	t_{time}^e	a reading of a point in time as given by a clock	"do you know what **time** it is?"
c_1^i	t_{kali}^i	kata untuk menyatakan kekerapan tindakan	"dalam satu minggu ini, dia sudah empat **kali** datang ke rumahku"
c_2^i	t_{kali}^i	kata untuk menyatakan salah satu waktu terjadinya peristiwa yg merupakan bagian dari rangkaian peristiwa yg pernah dan masih akan terus terjadi	"untuk **kali** ini ia kena batunya"
c_3^i	t_{waktu}^i	saat yg tertentu untuk melakukan sesuatu	"**waktu** makan"
c_4^i	t_{jam}^i	saat tertentu, pada arloji jarumnya yg pendek menunjuk angka tertentu dan jarum panjang menunjuk angka 12	"ia bangun **jam** lima pagi"
c_5^i	t_{kali}^i	sebuah sungai yang kecil	"air di **kali** itu sangat keruh"

Assuming the relation E, c_1^e represent similar concepts with c_1^i and c_2^i. Next, c_2^e and c_3^e represent similar concept with c_3^i and c_4^i, respectively. In Indonesian, $\omega^i(c_1^i)$ and $\omega^i(c_2^i)$ returns terms that convey c_1^i and c_2^i, that is t_{kali}^i. On the other hand, $\omega^i(c_3^i)$ returns t_{waktu}^i and $\omega^i(c_4^i)$ returns t_{jam}^i.

The task of bilingual term mapping is supposed to map English term t^e_{time} to the set of Indonesian terms $\{t^i_{kali}, t^i_{waktu}, t^i_{jam}, \ldots\}$. Notice that each of these Indonesian terms may convey distinct concepts, which have no relationship with the concepts conveyed by the Engish term. For example, $\chi^i(t^i_{kali})$ also return c^i_5 which has no relationship with the concepts conveyed by t^e_{time}.

3.1.2 Task of Bilingual Concept Mapping

According to the research objectives, the task of bilingual concept mapping is to map English concepts to their equivalent Indonesian concepts (see Section 1.3). Essentially, the task of bilingual concept mapping is to establish the relation $E: C^e \times C^i$. For example, given English and Indonesian concepts in Table 3.1, bilingual concept mapping attempts to map English concepts $\{c^e_1, c^e_2, c^e_3\}$ to the equivalent Indonesian concepts. Eventually, it should establish the set of English-Indonesian equivalent concept pairs $E = \{(c^e_1, c^i_1), (c^e_1, c^i_2), (c^e_2, c^i_3), (c^e_3, c^e_4)\}$.

Since each concept $c \in C$ represents distinct semantic entities, E should define a one-to-one relation. However, WordNets of different languages are likely to have different degree of concept granularity in analysis of polysemy. As a consequence, a concept in a WordNet may map to more than one concept in another WordNet of a different language.

For example, Princeton WordNet, as a repository of English concepts, has a concept "*the food served*" conveyed by the terms *meal* and *repast*, a concept "*a particular item of prepared food*" conveyed by the term *dish*, and a concept "*part of a meal served at one time*" conveyed by the term *course*. Using KBBI as the repositiory of Indonesian concepts, all these English concepts are mapped to a single Indonesian concept "*makanan yang dihidangkan*" conveyed by the term *hidangan.*

On the contrary, an English concept "*a light informal meal*" of the term *snack* is mapped to several distinct concepts conveyed by each of Indonesian terms *makanan ringan*, *makanan kecil*, and *kudapan*. Although these terms are synonyms, they convey distinct concepts, which basically have the same meaning. KBBI does not explicitly include them in a *synset.* Rather, KBBI associates each of these terms to a gloss which is very similar to each other. This fact may suggest that KBBI has a finer granularity than Princeton WordNet which has grouped the terms conveying the same concept in a *synset.*

Given an English concept, the task of bilingual concept mapping is then to find all Indonesian equivalent concepts and subsequently construct a *synset* consisting of the terms conveying all those concepts.

3.2 How LSA is Applied

LSA is a method which is capable of extracting and representing contextual-usage meaning of terms. Most likely, LSA is a potential method to automate human tasks related to understanding the meaning of terms and documents. Many successful applications of LSA suggest that LSA is able to match literate humans in doing those tasks (see Section 2.3.4).

Without human interventions, LSA is able to acquire knowledge of meaning of terms and documents by exploiting text of natural human language, typically a large corpus of text. As LSA analyses the corpus, which contains terms reflecting human knowledge, it derives knowledge of meaning of a term as a kind of average of meaning of all documents in which it appears. Conversely, LSA derives knowledge of meaning of a document as a kind of average of meaning of all terms it contains.

On the basis of its powerful mathematical foundation, LSA examines the similarity of contexts in which terms appears and creates a reduced-dimension space representation which discovers the important underlying relationships among terms of similar contextual-usage meaning. The reduced-dimension space reflects the relationships by locating the terms appearing in similar context near each other.

This special ability of LSA to automatically capture the similarity of contextual-usage meaning of term can be used to accomplish the automatic mapping. By applying SVD to an English-Indonesian bilingual term-document matrix and subsequently reconstructing the matrix with reduced dimension, English terms and Indonesian terms which are closely related should be located near each other in the reduced-dimension space. Moreover, terms which are translations to each other should be near each other. In other words, they should have high similarity. Therefore, LSA reduced-dimension space representation can be used to perform bilingual term mapping. Furthermore, LSA can be applied to perform bilingual concept mapping by specifying conceptual semantic vectors to be mapped.

3.2.1 Building Bilingual Term-Document Matrix

First of all, LSA is applied on a term-document matrix, i.e. a matrix containing the frequency of each unique term in each document (see Section 2.3.3). Each row of a term-document matrix represents each unique term appearing in a corpus, whereas each column represents each document in the corpus.

A parallel corpus P is defined as a set of document pairs $p_j = (d_j^e, d_j^i)$ where d^e denotes an English document and d^i denotes the Indonesian translation of that document. The two words t_x^e dan t_y^i are expected to appear consistently in document pairs p_j that are mutual translations, but not in other documents which at very least are semantically related. For example, in an English document d_j^e where the term $abacus$ appears, the term *abakus* is expected to appear in the corresponding Indonesian document d_j^i.

Let M_E denote a term-document matrix consisting of m English documents and n_E English terms. Let M_I denote a term-document matrix consisting of m Indonesian translation documents and n_I Indonesian terms. For $1 < i < n_E$, row i of M_E represents t_x^e. Similarly, for $1 < i < n_I$, row i of M_I represents w_y^i.

For $1 < j < m$, column j of M_E represents an English document d_j^e and column j of M_I represents d_j^i, the corresponding Indonesian document. Since d_j^e and d_j^i are translations of each other, d_j^e and d_j^i can be located at the same position in the semantic space. Thus, column j of M_E and M_I can be concatenated as a single column where the first n_E rows represents each English term appearing in English document j and the following n_I rows represents each Indonesian terms appearing in Indonesian document j. As a result, the bilingual term-document matrix

$$M = \begin{bmatrix} \mathrm{M_E} \\ \mathrm{M_I} \end{bmatrix}$$

is an $(n_E + n_I) \times m$ matrix in which cell (i, j) contains $f_{i,j}$, the frequency of appearance of t_i^e in d_j^e or t_i^i in d_j^i. Each row i of M forms a term vector of t_i^e for $i \leq n_E$ and of t_i^i for $i > n_E$. Conversely, each column j of M forms a vector representing the appearance of English terms and Indonesian terms in document j.

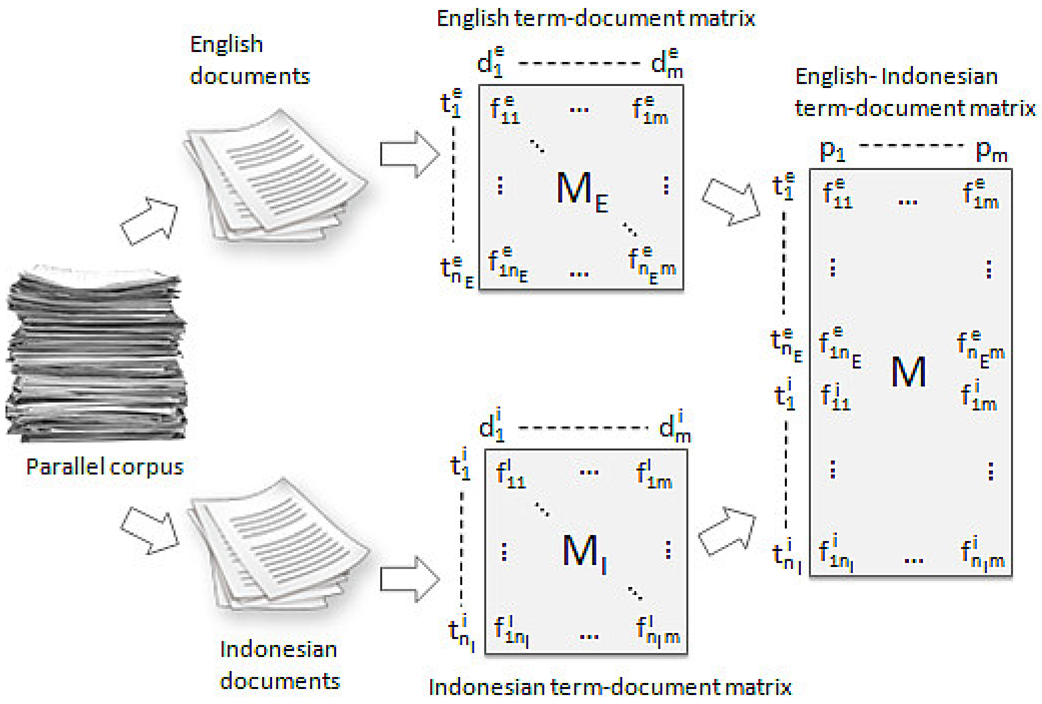
English term-document matrix
English documents
Parallel corpus
Indonesian documents
Indonesian term-document matrix
English- Indonesian term-document matrix
d_1^e d_m^e
t_1^e $t_{n_E}^e$
f_{11}^e f_{1m}^e $f_{1n_E}^e$ $f_{n_E m}^e$
M_E
d_1^i d_m^i
t_1^i $t_{n_I}^i$
f_{11}^I f_{1m}^I $f_{1n_I}^I$ $f_{n_I m}^I$
M_I
p_1 p_m
f_{11}^i f_{1m}^i $f_{1n_I}^i$ $f_{n_I m}^i$
M

Figure 3.1 Constructing English-Indonesian Term-Document Matrix

Figure 3.1 shows the diagram of the construction of English-Indonesian term-document matrix M. Generally, this construction is similar to that of multi-lingual term-document matrix in the experiments carried out by (Rehder, Littman, Dumais, & Landauer, 1997) (see Section 2.4.2). A sample of an English-Indonesian bilingual term-document matrix is given in Table 3.2 below.

Table 3.2 Example of Bilingual Term-Document Matrix

	p_1	p_2	p_3	…	p_m
$t_{aardvark}^e$	$f_{aardvark\ 1}$	$f_{aardvark\ 2}$	$f_{aardvark\ 3}$	…	$f_{aardvark\ \mathrm{m}}$
t_{abacus}^e	$f_{abacus\ 1}$	$f_{abacus\ 2}$	$f_{abacus\ 3}$	…	$f_{abacus\ \mathrm{m}}$
⋮	⋮	⋮	⋮	⋮	⋮
t_{zoo}^e	$f_{zoo\ 1}$	$f_{zoo\ 2}$	$f_{zoo\ 3}$	…	$f_{zoo\ n}$
t_{abad}^i	$f_{abad\ 1}$	$f_{abad\ 2}$	$f_{abad\ 3}$	…	$f_{abad\ \mathrm{n}}$
t_{abakus}^i	$f_{abakus\ 1}$	$f_{abakus\ 2}$	$f_{abakus\ 3}$	…	$f_{abakus\ \mathrm{m}}$
⋮	⋮	⋮	⋮	⋮	⋮
t_{zona}^i	$f_{zona\ 1}$	$f_{zona\ 2}$	$f_{zona\ 3}$	…	$f_{zona\ m}$

With the purpose of determining the degree of importance of a term i in document j, various weighting schemes can be applied to the raw frequency $f_{i,j}$ of the original term-document matrix. Basically, a weighting scheme consists of a local weighting and a global weighting. Local weighting is meant for determining the degree of importance of a term within a document. Conversely, global weighting is meant for determining the degree of importance of a term across the entire documents in a collection. That is, global weighting assigns low weight for terms appearing in many documents. (Landauer, McNamara, Dennis, & Kintsch, 2007)

Local weighting can be achieved by computing *term frequency (TF), binary frequency,* or *logarithmic,* i.e. log(TF +1). On the other hand, global weighting can be achieved by computing (inverse document frequency) *IDF*, *GFIDF* or *Entropy.* (Berry & Browne, 2005) In this work, two weighting schemes are used, namely *TF-IDF* and *Log-Entropy*.

The first weighting scheme, *TF-IDF*, is defined as

$$TF.IDF = f_{ij} \times \log\left(\frac{n}{\sum_{\mathrm{j}} df_{ij}}\right)$$

where TF denotes the frequency of appearance of a term i in document j, df_{ij} denotes the frequency of document containing the term i, and IDF denotes the inverse of frequency of documents containing the term i.

The second weighting scheme, *Log-Entropy*, is defined as

$$Log.Entropy = \log(TF) \times \left(1 + \sum_j \frac{p_{ij} \log 2\ (p_{ij})}{log 2\ n}\right)$$

where $p_{ij} = \frac{TF}{GF}$ and GF denotes the frequency of appearance of a term i in the entire documents in a collection.

3.2.2 Building Bilingual LSA Matrix

The next phase of LSA is applying SVD on an original or weighted bilingual term-document matrix. SVD decomposes the matrix into three matrices, which are a matrix U describing the original term vectors, a matrix V describing the original document vectors, and a diagonal matrix Σ of the singular values. A bilingual LSA matrix is built by reconstructing the three matrices with reduced dimension or rank-k approximation, i.e. the reconstruction only uses the k first columns or dimension of the three matrices. Figure

3.2 below depicts the decomposition of a term-document matrix M by SVD and the reconstruction of the bilingual LSA matrix M_k using rank-k approximation.

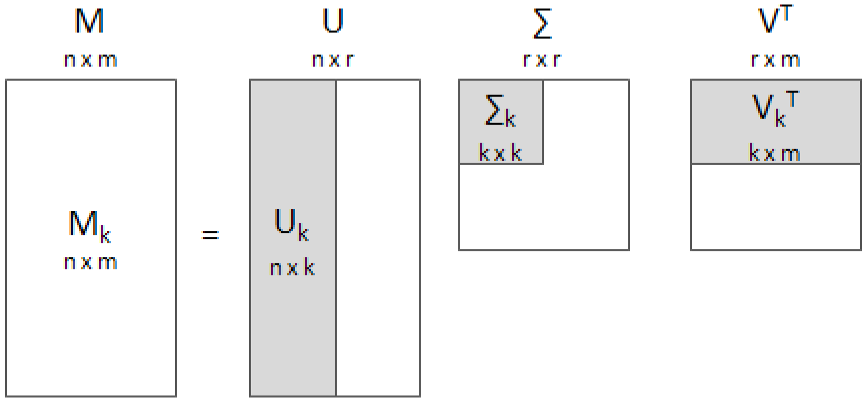

Figure 3.2 Construction of Rank *k* Approximation Term-Document Matrix

By reducing dimension, LSA removes the noise of irrelevant information and combines other information into an abstraction which captures underlying relationships among terms and documents. With respect to LSA's knowledge of the contextual-usage meaning of terms and the context of the documents, the frequency of each term in each document is estimated with a greater or lesser value. Hence, terms which did not appear in some documents might be estimated to appear in those documents. As a result, the term vectors with similar contextual-usage meaning are located near each other in the LSA reduced dimension-space (see Section 2.2.3).

3.3 Design of Bilingual Term Mapping

To approximate the bilingual term mapping task, the similarity between term vectors representing terms in T^e and T^i are compared. These vectors are obtained from the rows of a term-document matrix M. They can also be obtained from the rows of an LSA matrix, i.e. the reduced-dimension term-document matrix M_k.

Given a large enough corpus, all terms in T^e and T^i are expected to be represented by rows in M. The similarity of each of the first n_E rows of M, which represent terms in T^e, with each of the last n_I rows, which represent terms in T^e, is computed using the cosine measure of similarity. Given $\bar{t}_x^e$ and $\bar{t}_y^i$ as the term vectors for t_x^e and t_y^i, the similarity between them is computed as

$$\mathrm{sim}(t_x^e, t_y^i) = \cos\,\theta = \frac{(\bar{t}_x^e)^{\mathrm{T}} \cdot \bar{t}_y^i}{\|\bar{t}_x^e\| \cdot \|\bar{t}_y^i\|}$$

where the range of the similarity value (simval) is [0,1]. Value 1 means two semantic vectors are located exactly in the same point in the semantic space. As a result of bilingual term mapping, several Indonesian terms with the highest similarity values are designated, i.e. as the translation for each English term. The number of translations can be determined by a variety of ways, including choosing the top n terms or setting a minimum threshold of the similarity value.

Table 3.3 Sample of Documents: Titles for Topics on Time and River

Parallel document		Document Title
p_1	d_1^e	A ***Moment*** in ***Time***
	d_1^i	Suatu ***Saat*** dalam Suatu ***Waktu***
p_2	d_2^e	Managing Your ***Time***
	d_2^i	Mengatur ***Waktu*** Anda
p_3	d_3^e	Handling Difficult ***Moment*** Multiple ***Times*** at Chess
	d_3^i	Menangani ***Saat*** Sulit Berulang ***Kali*** dalam Permainan Catur
p_4	d_4^e	***River*** and ***Stream*** Monitoring
	d_4^i	Pengawasan ***Sungai*** dan ***Kali***
p_5	d_5^e	The Need for Clean ***Water*** Grows
	d_5^i	Kebutuhan ***Air*** Bersih Meningkat
p_6	d_6^e	Maintaining ***River Water*** Quality
	d_6^i	Menjaga Kualitas ***Air Kali*** dan ***Sungai***

The following is a simple example of bilingual term mapping. The text documents are taken from a small collection of document titles in Table 3.3. The collection consists of three parallel documents p_1-p_3 related to *time* and another three parallel documents p_4-p_6 related to *river*. Suppose each document only consist of the bold-italicised terms. The frequency of each term appearing in each corresponding document is enumerated in the term-document matrix shown in Table 3.4.

Next, the bilingual LSA matrix for the term-document matrix in Table 3.4 is computed with rank-2 approximation. The result is shown in Table 3.5. Notice that each term frequency is estimated with a lesser or greater value in the bilingual LSA matrix.

Table 3.4 Sample of English-Indonesian Term-Document Matrix Based on Table 3.3

	p_1	p_2	p_3	p_4	p_5	p_6
t^e_{moment}	1	0	1	0	0	0
t^e_{river}	0	0	0	1	0	1
t^e_{stream}	0	0	0	1	0	0
t^e_{time}	1	1	1	0	0	0
t^e_{water}	0	0	0	0	1	1
t^i_{air}	0	0	0	0	1	1
t^i_{kali}	0	0	1	1	0	1
t^i_{saat}	1	0	1	0	0	0
t^i_{sungai}	0	0	0	1	0	1
t^i_{waktu}	1	1	0	0	0	0

Using the bilingual LSA matrix, the similarity between each English term and each Indonesian term is computed. As a result of bilingual term mapping, for each English term, each Indonesian term are sorted with descending order, i.e. from the one with the highest similarity value to the lowest. For example, t^e_{moment} and t^e_{time} are properly map to t^i_{saat} and t^i_{waktu}.

Table 3.6 Table 3.6 lists the bilingual term mapping result using the bilingual LSA matrix in Table 3.5.

Table 3.5 Sample of English-Indonesian LSA Matrix with Rank-2 Approximation Based on Table 3.4

	p_1	p_2	p_3	p_4	p_5	p_6
t^e_{moment}	0.914	0.466	0.824	0.061	-0.042	0.022
t^e_{river}	-0.095	-0.068	0.177	0.773	0.339	1.016
t^e_{stream}	-0.028	-0.022	0.089	0.334	0.146	0.439
t^e_{time}	1.167	0.596	1.038	0.039	-0.071	-0.024
t^e_{water}	-0.116	-0.074	0.095	0.584	0.259	0.770
t^i_{air}	-0.116	-0.074	0.095	0.584	0.259	0.770
t^i_{kali}	0.328	0.146	0.579	0.861	0.345	1.105
t^i_{saat}	0.914	0.466	0.824	0.061	-0.042	0.022
t^i_{sungai}	-0.095	-0.068	0.177	0.773	0.339	1.016
t^i_{waktu}	0.744	0.382	0.637	-0.050	-0.077	-0.113

According to this small example, LSA is able to properly map the English terms to their appropriate translation in Indonesian by taking into account the Indonesian term with the highest similarity value as the translation. For example, t^e_{moment} and t^e_{time} are properly map to t^i_{saat} and t^i_{waktu}.

Table 3.6 Sample of English-Indonesian Term Mapping Result Based on Table 3.5

t^e_{moment}		t^e_{river}		t^e_{stream}		t^e_{time}		t^e_{water}	
t^i_y	simval	t^i_y	simval	t^i_y	simval	t^i_y	simval	t^i_y	simval
t^i_{saat}	1.000	t^i_{sungai}	1.000	t^i_{sungai}	0.999	t^i_{saat}	0.999	t^i_{air}	1.000
t^i_{waktu}	0.987	t^i_{air}	0.998	t^i_{air}	0.996	t^i_{waktu}	0.993	t^i_{sungai}	0.998
t^i_{kali}	0.431	t^i_{kali}	0.922	t^i_{kali}	0.934	t^i_{kali}	0.394	t^i_{kali}	0.896
t^i_{sungai}	0.047	t^i_{saat}	0.047	t^i_{saat}	0.080	t^i_{sungai}	0.007	t^i_{saat}	-0.015
t^i_{air}	-0.015	t^i_{waktu}	-0.115	t^i_{waktu}	-0.082	t^i_{air}	-0.056	t^i_{waktu}	-0.176

In practise, however, bilingual term mapping is a very difficult task to ask of LSA to accomplish. Bilingual term mapping can be seen as an extremely unconstrained task of *word alignment* in the machine translation field. Most word alignment systems employ parallel corpora, which have been aligned down to sentence level, to exploit some measure of syntactic information. LSA, however, treats documents as bags of words, and hence has no syntactic knowledge whatsoever.

3.4 Design of Bilingual Concept Mapping

To approximate the bilingual concept mapping task, the similarity between the conceptual semantic vectors representing concepts in C^e and C^i are compared. These vectors can be approximated by constructing a set of textual context representing a concept c. For example, given a row in Table 3.1, the terms conveying the concept, the gloss, and the example sentences are included in the subsets of the set of textual context for c.

More formally, the set of textual context for concept c is

$$\{T_c, G_c, X_c\}$$

where T_c is the set of terms conveying the concept c, i.e. $\omega(c)$, G_c is the set of terms of the gloss, and X_c is the set of terms of the example sentences. Given $\bar{c}$ as the conceptual semantic vector for c and $\bar{t}$ as a term vector for term t, then $\bar{c}$ is computed as follows:

$$\bar{c} = \left(\frac{W_{T_c}}{N_{T_c}} \sum_{t \in T_c} \bar{t} \right) + \left(\frac{W_{G_c}}{N_{G_c}} \sum_{t \in G_c} \bar{t} \right) + \left(\frac{W_{X_c}}{N_{X_c}} \sum_{t \in X_c} \bar{t} \right)$$

where N_x represents the number of term in a subset x. For each subset of the textual context set, the term vectors, i.e. the rows of M, are averaged and weighted. A weight W_x is given to a subset x according to the importance of that subset to express the concept meaning. Given a large enough corpus, these textual context terms are expected to be represented by rows in M to form an adequate conceptual semantic vector for the concept c. The conceptual semantic vector for c is the sum of the weighted average subset vectors.

Using the bilingual LSA matrix in Table 3.5 above, the conceptual semantic vectors for concept c_2^e, c_3^i, and c_5^i in Table 3.1 can be computed. In addition, c_4^e is defined as a concept that can be conveyed by t_{river}^e. The concept is associated with the gloss *"a large natural stream of water (larger than a creek)"* and an example sentence *"the river was navigable for 50 miles"*.

For each concept mentioned above, several terms are chosen as representatives for the gloss and the example sentence. These terms along with the term conveying the concept construct the textual concept set for that concept. Sample of these textual context sets is given in Table 3.7.

Table 3.7 Sample of Set of Textual Context and Conceptual Semantic Vector Based on English-Indonesian LSA Matrix in Table 3.5

Concept	Set of Textual Context	Conceptual Semantic Vector
c_2^e	$\{\{t_{time}^e\},\{t_{moment}^e\},\{t_{time}^e\}\}$	[0.364, 0.186, 0.325, 0.015, -0.021, -0.003]
c_4^e	$\{\{t_{river}^e\},\{t_{stream}^e,t_{water}^e\},\{t_{river}^e\}\}$	[-0.036, -0.025, 0.060, 0.272, 0.120, 0.358]
c_3^i	$\left\{\{t_{waktu}^i\},\{t_{saat}^i\},\{t_{waktu}^{\mathrm{i}}\}\right\}$	[0.265, 0.136, 0.231, -0.006, -0.022, -0.024]
c_5^i	$\left\{\{t_{kali}^i\},\{t_{sungai}^i\},\{t_{kali}^{\mathrm{i}}\}\right\}$	[0.067, 0.027, 0.153, 0.278, 0.114, 0.359]

In constructing a conceptual semantic vector, term vectors of textual context terms are taken from an LSA matrix. For the sample above, LSA matrix in Table 3.5 is used. For each subset of each textual context set, an average vector of the term vectors of its terms is computed. Then, a weight of 60%, 30%, and, 10% is applied to the average vectors of the subsets with respect to their ordering and importance. Finally, the sum of the three weighted average vectors of the subsets is computed as the conceptual semantic vector.

The similarity between each pair of bilingual conceptual semantic vectors is computed using the cosine measure of similarity. Like bilingual term mapping, the result of bilingual concept mapping can be obtained by choosing top n most similar conceptual semantic vector pairs for each concept to be mapped. That is, pairs of bilingual equivalent concepts represented by pairs of bilingual conceptual semantic vectors with the highest similarity values. Another way is by taking all pairs whose similarity values are above a custom threshold. There are a variety of ways to determine a threshold, e.g. average of all similarity values.

The result of computing the similarity between each pair of bilingual conceptual semantic vectors in Table 3.7 is shown in Table 3.8. According to this small example, LSA is able to map the English concepts to the equivalent Indonesian concepts in the first place.

Table 3.8 Sample of Bilingual Concept Mapping Result Based on Table 3.7

c_2^e		c_4^e	
c_3^i	0.997	c_5^i	0.955
c_5^i	0.306	c_3^i	-0.064

Bilingual concept mapping can be viewed as a generalisation of bilingual word sense disambiguation. Whilst the terms from both languages have been determined, the task is to disambiguate from the different word senses. A textual context set serves as a bag of words, which gives context to the word to be disambiguated. Hence, a conceptual semantic vector represents the word sense in a location which is near to other vectors representing similar word sense. Moreover, it should be near to one or more conceptual semantic vectors of the opposite language which represent the same word sense (see Section 2.4.3).