English-to-Indonesian Lexical Mapping using Latent Semantic Analysis

Eliza Margaretha, Franky, and Ruli Manurung
Faculty of Computer Science
University of Indonesia
elm40@ui.edu, fra43@ui.edu, maruli@cs.ui.ac.id

Abstract

This paper describes an attempt to automatically map English words and concepts, derived from the Princeton WordNet, to their Indonesian analogues appearing in an Indonesian lexicon. Using Latent Semantic Analysis (LSA), a semantic model is derived from an English-Indonesian parallel corpus. Given a particular word or concept, the semantic model is then used to identify its neighbours in a high-dimensional semantic space. Results from various experiments indicate that even with relatively small collections, LSA is able to discern some measure of implicit semantic knowledge from the training data.

1. Overview

Latent semantic analysis is a method that represents the contextual meaning of words appearing in a corpus as a vector in a high-dimensional semantic space [9]. As such, LSA is a powerful method for *word sense disambiguation*.

In this paper we present an extension to LSA into a bilingual context that is similar to [10, 2, 3]. We then apply it to two tasks, bilingual word mapping, which seeks to find translations of a lexical entry from one language to another, and bilingual concept mapping, which defines equivalence classes over concepts defined in two language resources. The specific lexical resources used are Princeton WordNet [4], an English semantic lexicon, and the *Kamus Besar Bahasa Indonesia* (KBBI)[1], considered by many to be the official dictionary of the Indonesian language.

In this paper, we first discuss relevant issues such as the construction of WordNets (Section 2) and the theory of LSA (Section 3) before outlining our approach and design (Sections 4 and 5). Our experiments and results are presented in Section 6.

[1]The KBBI is the copyright of the Language Centre, Indonesian Ministry of National Education.

2. WordNets

The original Princeton WordNet [4] is a lexical resource for the English language containing a large database of nouns, verbs, adjectives, and adverbs, which are clustered together based on their meaning into synonym sets, or *synsets*. Crucially, each instance of a word in a synset is marked as a separate *sense* of that word. WordNet further specifies semantic relations between synsets, such as hypernymy and meronymy.

As a freely-available semantic resource that comprehensively catalogues distinct English word senses, WordNet is widely used in NLP research such as word sense disambiguation, information retrieval, natural language generation, and is increasingly relevant in wider applications such as the Semantic Web[2].

WordNets have since been created for many other languages, with notable efforts including EuroWordNet [12] and the establishment of the Global WordNet Association[3]. A WordNet for Indonesian would serve as a useful resource for various research communities.The work presented in this paper is part of a larger ongoing project to develop an initial version of such a resource.

One area where WordNet is commonly used is *word sense disambiguation*, i.e. the task of identifying which particular sense of a word is intended in a given context [1]. Many approaches to this difficult problem exist. Whereas in most existing work WordNet is used as a knowledge source for performing disambiguation on words appearing in an external context, in this work we attempt to disambiguate senses appearing in a WordNet using knowledge obtained from an external source.

3. Latent semantic analysis

Latent semantic analysis, or simply LSA, is a method to discern underlying semantic information from a given cor-

[2]See, for example, the Suggested Upper Merged Ontology at http://www.ontologyportal.org

[3]http://www.globalwordnet.org/

pus of text, and to subsequently represent the contextual meaning of the words in the corpus [9]. As such, LSA is a powerful method for word sense disambiguation. LSA infers the meaning of words by observing cooccurrences of words over local meaning-bearing passages. Unlike conventional word frequency-based methods, LSA is able to establish the semantic relationship between words that do not occur in the same passage.

In LSA, the meaning of a word is represented as the average of the meaning of all the passages in which it occurs. Likewise, the meaning of a passage is the average of the meaning of all the words it contains. The ability of LSA to concurrently derive knowledge of words and passages in such a manner is provided by the singular value decomposition (SVD) as its mathematical foundation [6].

Initially, a corpus is represented as an $n \times m$ word-passage matrix M, where cell $[n, m]$ represents the occurrence of the n-th word in the m-th passage. Thus, each row of M represents a word and each column represents a passage.

The SVD is then applied to M, decomposing it such that $M = U\Sigma V^T$, where

- U is an $m \times m$ matrix of left singular vectors,
- V^T is an $n \times n$ matrix of right singular vectors, and
- Σ is an $n \times m$ matrix containing the singular values of M.

Crucially, this decomposition factors M using an orthonormal basis that produces an optimal reduced rank approximation matrix [7]. By reducing dimensions of the matrix, irrelevant information and noisy variability of word choice associated with passages is removed. The optimal rank reduction yields useful induction of implicit relations. However, finding the optimal level of rank reduction is an empirical issue.

LSA has been successfully applied on various tasks related to meaning-based cognition, such as rating the content adequacy of expository essays. It has also been shown to improve information retrieval by up to 30% [8].

4. Task definitions

For the purposes of our work, a WordNet can be viewed as a 4-tuple (C, W, χ, ω) as follows:

- A concept $c \in C$ is a **semantic entity**, which represents a distinct, specific meaning. Each concept is associated with a **gloss**, which is a textual description of its meaning. For example, we could define two concepts, c_1 and c_2, where the former is associated with the gloss *"a financial institution that accepts deposits and channels the money into lending activities"* and the latter with *"sloping land (especially the slope beside a body of water"*.
- A word $w \in W$ is an **orthographic entity**, which represents a word in a particular language (in the case of Princeton WordNet, English). For example, we could define two words, w_1 and w_2, where the former represents the orthographic string *bank* and the latter represents *spoon*.
- A word may convey several different concepts. The function $\chi : W \rightarrow \mathcal{P}(C)$ returns all concepts conveyed by a particular word. Thus, $\chi(w)$, where $w \in W$, returns $C_w \subset C$, the set of all concepts that can be conveyed by w. Using the examples above, $\chi(w_1) = \{c_1, c_2\}$.
- Conversely, a concept may be conveyed by several words. The function $\omega : C \rightarrow \mathcal{P}(W)$ returns all words that can convey a particular concept. Thus, $\omega(c)$, where $c \in C$, returns $W_c \subset W$, the set of all words that convey c. Using the examples above, $\omega(c_1) = \omega(c_2) = \{w_1\}$.

We can define different WordNets for different languages, e.g. $N^e = (C^e, W^e, \chi^e, \omega^e)$ and $N^i = (C^i, W^i, \chi^i, \omega^i)$. We also introduce the notation w_i^x to denote word i in W^x and c_j^x to denote concept j in C^x.For the sake of discussion, we will assume N^e to be an English WordNet, and N^i to be an Indonesian WordNet.

If we make the assumption that concepts are language-independent, N^e and N^i should theoretically share the same set of universal concepts, C. In practice, however, we may have two WordNets with different conceptual representations, hence the distinction between C^e and C^i. We introduce the relation $E : C^e \times C^i$ to denote the explicit mapping of equivalent concepts in C^e and C^i.

We now describe two tasks that can be performed between N^e and N^i, namely **bilingual concept mapping** and **bilingual word mapping**.

The task of bilingual concept mapping is essentially the establishment of the concept equivalence relation E. For example, given the example concepts in Table 1, bilingual concept mapping seeks to establish $E = \{(c_1^e, c_1^i), (c_1^e, c_2^i), (c_2^e, c_3^i), (c_3^e, c_4^i), \}$.

Note that since the members of a concept set C should denote distinct semantic elements, E should define a one-to-one relation. However, different WordNets employ different degrees of granularity in analysis of polysemy, hence a concept in one WordNet may map to more than one concept in another. If this occurs in both directions E may result in the conflation of several distinct concepts within a WordNet.

Concept	Gloss	Example
c^e_1	an instance or single occasion for some event	*"this **time** he succeeded"*
c^e_2	a suitable moment	*"it is **time** to go"*
c^e_3	a reading of a point in time as given by a clock	*"do you know what **time** it is?"*
c^i_1	kata untuk menyatakan kekerapan tindakan (a word signifying the frequency of an event)	*"dalam satu minggu ini, dia sudah empat **kali** datang ke rumahku" (this past week, she has come to my house four **times**)*
c^i_2	kata untuk menyatakan salah satu waktu terjadinya peristiwa yg merupakan bagian dari rangkaian peristiwa yg pernah dan masih akan terus terjadi (a word signifying a particular instance of an ongoing series of events)	*"untuk **kali** ini ia kena batunya" (this **time** he suffered for his actions)*
c^i_3	saat yg tertentu untuk melakukan sesuatu (a specific time to be doing something)	*"**waktu** makan" (eating **time**)*
c^i_4	saat tertentu, pada arloji jarumnya yg pendek menunjuk angka tertentu dan jarum panjang menunjuk angka 12 (the point in time when the short hand of a clock points to a certain hour and the long hand points to 12)	*"ia bangun **jam** lima pagi" (she woke up at five **o'clock**)*
c^i_5	sebuah sungai yang kecil (a small river)	*"air di **kali** itu sangat keruh" (the water in that **small river** is very murky)*

Table 1. Sample concepts in C^e and C^i

The task of bilingual word mapping is to find, given word $w^e_x \in W^e$, the set of all its plausible translations in W^i, *regardless* of the concepts being conveyed. We can also view this task as computing the union of the set of all words in W^i that convey the set of all concepts conveyed by w^e_x. More formally, we compute the set $\{w^i_y : w^i_y \in \omega^i(c^i) \text{ where } (c^e, c^i) \in E \text{ and } c^e \in \chi^e(w^e_x)\}$.

For example, in Princeton WordNet, given w^e_{time} (i.e. the English orthographic form $time$), $\chi^e(w^e_{time})$ returns more than 15 different concepts, among others $\{c^e_1, c^e_2, c^e_3\}$ (see Table 1).

In Indonesian, assuming the relation E as defined above, the set of words that convey c^i_1, i.e. $\omega^i(c^i_1)$, includes w^i_{kali} (as in *"**kali** ini dia berhasil" = "this **time** she succeeded"*). On the other hand, $\omega^i(c^i_3)$ may include w^i_{waktu} (as in *"ini **waktu**nya untuk pergi" = "it is **time** to go"*) and w^i_{saat} (as in *"sekarang **saat**nya menjual saham" = "now is the **time** to sell shares"*), and lastly, $\omega^i(c_4)$ may include w^i_{jam} (as in *"apa anda tahu **jam** berapa sekarang?" = "do you know what **time** it is now?"*).

Thus, the bilingual word mapping task seeks to compute, for the English word w^e_{time}, the set of Indonesian words $\{w^i_{kali}, w^i_{waktu}, w^i_{saat}, w^i_{jam}, \ldots\}$. Note that each of these Indonesian words may convey different concepts, e.g. $\chi^i(w^i_{kali})$ may include c^i_5 in Table 1.

5. Automatic mapping with LSA

In this section we describe a method that exploits linguistic information implicitly encoded within a parallel corpus to automatically perform bilingual word and concept mapping.

We define a parallel corpus P as a set of pairs $p = (d_e, d_i)$, where d_e is a document written in the language of N^e, and d_i is its translation in the language of N^i.

Intuitively, we would expect that if two words w^e_x and w^i_y consistently occur in documents that are translations of each other, but not in other documents, that they would at the very least be semantically related, and possibly even be translations of each other. For instance, imagine a parallel corpus consisting of news articles written in English and Indonesian: in English articles where the word *Japan* occurs, we would expect the word *Jepang* to occur in the corresponding Indonesian articles.

This intuition can be represented in a word-document matrix as follows: let M_E be a word-document matrix of m English documents and n_E English words, and M_I be a word-document matrix of m Indonesian documents and n_I Indonesian words. The documents are arranged such that, for $1 \leq j \leq m$, the English document represented by column j of M_E and the Indonesian document represented by column j of M_I form a pair of translations. Since they are translations, we can view them as occupying exactly the same point in semantic space, and could just as easily view column j of both matrices as representing the union, or concatenation, of the two articles.

Consequently, we can construct the bilingual word-document matrix

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

which is an $(n_E + n_I) \times m$ matrix where cell $[i, j]$ contains the number of occurrences of word i in article j. Row i forms the semantic vector of, for $i \leq n_E$, an English word, and for $i > n_E$, an Indonesian word. Conversely, column j forms a vector representing the English and Indonesian words appearing in translations of document j.

This approach is similar to that of [10, 2, 3].

The LSA approach described in Section 3 can be applied to this bilingual word-document matrix. Computing the SVD of this matrix and reducing the rank should unearth implicit patterns of semantic concepts. The vectors representing English and Indonesian words that are closely related should have high similarity; word translations more

so.

To approximate the bilingual word mapping task, we compare the similarity between the semantic vectors representing words in W^e and W^i. Specifically, for the first n_E rows in M which represent words in W^e, we compute their similarity to each of the last n_I rows which represent words in W^i. Given a large enough corpus, we would expect all words in W^e and W^i to be represented by rows in M.

To approximate the bilingual concept mapping task, we compare the similarity between the semantic vectors representing concepts in C^e and C^i. These vectors can be approximated by first constructing a set of textual context representing a concept c. For example, we can include the words that convey it ($\omega(c)$) together with the words from its gloss and example sentences. The semantic vector of a concept is then a weighted average of the semantic vectors of the words contained within this context set, i.e. rows in M. Again, given a large enough corpus, we would expect enough of these context words to be represented by rows in M to form an adequate semantic vector for the concept c.

Although this approach does not yield any explicit mappings in the shape of the functions χ, ω and the relation E, nevertheless LSA encodes this information implicitly in the semantic vectors of words.

These are both very difficult tasks to ask of LSA to accomplish. For instance, given this setup, bilingual word mapping can be seen as an extremely unconstrained instance of the *word alignment* task in the machine translation field. Most word alignment systems employ parallel corpora that are aligned down to the sentence level to exploit some measure of syntactic information. LSA, however, treats articles as bags of words, and hence has no syntactic knowledge whatsoever. Nevertheless, LSA has been shown to be valuable as additional probabilistic bias within a word alignment system employing HMM-based syntactic models [3]. Bilingual concept mapping can be viewed as a generalization of bilingual word sense disambiguation, where at least the word forms from both languages have been determined, and the task is to disambiguate from the different word senses.

6. Experiments and Discussion

6.1. Existing resources

In this section we describe the various lexical resources used in our experiments:

- For the English lexicon, we used the most current version of WordNet, version 3.0[4]. For each of the 117659 distinct synsets, we only use the following data: the set of words belonging to the synset, the gloss, and example sentences, if any. The union of these resources yields a set 169583 unique words.
- For the Indonesian lexicon, we used an electronic version of the KBBI, which was developed at the Faculty of Computer Science, University of Indonesia, during the mid-90s. For each of the 85521 distinct word sense definitions, we use the following data: the list of sublemmas, i.e. inflected forms, along with gloss and example sentences, if any. The union of these resources yields a set of 87171 unique words.
- Our main parallel corpus consists of 3273 English and Indonesian article pairs taken from the ANTARA news agency. This collection was developed at the Information Retrieval Lab, University of Indonesia. The documents have been paired using a statistical approach developed by Mirna Adriani and Monica Lestari Paramita[5].
- A bilingual English-Indonesia dictionary was constructed using various online resources, including a handcrafted dictionary by Hantarto Widjaja[6], kamus.net, and Transtool v6.1, a commercial translation system. In total, this dictionary maps 37678 unique English words to 60564 unique Indonesian words.

6.2. Bilingual word mapping

Our experiment with bilingual word mapping was set up as follows: firstly, we define a collection of article pairs derived from the ANTARA collection, and from it we set up a bilingual word-document matrix (see Section 5). The LSA process is subsequently applied on this matrix, i.e. we first compute the SVD of this matrix, and then use it to compute the optimal k-rank approximation. Finally, based on this approximation, for a randomly chosen set of vectors representing English words, we compute the n nearest vectors representing the n most similar Indonesian words. This is conventionally computed using the cosine of the angle between two vectors.

Within this general framework, there are several variables that we experiment with, as follows:

1. **Collection size.** Three subsets of the parallel corpus were randomly created: P_{100} contains 100 article pairs, P_{500} contains 500 article pairs, and P_{1000} contains 1000 article pairs. Each subsequent subset wholly contains the previous subsets, i.e. $P_{100} \subset P_{500} \subset P_{1000}$.

[4] More specifically, the SQL version available from `http://wnsqlbuilder.sourceforge.net`

[5] publication forthcoming

[6] `http://hantarto.definitionroadsafety.org`

film	0.814
filmnya	0.698
sutradara	0.684
garapan	0.581
perfilman	0.554
penayangan	0.544
kontroversial	0.526
koboi	0.482
irasional	0.482
frase	0.482

(a)

pembebanan	0.973
kijang	0.973
halmahera	0.973
alumina	0.973
terjadwal	0.973
viskositas	0.973
Tabel	0.973
royalti	0.973
reklamasi	0.973
penyimpan	0.973

(b)

Table 2. The 10 most similar Indonesian words for the English words (a) film and (b) million

2. **Rank reduction.** For each collection, we applied LSA with different degrees of rank approximation, namely 10%, 25%, and 50% the number of dimensions of the original collection. Thus, for P_{100} we compute the 10, 25, and 50-rank approximations, for P_{500} we compute the 50, 125, and 250-rank approximations, and for P_{1000} we compute the 100, 250, and 500-rank approximations.

As an example, Table 2 presents the results of mapping w^e_{film} and $w^e_{billion}$, i.e. the two English words *film* and *billion*, using the P_{1000} training collection with 500-rank approximation. The former shows a successful mapping, while the latter shows an unsuccessful one. Bilingual LSA correctly maps w^e_{film} to its translation, w^i_{film}, despite the fact that they are treated as separate elements, i.e. their shared orthography is completely coincidental. Additionally, the other Indonesian words it suggests are semantically related, e.g. *sutradara* (director), *garapan* (creation), *penayangan* (screening), etc. On the other hand, the suggested word mappings for $w^e_{billion}$ are incorrect, and the correct translation, *milyar*, is missing. We suspect this may be due to several factors. Firstly, *billion* does not by itself invoke a particular semantic frame, and thus its semantic vector might not suggest a specific conceptual domain. Secondly, *billion* can sometimes be translated numerically instead of lexically. In general, however, we believe this failure is simply due to the lack of data: the collection is simply too small to provide useful statistics that represent semantic context. Similar LSA approaches are commonly trained on collections of text numbering in the tens of thousands of articles.

Note as well that the absolute vector cosine values do not accurately reflect the correctness of the word translations. To properly assess the results of this experiment, evaluation against a gold standard is necessary. This is achieved by comparing its precision and recall against the Indonesian words returned by the bilingual dictionary, i.e. how isomorphic is the set of LSA-derived word mappings with a human-authored set of word mappings?

We provide two baselines as comparison. The first is a truly random baseline: given an English word, we randomly select n Indonesian words appearing in the training collection. The second baseline computes vector nearness between English and Indonesian words on the original word-article occurrence matrix, or in other words, applying LSA without any rank approximation. Other approaches are possible, e.g. mutual information [11].

The results are presented in Table 3. For each approach, we experimented with taking the top 1, 10, 50, and 100 mappings, and computing precision and recall against the mappings returned by our bilingual dictionary.

For each collection size (P_{100}, P_{500}, P_{1000}) we also show the results of the random baseline (RND), the non-LSA vector comparison baseline (FRQ), and LSA with dimensional reduction of 10%, 25%, and 50%.

As expected, all the vector comparison-based approaches comfortably outperform the random baseline.

Additionally, the larger the training collection, the better the performance. Consistently, all other parameters being equal, the precision and recall values when using P_{1000} are better than the values when using P_{500}, which in turn are better than when using P_{100}.

For small training collections, the original word-document cooccurrence matrix yields better results than the LSA-based approximation. However, as the size of the training collection increases, LSA gradually outperforms this baseline. The best performance was obtained using LSA on the P_{1000} collection and applying a 500-rank approximation.

Unfortunately, due to time restrictions and computational resource constraints, we were unable to carry out experiments on larger collections.

We also experimented with a common technique in Information Retrieval, i.e. preprocessing the collection by removing all instances of *stopwords* beforehand, but the results were consistently worse.

6.3. Bilingual concept mapping

Using the same resources from the previous experiment, we ran an experiment to perform bilingual concept mapping by replacing the vectors to be compared with semantic vectors for concepts (see Section 5). For concept $c^e \in C^e$, i.e. a WordNet $synset$, we constructed a set of textual context as the union of $\omega(c)$, the set of words in the gloss of c^e, and the set of words in the example sentences associated with c^e. To represent our intuition that the words in $\omega(c)$ played more

Precision / Recall	P_{100}					P_{500}					P_{1000}				
	RND	FRQ	10%	25%	50%	RND	FRQ	10%	25%	50%	RND	FRQ	10%	25%	50%
Top 1 Precision	0.000	**0.140**	0.030	0.110	0.090	0.000	**0.330**	0.250	0.250	0.290	0.000	**0.410**	0.310	0.290	0.330
Top 1 Recall	0.000	**0.056**	0.012	0.026	0.028	0.000	**0.158**	0.105	0.107	0.123	0.000	**0.181**	0.106	0.105	0.131
Top 10 Precision	0.000	0.025	0.014	**0.028**	0.022	0.001	0.045	0.043	0.047	**0.049**	0.000	0.056	**0.059**	0.057	0.055
Top 10 Recall	0.000	**0.100**	0.041	0.087	0.087	0.001	**0.200**	0.151	0.193	0.207	0.000	**0.233**	0.209	0.232	0.214
Top 50 Precision	0.001	**0.010**	0.006	0.009	0.009	0.001	0.013	0.014	0.014	**0.014**	0.000	0.015	0.017	0.016	**0.017**
Top 50 Recall	0.002	**0.186**	0.086	0.141	0.149	0.002	0.260	0.235	0.253	**0.272**	0.000	0.264	0.275	0.277	**0.283**
Top 100 Precision	0.001	0.006	0.004	0.006	**0.006**	0.001	0.007	**0.009**	0.008	0.008	0.001	0.008	0.009	0.010	**0.010**
Top 100 Recall	0.009	0.221	0.118	0.200	**0.22**	0.005	0.275	0.275	**0.293**	0.296	0.006	0.279	0.294	0.293	**0.302**

Table 3. Precision and recall results for bilingual word mapping experiment

of an important role in defining the semantic vector than the words in the gloss and example, we applied a weight of 60%, 30%, and 10% to the three components, respectively. Similarly, a semantic vector representing a a concept $c^i \in C^i$, i.e. an Indonesian word sense in the KBBI, was constructed from a textual context set composed of the sublemma, the definition, and the example of the word sense, using the same weightings. Note that we only average word vectors if they appear in the collection (depending on the experimental variables used).

Initially, for each synset, we computed the n nearest vectors representing the n most similar Indonesian word senses. However, manual inspection revealed that the results were far from satisfactory. As before, the vector cosine values did not accurately reflect the semantic similarity. We believe this is due to the lack of context provided by the small training collection. However, the problem is more severe than during the word mapping exercise, as the LSA matrix is consulted for semantic information regarding not just a single word, but for a whole set of words.

To illustrate, Table 4(a) and 4(b) presents a successful and unsuccessful example of mapping a WordNet synset. For each example we show the synset ID and the **ideal** textual context set, i.e. the set of words that convey the synset, its gloss and example sentences. We then show the **actual** textual context set with the notation $\{\{X\}, \{Y\}, \{Z\}\}$, where X, Y, and Z are the subset of words that appear in the training collection. We then show the two most similar Indonesian word senses. For each sense we show the vector similarity score, the KBBI ID and its ideal textual context set, i.e. the sublemma, its definition and example sentences. We then show the actual textual context set with the same notation as above.

In the first example, the textual context sets from both the WordNet synset and the KBBI senses are fairly large, and provide sufficient context for LSA to choose the correct KBBI sense. However, in the second example, the textual context set for the synset is very small, due to the words not appearing in the training collection. Furthermore, it does not contain any of the words that truly convey the concept. As a result, LSA is unable to identify the correct KBBI sense.

These results suggested that the task was too much to ask

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*}}

1. **Similarity:** 0.804, **KBBI ID:** k39607, **Sublemma:** mengejar, **Definition:** berlari untuk menyusul menangkap dsb memburu, **Example:** ia berusaha mengejar dan menangkap saya, **Textual context set:** {{*mengejar*}, {*memburu, berlari, menangkap, untuk, menyusul*}, {*berusaha, dan, ia, mengejar, saya, menangkap*}}

2. **Similarity:** 0.781, **KBBI ID:** k14029, **Sublemma:** memburu, **Definition**: mengejar untuk menangkap binatang dl hutan dsb, **Example**: memburu di daerah suaka margasatwa adalah terlarang, **Textual context set:** {{*memburu*}, {*mengejar, hutan, menangkap, binatang, untuk*}, {*adalah, memburu, suaka, di, terlarang, daerah, margasatwa*}}

(a)

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*}}

1. **Similarity:** 0.695, **KBBI ID:** k02421, **Sublemma:** alur, **Definition:** jalinan peristiwa dl karya sastra untuk mencapai efek tertentu pautannya dapat diwujudkan 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*}, {}}

2. **Similarity:** 0.688, **KBBI ID:** k26302, **Sublemma:** gelugur, **Definition**: pohon mangga hutan buahnya berwarna merah kekuning-kuningan dipakai untuk mengasami gulai garcinia macrophylla, **Example**: (none), **Textual context set:** {{}, {*mangga, berwarna, merah, hutan, pohon, dipakai, untuk*}, {}}

(b)

Table 4. Example of (a) successful and (b) unsuccessful concept mappings

of LSA. We reformulated another experiment which more closely resembles the word sense disambiguation problem. Given a WordNet synset, the task is to select the most appropriate Indonesian sense from a subset of senses that have been selected based on their words appearing in our bilingual dictionary. These specific senses are called *suggestions*. For instance, instead of comparing the vector representing *communication* with every single Indonesian sense in the KBBI, in this task we only compare it against *suggestions* with a limited range of sublemmas, e.g. *komunikasi*,

Judges	Synsets	Fleiss kappa values		
		Judges only	Judges + LSA	Judges + Random
≥ 2	140	0.427	0.210	0.174
≥ 3	25	0.467	0.276	0.254
≥ 4	10	0.576	0.359	0.364
≥ 5	6	0.464	0.331	0.325
Average:		0.483	0.294	0.279

Table 5. Results of concept mapping

perhubungan, *hubungan*, etc.

This setup is thus identical to that of an ongoing experiment here to manually map WordNet synsets to KBBI senses. Consequently, this facilitates assessment of the results by computing the level of agreement between the LSA-based mappings with human annotations.

As a baseline, we select a random suggested Indonesian word sense as a mapping for each English word sense. Subsequently, we compute the Fleiss kappa [5] of this result together with the human judgements.

For this experiment, we used the P_{1000} training collection with a rank reduction of 50%, as this was the configuration that yielded the best results for the word mapping experiment (Section 6.2). The results are presented in Table 5. The reported random baseline is an average of three separate runs.

The average level of agreement between the LSA mappings and the human judges (0.294) is not as high as between the human judges themselves (0.483). Nevertheless, in general it is better than the baseline (0.279), which suggests that LSA is indeed managing to capture some measure of bilingual semantic information implicit within the parallel corpus.

7. Summary

We have presented a model of computing bilingual word and concept mappings between two WordNets, namely Princeton WordNet and the KBBI, using an extension to LSA that exploits implicit semantic information contained within a parallel corpus.

The results, whilst far from conclusive, indicate that with this approach, LSA is able to discern some measure of semantic information that enables it to perform better than random baselines, but is not yet able to attain levels comparable to human judgments.

In general, we believe the failure is down to lack of context provided by the collection size – this suggests obvious further work of experimenting with larger corpora. Furthermore, there are other variables we have yet to test, such as the weighting function for the word-document cooccurrence matrix (typically, some form of log-entropy weighting, or TF-IDF [10]).

Specifically for bilingual word mapping, a finer granularity of alignment, e.g. at the sentential level, should greatly increase accuracy [3].

Acknowledgment

The work presented in this paper is supported by an RUUI (Riset Unggulan Universitas Indonesia) 2007 research grant from DRPM UI (Direktorat Riset dan Pengabdian Masyarakat Universitas Indonesia). We would also like to thank Desmond Darma Putra for help in computing the Fleiss kappa values in Section 6.3.

References

[1] E. Agirre and P. Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2007.

[2] K. Clodfelder. An lsa implementation against parallel texts in french and english. In *Proceedings of the HLT-NAACL Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 111–114, 2003.

[3] Y. Deng and Y. Gao. Guiding statistical word alignment models with prior knowledge. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1–8, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[4] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[5] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[6] G. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(2):205–224, 1965.

[7] D. Kalman. A singularly valuable decomposition: The svd of a matrix. *The College Mathematics Journal*, 27(1):2–23, Jan 1996.

[8] T. Landauer. *Handbook of Latent Semantic Analysis*. Routledge, 2007.

[9] T. Landauer, P. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[10] B. Rehder, M. Littman, S. Dumais, and T. Landauer. Automatic 3-language cross-language information retrieval with latent semantic indexing. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, pages 233–239, 1997.

[11] S. Sari. Perolehan informasi lintas bahasa indonesia-inggris berdasarkan korpus paralel dengan menggunakan metoda mutual information dan metoda similarity thesaurus. Master's thesis, Faculty of Computer Science, University of Indonesia, 2007. Call number: T-0617.

[12] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998.