A study in explaining unseen words in Indonesian using analogical clusters

Rashel Fam Yves Lepage Susanti Gojali Ayu Purwarianti

Graduate School of IPS, Waseda University *Institut Teknologi Bandung*

fam.rashel@fuji.waseda.jp yves.lepage@waseda.jp susantigojali@hotmail.com ayu@informatika.org

Abstract

We propose a pipeline to explain, on the level of form, the unseen words contained in an Indonesian test set, by using analogical clusters. Analogical clusters are extracted from a training set by relying on formal relations between words. The unseen words which can be explained on the level of form are then verified on two other representation levels: morphological and semantic. In our experiments on the BPPT corpus, 98 % of unseen words were explained on the level form, out of which 58 % could also be explained on the two levels of morphological and semantic representations.

1. Introduction

The problem of unseen words, or out-of-vocabulary (OOV) words, or new words, is one of the big issues in natural language processing (NLP). This is the case for tasks such as speech recognition or machine translation. The vocabulary of an NLP system is usually limited by the words learnt by the system in the preliminary step, for example, that of extracting knowledge from a training corpus.

This paper addresses the issue of predicting unseen words. Given an unseen word, how to find all other words in the known vocabulary which may explain it. We consider computational analogy as a possible way to answer this problem. For example *inexhaustivity* may be explained by the following three words in the following manner: $active : inactivity :: exhaustive : x \quad \Rightarrow \quad x = inexhaustivity$. Previous work, like [8], focuses on the formal aspect of the problem. They perform experiments on 12 different languages using translations of the Bible to explain unseen words by using paradigm tables. On the contrary, works like [9] or [7] take into consideration the meaning of words. In the present paper, we first explain words on the formal level. We then confirm the explanation on the level of form by checking it on two other levels: morphological representation and semantic representation. We choose to specifically work on Indonesian as it is a language known for its relative richness in derivational morphology.

The paper is organized as follows: Section 2 introduces our method to produce analogical clusters. Section 3 presents a survey on the data we used to carry our experiments. Section 4 explains the experimental protocol. Section 5 presents and analyzes the results obtained in the experiments. Section 6 gives conclusions.

2. Explaining unseen words

In the next following sections, we introduce our method to produce analogical clusters by using the notion of computational analogy between strings of symbols proposed in [3].

2.1. Word ratios

We define the ratio between two words as a vector of features made of all the differences in the two words in number of occurrences of all characters, whatever the writing system, plus the edit distance between the two words. The following formula explains the ratio between two words A and B.

$$A : B \triangleq \begin{pmatrix} |A|_a - |B|_a \\ \vdots \\ |A|_z - |B|_z \\ \mathrm{d}(A, B) \end{pmatrix} \qquad (1)$$

The notation $|S|_c$ stands for the number of occurrences of character c in string S. The last dimension, written as $\mathrm{d}(A, B)$, is the edit distance between the two strings. The edit distance is computed using only two edit operations: insertion and deletion. It indirectly gives the number of common characters appearing in the same order in A and B. This definition of word ratios captures prefixing and suffixing and more generally infixing. However, this definition does not capture reduplication nor repetition. The latter one would be needed to capture marked plurals in Indonesian, for instance *meja-meja* is to *meja* in Indonesian as ‘tables’ is to ‘table’ in English. Figure 1 shows the word ratio for the Indonesian words *makan* and *makanan* which can be translated into ‘to eat’ and ‘food’ in English.

$$\text{makan} : \text{makanan} \triangleq \begin{pmatrix} -1 \\ \vdots \\ 0 \\ 2 \end{pmatrix}$$

Figure 1. Word ratio for *makan* and *makanan*

The above definition of is found in the characterization of the notion of proportional analogy in [1] or [2]. There, proportional analogy is defined as a relationship between four objects where two properties are met: (a) equality of ratios between the first and the second terms on one hand and the third and the fourth terms on the other hand, and (b) exchange of the means. The exchange of the means states that the second and the third terms can be exchanged in a proportional analogy. Formula (2) gives the notation and the definition of a proportional analogy.

$$A : B :: C : D \overset{\triangle}{\Leftrightarrow} \begin{cases} A : B = C : D \\ A : C = B : D \end{cases} \qquad (2)$$

2.2. Analogical clusters

We compute all ratios and group pairs of words by equal ratio. A set of pairs of words with the same ratio is called an analogical cluster. Using the Formula (2), we define an analogical cluster in Formula (3). Notice that the order of word pairs in analogical clusters has no importance.

$$\begin{matrix} A_1 : B_1 \\ A_2 : B_2 \\ \vdots \\ A_n : B_n \end{matrix} \overset{\triangle}{\Leftrightarrow} \begin{matrix} \forall (i, j) \in \{1, \dots, n\}^2, \\ A_i : B_i :: A_j : B_j \end{matrix} \qquad (3)$$

Practically, it would be too long to compute all possible ratios between all pairs of words directly so that a strategy in two steps is adopted following a method proposed in [10]. First, the set of all words is represented by a tree where each level stands for a character. All the words in the text are hierarchically grouped by their number of occurrences of characters. Then, a top-down exploration of the tree against itself is performed to group pairs of words by equal difference of number of occurrences of characters. We then test for equality of distance for each word pair. This may split the group of word pairs into smaller groups where all word pairs in a group will have the same ratio. Finally, for each group of word pairs with equal ratio, we test for equality in edit distance vertically, $A_i : B_i = A_j : B_j$, for any pair of word pairs (i, j). This, again, may split the group into smaller ones (see Algorithm 1).

Algorithm 1. Extracting set of analogical clusters from set of words

```
function BUILD_CLUSTERS(set of words)
   tree ← from the set of words    ◃ Hierarchically group words by their
                                     ◃ number of occurrences of characters.
   repeat top-down exploration of the tree against itself
      group pairs of words by equal difference
         of number of occurrences of characters
   until last character
   for all set of word pairs with equal number of occurrences of characters do
      CHECK_DISTANCE(set of word pairs)

function CHECK_DISTANCE(set of word pairs $(A_1, B_1), \ldots, (A_n, B_n)$)
   for all i ∈ $\{1, \ldots, n\}$ do
      compute d($A_i$,$B_i$) end for
   for all set of word pairs ($A_i$,$B_i$) with same distance do
      CHECK_CLUSTER(set of word pairs)

function CHECK_CLUSTER(set of word pairs $(A_1, B_1), \ldots, (A_n, B_n)$)
   $V \leftarrow \{1, \ldots, n\}$                          ◃ Vertices of the graph.
   $E \leftarrow \{ (i, j) \in V^2 \ / \ A_i : B_i = A_j : B_j \}$          ◃ Edges of the graph.
   list ← nodes in $V$ sorted by non-increasing number of edges
   not_yet_covered ← V
   repeat
      i ← first node in list
      delete i from list
      if i ∈ not_yet_covered then
         clique ← {i} ◃ Initialize clique to singleton of not yet explored vertex.
         clique, not_yet_covered ← EXPAND_CLIQUE(clique, list, not_yet_covered)
         return clique                     ◃ clique is an analogical cluster.
until not_yet_covered = ∅

function EXPAND_CLIQUE(clique, list, not_yet_covered)
   for all i in list do
      if i is connected with all vertices in the clique then
         add i to the clique                ◃ Remains a clique.
         delete i from not_yet_covered
return clique, not_yet_covered
```

Individual analogical clusters give some insight of the organization of the lexicon of a language. The top left cluster in Figure 2 shows an analogical cluster which consists of three ratios. This analogical cluster illustrates the morphological phenomenon in Indonesian where nouns are derived from active verbs, e.g *makan* 'to eat', by suffixing *–an*. The derived nouns, e.g *makanan* 'food' are the object of the respective verbs. As for the top right cluster in Figure 2, it illustrates the phenomenon of deriving passive verbs, e.g *dimakan* (be eaten) from active verbs, e.g. *makan* 'to eat' by prefixing *di-*. The bottom cluster is an analogical cluster which consists of only two ratios. As mentioned on the previous section, because it has only two ratios, this is a proportional analogy.

makan : makan**an**
minum : minum**an**
main : main**an**

makan : **di**makan
minum : **di**minum
beli : **di**beli

makan : **me**makan
minum : **me**minum

Figure 2. Examples of analogical clusters

3. Data used

We carried out experiments using the BPPT corpus provided by PAN Localization[1]. BPPT is an Indonesian-English aligned parallel corpus of news articles. It contains almost half million tokens (words in the corpus) representing twenty-seven thousand types (number of different words). The average length of a token is around six characters while the average length for types is almost eight characters. Less than half of the tokens are hapaxes, i.e, words which appear only once in the corpus. Table 1 shows the statistics of the BPPT corpus.

Table 1. Statistics of BPPT corpus

Number of tokens	486,936
Avg length of tokens	6.199
Number of types	27,315
Avg length of types	7.946
Number of hapaxes	44,309
Type-Token Ratio	0.056

4. Experiments

The goal of the experiments is first to explain unseen words by analogy on the level of form. We then confirm the analogies on the level of form by checking them on two other levels of representation: morphological and semantic. In this section, we present the experimental protocol that we used, how we build the morphological and semantic representation, and how we check the analogy on these levels.

4.1. Experimental protocol

We shuffle the sentences in the BPPT corpus and divide the corpus into two parts: 90% for training set and 10% for test set. We have 1,276 unseen words out of 8,629 words in the test set (less than 15%). For each unseen word in the test set, we extract all possible analogical clusters which include it. After that, we confirm the validity of the ratios in the analogical clusters on two other levels of representation: morphological and semantic. In this way, we count how many unseen words can be explained on the three levels at the same time: form, morphological representation, and semantic representation.

Table 2. Number of types

	Number of types
Training set	26,039
Test set	8,629
Unseen words	1,276

4.2. Unseen words

A rough characterization and estimation of the categories of unseen words was conducted. Hundred unseen words were sampled out of all unseen words and classified by hand. Although there are some abbreviations, foreign words, and typos, unseen words are mainly proper nouns and valid Indonesian words. Of course, valid Indonesian words are of greater interest than the other categories. Table 3 shows the categories and the number of unseen words in each category.

[1] http://www.panl10n.net/indonesia/

Table 3. Categories of unseen words

Category	Percentage (%)	Example	Description
Abbreviation	4	BBMI	stock exchange symbol for *Bank Muamalat Indonesia*
Foreign word	8	squirrel	English word
Proper noun	32	Alexandr	Person's name
Typo	13	pemeliti	for *peneliti* \`researcher\`
Valid Indonesian	43	ibu-ibu memayungi	marked plural of *ibu* \`mother\` *meN+payung+i* \`to shelter with an umbrella\`
Total	100		

4.2. Morphological representation

We use a stemmer [4] and an HMM-based part-of-speech tagger [5] for Indonesian to obtain the morphological representation for each word. Each word is represented by its lexeme and exponent(s) which construct the word, accompanied with its part-of-speech tag.

$$\mathit{makan\_VB} : \mathit{makan{+}an\_NN} :: \mathit{minum\_VB} : \mathit{minum{+}an\_NN}$$

Figure 3. Confirming the analogy on the level of morphological representation

To verify a proportional analogy on the level of morphological representations, we verify the proportional analogy on the strings of the representations themselves using Formula (2).

We verify analogies on the level of morphological representation for at most 30 ratios in each analogical cluster that includes an unseen word. If more than 50% (at least 15 ratios out of 30 ratios) of the analogies are verified, we assume that it is sufficient to state that the analogy on the morphological level holds for that cluster. We consider that it is sufficient that one analogical cluster pass the previous criterion to explain an unseen word on the level of morphological representation.

4.3. Semantic representation

Linguistic regularities can also be captured in continuous word representations [6][7] where words are represented in a vector space to perform tasks such as solving semantic analogical equations. As a famous example, the vector for *queen* can be approximated by summing the vectors for *king* and *woman* and subtracting the vector for *man*: $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$.

$$\overrightarrow{makan} - \overrightarrow{makanan} + \overrightarrow{minum} \approx \overrightarrow{minuman}$$

Figure 4. Confirming the analogy on the level of semantic representation

We train a model for all the words contained in the same corpus. The vector dimension is 300 with a window size of 5 words on the left and on the right of the current word. For the ratios in our analogical clusters, we solve the analogical equations using vectors and check whether the unseen word comes out as an answer. In this experiment, we take the top 100 answers for each equation. If the unseen word comes out at least once in the top 100 answers, we consider that the analogy holds on the level of semantic representation.

5. Results

In our experiments, 98 % of the unseen words were explained on the level of form. It means that for 98 % of the unseen words, we found at least one analogical cluster that includes the unseen word. The remaining 2 % unexplained unseen words are mostly marked plurals. This confirms what we stated in Section 2.1 concerning the power of our definition of word ratio. Table 4 shows the exact number of unseen words explained on the levels of form, morphological representation, and semantic representation.

We turn now to morphological representation. We transformed the ratios in the analogical clusters extracted on the level of form into their morphological representations, and verified the analogies against the criterion mentioned in Section 4.2. Up to 81 % of the unseen words explained on the level of form were also explained on the level of morphological representation. Most of these explained unseen words have a part-of-speech tag which show a derivational phenomenon, such as verbs, adjectives, and nouns. The remaining unexplained 19 % are mostly marked plurals, typos and nouns.

For the semantic representation, 63 % out of the 98 % unseen words explained on the level of form, were explained on the level of semantic representation. Further experiments with more parameters varying when building vector models may led to different results. In this experiment, we trained the word models on our relatively small training set. The use of a larger corpus, like the Indonesian Wikipedia, would have surely led to higher percentages.

Table 4. Number of unseen words explained

Form	Morphology	Semantics	Total
✓			1,249
✓	✓		1,010
✓		✓	791
✓	✓	✓	724
			27

Table 4 shows the accumulated results of how many unseen words which were explained on the level of form can be explained on the two additional levels of morphological and semantic representations. In summary, 58 % out of 98 % unseen words explained on the level of form can be explained on these additional two levels.

Table 5. Examples of unseen words explained or not on each level of representation

Form	Morphology	Semantics	Number	Examples	English translation

Yes	No	No	172	*terenggut*
				ilustrasi
				Montolivo
					`wrenched`
					`illustration`
					person's name

Yes	Yes	No	286	*bercampur*
				disewakan
				menyepakatinya
					`mixed`
					`rent`
					`to agree`

Yes	No	Yes	67	*perfeksionis*
				endoplasma
				radjawali
					`perfectionist`
					`endoplasm`
					name of a kind of bird

Yes	Yes	Yes	724	*terkoordinasi*
				persilangan
				pembelajaran
					`coordinated`
					`cross`
					`learning`

6. Conclusions

We proposed a pipeline to predict unseen words. On the level of form, we explained unseen words contained in a test set by using analogical clusters extracted from a training set. The method relies on the formalisation of a relationship between words used in works dealing with proportional analogy.

We performed experiments on the Indonesian language with the BPPT corpus. Our experimental results gave a high percentage of 98 % of unseen words explained on the level form. Further verification showed that 58 % out of these 98 % unseen words explained on the level of form can also be explained on two other levels of morphological and semantic representations.

Because our method works on the level of form, it is in practice language-independent. It would be interesting to perform similar experiments on different languages to compare the results. Comparing similarities and differences in results could lead to interesting conclusions across languages.

Acknowledgements

This work was supported by a grant from the Japanese Society for the Promotion of Science (JSPS): grant number 15K00317 entitled ‘Language productivity: fast extraction of productive analogical clusters and their evaluation using statistical machine translation.’

References

[1] Philippe Langlais and Francois Yvon. 2008. Scaling up analogical learning. In *Coling 2008: Companion volume: Posters*, pages 51–54, Manchester, UK, August. Coling 2008 Organizing Committee.

[2] Nicolas Stroppa and Francois Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 120–127, Ann Arbor, Michigan, June. Association for Computational Linguistics.

[3] Yves Lepage. 2004. Lower and higher estimates of the number of “true analogies” between sentences contained in a large multilingual corpus. In *Proceedings of COLING-2004*, volume 1, pages 736–742, Gene`ve, August.

[4] Ayu Purwarianti. 2011. A non-deterministic Indonesian stemmer. In *International Conference on Electrical Engineering and Informatics (ICEEI-2011)*, pages 1–5. IEEE.

[5] Alfan Fariski Wicaksono and Ayu Purwarianti. 2010. HMM based part-of-speech tagger for bahasa Indonesia. In *Fourth International MALINDO Workshop, Jakarta*.

[6] Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR, abs/1301.3781*.

[7] Mikolov, T., Yih, W.-T., and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

[8] Rashel Fam and Yves Lepage. 2016. Morphological predictability of unseen words using computational analogy. In *Proceedings of Workshops on Computational Analogy at The Twenty-Forth International Conference on Case-Based Reasoning (ICCBR 2016)*, pages 51–60, Atlanta, Georgia.

[9] Tekauer, P. 2005. *Meaning Predictability in Word Formation: Novel, Context-free Naming Units*. John Benjamins Publishing.

[10] Yves Lepage. 2014. Analogies between binary images: Application to Chinese characters. In *Prade, H., Richard, G. (eds.) Computational Approaches to Analogical Reasoning: Current Trends*, pages 25–57, Springer, Berlin, Heidelberg (2014).