Plausibility of Word Forms Generated from Analogical Grids in Indonesian

Rashel Fam[†], Ayu Purwarianti[*], Yves Lepage[+]

[†+]*Graduate School of IPS Waseda University, Fukuoka, Japan*

[*]*Bandung Institute of Technology, Bandung, Indonesia*

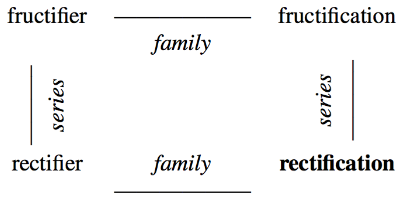
[†]fam.rashel@fuji.waseda.jp, [*]ayu@informatika.org, [+]yves.lepage@waseda.jp

Abstract

The vocabulary of a natural language processing (NLP) system is usually limited by the word forms learnt by the system in the preliminary step, for example, word forms seen in the training corpus. Thus, out-of-vocabulary (OOV) problem is an important issue in NLP. In this paper, we study the plausibility of unseen word forms generated from analogical grids on Indonesian, a language known for its richness in derivational morphology. We construct analogical grids from a list of word forms contained in an annotated Indonesian corpus. We generate new word forms by filling the empty cells in the analogical grids. We verify these generated word forms using morphological analyzer and count how many of them are valid Indonesian word forms.

1. Introduction

Analogical grids give a compact view of the organization of a lexicon in a language up to certain extent. Such analogical grids can be built from a list of word forms contained in a corpus. They can be used to study the productivity of a language such as in [1], [2] and [3]. [4] performed such an analysis across 12 languages using analogical grids built from the Bible corpus. As another example of use, [5] showed how to produce the French word form *rectification* by analogy from the neighboring word forms *fructifier*, *fructification*, and *rectifier* in the same series (see **Figure 1**).



fructifier : *fructification* :: *rectifier* : *rectification*

Figure 1. Producing a new word form from neighboring word forms. Example taken from [5]

In this paper, we compare the productivity of analogical grids built from two different kinds of feature vector. We generate new word forms by filling the empty cells in the grids. The plausibility of these newly generated word forms is tested against morphological analyzers. We count how many of them are recognized by at least one morphological analyzer.

The paper is organized as follows: Section 2 introduces some notions related to production of analogical grids. Section 3 explains the data used in this work. Section 4 presents the experiment and results obtained. Section 5 gives the conclusion.

2. Analogical Grids

makan	:	memakan	:	dimakan	:	makanan
minum	:	meminum	:	diminum	:	minuman
main	:		:		:	mainan
beli	:		:	dibeli	:	
lihat	:	melihat	:	dilihat	:	

Figure 2. An analogical grid in Indonesian

An analogical grid is a matrix of word forms where four word forms from two rows and two columns are a proportional analogy. Equation (1) gives the definition of an analogical grid. Such analogical grid can be built from list of word forms contained in a corpus using the methods previously presented in [6], [7] and [4].

$$\begin{matrix} P_1^1 & P_1^2 & \cdots & P_1^m \\ P_2^1 & P_2^2 & \cdots & P_2^m \\ \vdots & & & \vdots \\ P_n^1 & P_n^2 & \cdots & P_n^m \end{matrix} \overset{\Delta}{\Leftrightarrow} \begin{matrix} \forall(i,k) \in \{1,\dots,n\}^2, \\ \forall(j,l) \in \{1,\dots,m\}^2, \\ P_i^j : P_i^l :: P_k^j : P_k^l \end{matrix} \qquad (1)$$

Analogy is defined from feature vectors representing word forms, through equality of ratios. A ratio is the difference between two feature vectors plus the edit distance between the word forms as defined in Equation (2). We refer the reader to [4] for further details.

$$A : B \triangleq \begin{pmatrix} |A|_a\text{-}|B|_a \\ \vdots \\ |A|_z\text{-}|B|_z \\ \mathrm{d}(A,B) \end{pmatrix} \qquad (2)$$

The size of an analogical grid is defined as its number of rows multiplied by its number of columns (See Equation (3)). For instance, the analogical grid in **Figure 2** has a size of 5 × 4 = 20.

$$\text{Size} = \text{Number of rows} \times \text{Number of columns} \qquad (3)$$

The number of empty cells of an analogical grid can be roughly considered the number of possible forms not present in the data. We call it *saturation*, the ratio of the number of non-empty cells and the size of an analogical grid in percentage. According to Equation (4), the analogical grid in **Figure 2** has 5 empty cells; its saturation is thus: 75%.

$$\text{Saturation} = 100 - \frac{\text{Number of empty cells} \times 100}{\text{Total number of cells}} \qquad (4)$$

It is worth noticing that, when creating all possible analogical grids from a text, not all of the word forms will necessarily appear in analogical grids. Reciprocally, analogical grids extracted from texts may contain empty cells. An analogical grid which does not contain any empty cell is not productive as no new word form can be entered in it. On the contrary, we will call any analogical grids which contains at least one empty cell a *productive analogical grid*. We will call a word form that may fill a blank cell in a productive analogical grid a *generated word form*. We can test the validity of these generated word forms by testing them against morphological analyzer.

2.1. Character-Based Word Vector Representation

A word form is converted into a vector of features which are the number of occurrences for all the characters in the alphabet. For instance, in lowercase Indonesian, the dimension of the vector is 26 (from a to z) as illustrated in Equation (5). The notation $|S|_c$ stands for the number of occurrences of character c in string S.

$$A = \begin{pmatrix} |A|_a \\ |A|_b \\ \vdots \\ |A|_z \end{pmatrix} \qquad (5)$$

According to Equation (5) the Indonesian word, *makan*, which means *to eat* in English, will have the following vector representation.

$$\text{makan} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

2.2. Richer Word Vector Representation

The previous feature vector is constructed based on the character information of the word form only. It is possible for us to define richer vector representation. For instance, we can embed more information about the word form, such as part-of-speech (POS) tag, as additional features of the vector. Equation (6) shows how to convert the POS tag into integer values inside the feature vector for word form *makan* which is a verb (VB).

$$A = \begin{pmatrix} |A|_a \\ |A|_b \\ \vdots \\ |A|_z \\ is\_NN(A) \\ is\_VB(A) \\ \vdots \\ is\_ADJ(A) \end{pmatrix} \qquad (6)$$

Using such feature vectors will give more constraint and lead to producing regular paradigm tables. Now, we will have the following vector representation for the Indonesian word, *makan*. Please compare with the previous character-based vector representation at Section 2.1.

$$\text{makan} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

3. Data used

We carried out experiments using the first thousand sentences of idn-tagged[1] corpus. This corpus is an effort by [8] to manually annotate the BPPT corpus provided by PAN Localization[2]. BPPT corpus is an Indonesian-English aligned parallel corpus of news articles. In this work, we will only use the Indonesian part. It contains almost thirty thousand tokens (word forms in the corpus) representing almost five thousand types (number of different word forms). The average length of a token is around five characters while the average length for types is almost seven characters. More than half of the tokens are hapaxes, i.e, word forms which appear only once in the corpus. **Table 1** shows the statistics of the corpus.

Table 1. Statistics of the first thousand sentences of idn-tagged corpus

Number of tokens	27,545
Avg. length of tokens	5.17±2.95
Number of types	4,768
Avg. length of types	6.73±2.84
Type-token ratio	0.17
Hapax	53.48

4. Experiment and Results

Each word form from the first thousand sentences of idn-tagged corpus is converted into two different feature vectors, the standard character-based and richer representation with POS information. We then produce the analogical grids with different saturation

[1] github.com/famrashel/idn-tagged-corpus

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

Table 2. Statistics of analogical grids obtained

	Saturation threshold (%)	# analogical grids produced	# productive analogical grids	Average size	Average saturation (%)
char	0	951	149	8,745	65.56
	50	2,621	1,366	54	52.83
	90	5,825	137	12	91.73
char+POS	0	370	63	3,548	68.64
	50	625	204	47	63.04
	90	1,133	17	13	91.47

thresholds and generate new word forms from the empty cells inside the grids. The plausibility of these newly generated word forms is then tested against two Indonesian morphological analyzers, MorphInd [9] and InaNLP [10, 11]. We assume that a word form is valid if at least one of the analyzer is able to recognize the word form.

4.1. Productive Analogical Grids Obtained

Table 2 shows the statistics of the analogical grids obtained from the two configurations used, character-based feature vector (char) and richer feature vector (char+POS). We controlled the saturation of the analogical grids during their production above some thresholds (no threshold, 50% and 90%). As we are interested with the generated word forms, we left out non-productive analogical grids. Productive analogical grids have saturation such that: threshold $\leq$ saturation $<$ 100.

In all settings, we obtained more analogical grids from character-based feature vector representation than richer feature vector representation. This meets intuition because it is less constrained. These analogical grids also tend to have bigger size but they are hollower as their saturation are lower.

We now turn into the saturation threshold of the production of the analogical grids. Without setting the saturation threshold, we produced bigger analogical grids with more empty cells in it. We produce the most productive analogical grids with the saturation threshold of 50%. The least productive analogical grids are produced under saturation threshold of 90%. As we controlled the analogical grids to be almost filled, with this setting, we usually ended up with non-productive analogical grids. If we also count the non-productive analogical grids, we obtained the most number of analogical grids with saturation threshold of 90%. However, only around 2% of the analogical grids produced are productive analogical grids.

4.2. Word Forms Generated from Productive Analogical Grids

Table 3 shows the results of newly generated word forms from different kind of analogical grids. As expected, we generate a lot more word forms under character-based setting. This is caused by having more and bigger productive analogical grids. We are able to generate 8 times more word forms in compare to richer feature vector representation.

Table 3. Number of word forms generated from productive analogical grids obtained

	Saturation threshold (%)	# generated word forms	Valid generated word forms	
			Total	Ratio
char	0	161,338	14,406	9%
	50	15,886	3,073	19%
	90	93	26	28%
char+POS	0	22,876	4,290	19%
	50	2,401	1,023	43%
	90	16	8	50%

Although we produced more productive analogical grids under saturation threshold of 50%, the most number of generated word forms are obtained from analogical grids built with no saturation threshold. This is due to the advantage of bigger size of analogical grids even when the saturation is slightly higher. These analogical grids have more empty cells and thus generate more word forms. The number of generated word forms is around ten times more than the analogical grids built with saturation threshold of 50% and above thousand times more in compare to 90% saturation threshold. We generated the least number of word forms with saturation threshold of 90% because we have less empty cells and less productive analogical grids.

For the plausibility of the generated word form, the use of higher saturation threshold leads to higher ratio of valid word forms. As analogical grids are almost complete (small number of empty cells), we have more evidences from the neighboring cells which rise the confidence to fill in the empty cells.

From the point of view of feature vector, the ratio of valid word forms is obtained when using productive analogical grids built from richer feature vector. The highest one is when using saturation threshold of 90%: half of the generated word forms are recognized by morphological analyzer. In summary, the analogical grids built from richer feature vector are more reliable to produce plausible word form than character-based feature vector. The difference is up to two times better.

4.3. Categories of Valid Word Forms Generated from Productive Analogical Grids

We sampled one hundred generated word forms which can be recognized by at least one of the morphological analyzers and annotated them by hand. Table 4 shows that 73 word forms are real Indonesian words, like:

- verbs (43 word forms),
- adjectives (4 word forms), and
- nouns (27 word forms).

Some of them are word forms derived from adding prefix or suffix which is meets our intuition.

22 word forms are categorized as *unknown*. They are consisted of invalid word forms and potential new word forms. The invalid word forms come from the confusion in parsing the word form due to the limitation of the morphological analyzers. However, we also observed several word forms that annotator never seen before but seems to be valid and understandable by looking at the outputs of the morphological analyzers. Therefore, we think of these word forms as potential new word forms generated by the analogical grids.

The rest of the word forms are foreign words.

Table 4. Distribution of generated word forms is recognized by morphological analyzers.

Category	Number	Example	Description
Verb	43	*mencerminkan*	*meN* + ***cermin***<**noun**> + *kan*: 'to reflect'
Adjective	4	*putus*	'cut off'
Noun	27	*proyek*	'project'
Foreign words	4	*use*	English word
Unknown	22	*terp*	*ter* + ***p***<**noun**>: Invalid form
		berpelebar	*ber* + *peN* + ***lebar***<**adjective**>: Potential new form?
		termengerti	*ter* + ***mengerti***<**adjective**>: Potential new word form?
	100		

5. Conclusions

We constructed analogical grids from the first thousand sentences of idn-tagged corpus using two different kinds of feature vector as word form representation, character-based with and without part-of-speech information. We generated new word forms from empty cells in the analogical grids. The plausibility of these generated word forms is evaluated using morphological analyzers.

Results show that we are able to generate a lot more word forms under the character-based feature vector representation. In terms of saturation threshold, we generated more word forms with lower threshold. This meets our intuition because we have less constraints to satisfy and thus obtained analogical grids with more empty cells to fill.

However, we produce more valid word forms when using the richer feature vector representation. The additional information, part-of-speech of the word form, concatenated in the feature vector were proved to put more constrains when building the analogical grids. The analogical grids are then forced to generate a more plausible word form from each of its empty cells. Results show that the chance to generate a valid word form is up to two times higher when using the richer word feature vector in compare to character-based feature vector.

Acknowledgments

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

References

[1] Singh, R. and Ford, A., "In praise of Sakatayana: some remarks on whole word morphology", *In Rajendra Singh, editor, The Yearbook of South Asian Languages and Linguistics-200*, Sage, Thousand Oaks, 2000.

[2] Neuvel, S. and Fulop, S. A., "Unsupervised learning of morphology without morphemes", *In Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pp 31–40, Association for Computational Linguistics, July 2002.

[3] Hathout, N., "Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy", *In Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing*, pp 1–8, Manchester, UK, August, Coling 2008 Organizing Committee, 2008.

[4] Fam, R. and Lepage, Y., "Morphological predictability of unseen words using computational

analogy", *In Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-CA-16)*, pp 51–60, Atlanta, Georgia, 2016.

[5] Hathout, N., "Acquisition of morphological families and derivational series from a machine readable dictionary", 2009, CoRR, abs/0905.1609.

[6] Lepage, Y., "Solving analogies on words: an algorithm", *In Proceedings of the 17th international conference on Computational linguistics (COLING 1998), volume 1*, pp 728–734, Association for Computational Linguistics, 1998.

[7] Lepage, Y., "Analogies between binary images: Application to Chinese characters", *In Henri Prade et al., editors, Computational Approaches to Analogical Reasoning: Current Trends*, pp 25–57, Springer, Berlin, Heidelberg, 2014.

[8] Dinakaramani, A., Rashel, F., Luthfi, A. and Manurung, R, "Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus", *In Proceedings of the International Conference on Asian Language Processing (IALP 2014)*, pp 70-73, Kuching, Malaysia, October 2014.

[9] Larasati, S., Kuboň, V. and Zeman, D., "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus", *In Proceedings of the Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011),* Springer CCIS*,* Zurich, Switzerland, 2011.

[10] Wicaksono, A. and Purwarianti, A., "HMM based part-of-speech tagger for bahasa Indonesia", *In Fourth International MALINDO Workshop*, Jakarta, 2010.

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