Validating analogically generated Indonesian words using Fisher's exact test

Rashel Fam　　Yves Lepage
早稲田大学大学院情報生産システム研究科

fam.rashel@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract

We address the issue of generating previously unseen word forms by filling empty cells inside analogical grids. We verify the plausibility of these generated word forms using a morphological analyzer and count how many of them are valid word forms. In this paper, we compare the ratio of valid word forms generated from analogical grids constructed from the list of word forms contained in an annotated Indonesian corpus. We use different word vector representations and saturation thresholds to construct the analogical grids and compare with the use of Fisher's exact test to measure the confidence of filling an empty cell. The experimental results show that although the confidence is low in general, using Fisher's exact test gives twice more confidence when generating validated word forms.

1 Introduction

makan	:	***di**makan*	:	***me**makan*	:	*makan**an***
minum	:	***di**minum*	:	***me**minum*	:	*minum**an***
main	:		:		:	*main**an***
beli	:	***di**beli*	:		:	

Figure 1: An analogical grid in Indonesian

Figure 1 shows an analogical grid in Indonesian. It gives a compact view of how the lexicon in the language are organized, up to some extent. Previous works, like [9, 7, 4], shows how to use such table to study word productivity in a given language. [2] reported experiments in various languages on predicting previously unseen word forms in a test set by constructing analogical grids from list of words contained in a training corpus.

In this paper, we address the issue of filling empty cells inside analogical grids to generate previously unseen word forms. We study the number of newly generated word forms obtained from filling empty cells inside the analogical grids built from different word vector representations and saturation thresholds. We also perform a Fisher's exact test to measure the confidence of filling an empty cell.

The paper is organized as follows: Section 2 introduces basic notions about analogical grids and Fisher's exact test. Section 3 presents a survey on the data we used to carry our experiments. Section 4 explains the experimental protocol and experimental results. Section 5 gives the conclusions on our work.

2 Basic notions

2.1 Analogical grids

An analogical grid is a matrix of word forms where any four words from two rows and two columns are a proportional analogy. Formula (1) gives the definition of an analogical grid.

$$\begin{array}{c} P_1^1 : P_1^2 : \cdots : P_1^m \\ P_2^1 : P_2^2 : \cdots : P_2^m \\ \vdots \quad \vdots \qquad\quad \vdots \\ P_n^1 : P_n^2 : \cdots : P_n^m \end{array} \quad \overset{\Delta}{\iff} \quad \begin{array}{c} \forall (i,k) \in \{1,\ldots,n\}^2, \\ \forall (j,l) \in \{1,\ldots,m\}^2, \\ P_i^j : P_i^l :: P_k^j : P_k^l \end{array} \tag{1}$$

The size of an analogical grid is simply the total number of cells inside the matrix. The saturation of an analogical grid is defined as the ratio of non-empty cells to the size of the analogical grid.

2.2 Word vector representation

Each word in an analogical grid is represented as a vector of features. The features are basically free to determine. In this paper, we consider to use two kinds of features for the word vector representation:

- characters-only;
- characters + part-of-speech (POS).

2.2.1 Characters-only feature vector

For the characters-only feature vector, we only consider the number of occurrences of all characters in the alphabet as shown in Formula 2. Here, the notation $|A|_c$ stands for the number of occurrences of character c in string A.

$$A = \begin{pmatrix} |A|_a \\ |A|_b \\ \vdots \\ |A|_z \end{pmatrix} \qquad \begin{matrix}\text{makanan} \\ \text{'food'}\end{matrix} = \begin{pmatrix} 3 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2}$$

2.2.2 Characters + POS feature vector

In contrast with characters-only feature vector, we add more information about the word form as features in the vector (e.g. the part-of-speech, case, tense, etc). Such information are available from annotated data, like the Unimorph Project[1]. Formula 3 shows how we can embed the part-of-speech information as additional features in the word vector representation. In this paper, we decided to add the part-of-speech tag available in the annotated corpus, as additional features in the word vector.

$$A = \begin{pmatrix} |A|_a \\ \vdots \\ |A|_z \\ is\_VB(A) \\ is\_NN(A) \\ \vdots \\ is\_ADJ(A) \end{pmatrix} \qquad \begin{matrix}\text{makanan} \\ \text{'food'}\end{matrix} = \begin{pmatrix} 3 \\ \vdots \\ 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \tag{3}$$

2.3 Proportional analogy

Proportional 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. Formula 4 shows the definition of ratio between two word forms A and B which are represented by characters-only feature vector. We refer the reader to [2] for further details.

$$A : B \quad \triangleq \quad \begin{pmatrix} |A|_a - |B|_a \\ \vdots \\ |A|_z - |B|_z \\ \mathrm{d}(A, B) \end{pmatrix} \tag{4}$$

The above definition is found in or implied by the characterization of the notion of proportional analogy between sequences of characters in [5] or [10]. From Formula 5 we can see that the term B and C are exchangeable. This is a property of proportional analogy called exchange of the means.

$$A : B :: C : D \overset{\Delta}{\Longleftrightarrow} \begin{cases} A : B & = \quad C : D \\ A : C & = \quad B : D \end{cases} \tag{5}$$

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

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

2.4 Empty cells inside analogical grids

Analogical grid that have at least one empty cell (saturation < 100 %) are productive analogical grid. Empty cells inside analogical grids are interesting because they can be filled by potential word forms. For example, the word form *belian* is a potential word form to fill the empty cell on the fourth row of the fourth column in Figure 1. It can be retrieved by solving the following analogical equation:

$$makan : makanan :: beli : x \qquad \Rightarrow \quad x = belian$$

2.5 Fisher's exact test

Fisher's exact test is a statistical test to analyze a contingency table. [3] showed that the hypergeometric distribution of the numbers in the tables can be used to calculate the significance of the observation from a null hypothesis. It is usually used for 2 x 2 contingency tables, but it is not limited to them.

[8] reported that Fisher's exact test is a more appropriate test to identify dependent word pairs in comparison to other statistical methods. Here, we use Fisher's exact test to measure the confidence of filling an empty cell. Before filling an empty cells P_i^j, we create a 2 x 2 table by observing the row$_i$ and column$_j$. The p-value p is calculated as follows.

	row$_i$	column$_j$
# non-empty cells	a	b
# empty cells	c	d

$$p = \frac{(a+b)! \quad (c+d)! \quad (a+c)! \quad (b+d)!}{a! \quad b! \quad c! \quad d! \quad (a+b+c+d)!} \tag{6}$$

3 Data used

We carried out experiments on the first thousand lines of idn-tagged-corpus[2] which is freely available. It is an effort described in [1] to manually annotate

[1] unimorph.org

[2] https://github.com/famrashel/idn-tagged-corpus

Feature vector	Saturation threshold (%)	# analogical grids produced	# productive analogical grids	Average size	Average saturation
char	≥ 50	2,621	1,366	54	52.83
	≥ 90	5,825	137	12	91.73
char + POS	≥ 50	625	204	47	63.04
	≥ 90	1,133	17	13	91.47

Table 2: Statistics of analogical grids obtained

the BPPT corpus[3], an Indonesian-English aligned parallel corpus of news articles. It contains around a quarter of a hundred thousand tokens (words in the corpus) representing around five thousand types (number of different words). More than half of the tokens (44.3 %) are hapaxes, which meets the intuition of being a news articles. Table 1 shows the statistics on the data.

4 Experiments

In this section, we present our experimental protocol which uses different word vector representations saturation threshold to construct analogical grids. We then investigate the performance of Fisher's exact test on measuring the confidence of filling the empty cells inside the analogical grids. We also present the results obtained on the data introduced in Section 3.

4.1 Experimental protocol

From the list of word form contained in the first thousand lines of idn-tagged-corpus, we extract all of the analogical grids by using two different word vector representations, characters-only and characters + POS feature vectors. For each of word vector representation, we construct the analogical grids while maintaining a saturation threshold. We choose to use 50 % and 90 % as our saturation threshold when building the analogical grids with intuition that empty cells in grids with higher saturation will be more reliable to fill. We then use Fisher's exact test to measure the confidence of filling an empty cell.

Finally, we count how many newly generated word forms are valid by testing them against MorphInd [6], a rule-based morphological analyzer for Indonesian. We define that a newly generated word form is a valid Indonesian word form if it can be recognized (parsed) by the morphological analyzer.

4.2 Analogical grids obtained

Table 2 shows the statistics of the analogical grids obtained from different configurations. We constructed more analogical grids (and also more productive grids) when using characters-only feature vectors. The characters-only feature vectors introduced more freedom when constructing the analogical grids. However, we obtained analogical grids with higher saturation when using the characters + POS feature vectors in average.

Higher saturation threshold produced a very small number of productive analogical grids. This is natural because we stressed more constraints when building the analogical grids which led us on producing complete and smaller analogical grids at the end.

4.3 Newly generated word forms

We can see the similar trend with the number of newly generated word forms. Analogical grids with higher saturation generates less number of new word forms because we have less and smaller productive analogical grids. This led us to less number of empty cells to fill. However, the ratio of plausible generated word forms are higher than using the lower saturation threshold.

From the point of view of feature vectors, we can see that using the characters-only feature vectors will give us analogical grids with more empty cells. Thus, we generate more new word forms but facing the drawbacks of producing more invalid word forms. On the contrary, characters + POS feature vectors deliver a smaller number of newly generated word forms but they are around twice better in terms of ratio of valid generated word forms. Table 3 shows the number of newly generated word forms obtained from filling analogical grids built under different configurations.

4.4 Fisher's test performance

As can be seen from Table 3, the use of Fisher's exact test under the condition of p-value ≤ 5 % gives 29 % (= 24/82) and 65 % (= 11/17) ratio of validated generated word forms. It is around two times better performance in comparison to the configuration without Fisher's exact test, 15 % (= 2,426/15,886) and 38 % (= 904/2,401). The p-value from Fisher's exact test leads to be very cautious in filling the empty cells.

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

Feature vector	Saturation threshold (%)	# empty cells	# generated word forms			# valid generated word forms		
			w/o Fisher	w/ Fisher		w/o Fisher	w/ Fisher	
				$p \leq 5$ %	$p > 5$ %		$p \leq 5$ %	$p > 5$ %
char	≥ 50	34,914	*15,886	**82**	15,824	*2,426	**24**	2,409
	≥ 90	140	93	0	92	23	0	23
char + POS	≥ 50	4,313	*2,401	**17**	2,387	*904	**11**	897
	≥ 90	19	16	0	15	7	0	7

Table 3: Number of newly generated word forms

A consequence of that is that no empty cell is filled under a saturation threshold of 90 %.

5 Conclusion

We addressed the issue of generating previously unseen word forms in Indonesian by filling empty cells inside analogical grids. We constructed analogical grids with different word vector representations and saturation thresholds. We also performed experiments in using Fisher's exact test to measure beforehand the confidence in filling empty cells. The result from Fisher's exact test gives around two times more confidence that the generated word forms will be valid when filling the empty cells.

As future work, we want to use different features for the computation of the p-value in Fisher's exact test, (e.g. frequency of the word forms in the corpus). Similar experiments on other languages should also be conducted.

Acknowledgements

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] Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. In *Proceedings of 2014 International Conference on Asian Language Processing (IALP-14)*, pages 66–69, Kuching, Malaysia, October 20-22 2014.

[2] Rashel Fam and Yves Lepage. 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)*, pages 51–60, Atlanta, Georgia, 2016.

[3] R. A. Fisher. On the interpretation of X^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[4] Nabil Hathout. Acquistion 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*, pages 1–8, Manchester, UK, August 2008.

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

[6] S.D. Larasati, V. Kuboň, and D. Zeman. Indonesian morphology tool (morphind): Towards an indonesian corpus. *Systems and Frameworks for Computational Morphology*, pages 119–129, 2011.

[7] Sylvain Neuvel and Sean A. Fulop. Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 31–40, July 2002.

[8] Ted Pedersen. Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*, Austin, TX, October 27-29 1996.

[9] Rajendra Singh and Alan Ford. 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.

[10] Nicolas Stroppa and François Yvon. 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 2005.