Indonesian unseen words explained by form, morphology and distributional semantics at the same time

Rashel Fam　Yves Lepage　Susanti Gojali　Ayu Purwarianti
早稲田大学大学院 情報生産システム研究科　Institut Teknologi Bandung
{fam.rashel@fuji., yves.lepage@}waseda.jp　susantigojali@hotmail.com, ayu@informatika.org

Abstract

We address the issue of explaining previously unseen words on different levels at the same time.

We explain unseen words on the level of form by using analogical clusters extracted from a given training set by relying on formal relations between words. The analogies which explain unseen words on the level of form are then verified on two other representation levels: morphological and semantic.

In this paper, we carry out a ten-fold cross-validation experiment on Indonesian. The experimental results show that almost half of the unseen words can be explained on the three different levels of representation at the same time.

1 Introduction

The vocabulary of a natural language processing (NLP) system is usually limited by the words learnt during the training step. Thus, unseen words are an important issue for NLP systems. Speech recognition and machine translation face this issue.

In this paper, we address the issue of explaining unseen words. We consider computational analogy as one possible way to answer this problem. For example, the word *inexhaustivity* may be explained in the following manner: $\textit{active} : \textit{inactivity} :: \textit{exhaustive} : x \quad \Rightarrow \quad x = \textit{inexhausitivity}$

Some previous works are restricted either to the formal aspect of the problem, as in [1, 2], or to its semantical aspect, like [10, 7]. In the present paper, we explain unseen words on such levels and even additional levels. For that, we first explain unseen words on the formal level. We then confirm the explanation 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 and morphological analyzers are available for this language.

	Tokens	Types	Type-Token-Ratio
Number	486,936	27,315	0.056
Avg length	6.1	8.0	

Table 1: Statistics for BPPT Corpus

2 Data used

We carried out experiments using the BPPT[1] corpus provided by PAN Localization[2]. BPPT is an Indonesian-English aligned parallel corpus of news articles. The Indonesian part 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. Almost half of the tokens (44.3 %) are hapaxes. Table 1 shows the statistics on the BPPT corpus.

We carried out a ten-fold cross-validation experiment using the BPPT corpus. Each of the ten test sets contains around 1,200 unseen words (almost 15 % of the test set). The statistics for the data, training and test sets, are shown in Table 2.

A rough estimation of the categories of the unseen words was conducted. Hundred unseen words were sampled out of the unseen words of one test set and classified by hand. 40 % of the unseen words are valid Indonesian words while around 30 % are proper nouns. The remaining unseen words are either abbreviations, typos, or foreign words.

3 Level of form

In the next following sections, we once again introduce our method to produce analogical clusters. This method has already been presented elsewhere [5, 1].

[1] Licence: Creative Commons BY-NC-SA 3.0
[2] http://www.panl10n.net/indonesia/

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

Figure 1: Word ratio for word *makan* ('to eat') and *makanan* ('food').

It relies on the notion of computational analogy between strings of symbols proposed in [4].

3.1 Word ratios

The ratio between two words is defined 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, and Figure 3.1 illustrates it on an example.

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

The notation $|S|_c$ stands for the number of occurrences of character c in string S. The last dimension, written as $d(A, B)$, is the edit distance between the two strings with 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* ('tables') for *meja* ('table').

The above definition is found in or implied by the characterization of the notion of proportional analogy between sequences of characters in [3] or [9]. 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 always be exchanged. Formula 2 gives the notation and the definition of a proportional analogy.

$$A : B :: C : D \quad \overset{\Delta}{\Longleftrightarrow} \left\{ \begin{array}{lcl} A : B & = & C : D \\ A : C & = & B : D \end{array} \right. \qquad (2)$$

$$\begin{array}{rl} \text{makan} : & \text{makanan} \\ \text{minum} : & \text{minuman} \\ \text{main} : & \text{mainan} \end{array} \qquad \begin{array}{rl} \text{minum} : & \text{meminum} \\ \text{makan} : & \text{memakan} \end{array}$$

$$\begin{array}{rl} \text{minum} : & \text{diminum} \\ \text{makan} : & \text{dimakan} \\ \text{beli} : & \text{dibeli} \end{array} \qquad \begin{array}{rl} \text{makan} : & \text{minum} \\ \text{makanan} : & \text{minuman} \\ \text{dimakan} : & \text{diminum} \\ \text{memakan} : & \text{meminum} \end{array}$$

Figure 2: Four analogical cluster of different sizes: three ratios for the clusters on the *left*, two and four ratios respectively for the clusters on the *right*.

3.2 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 Formula 2, we define an analogical cluster in Formula 3. Notice that the order of word pairs in an analogical cluster has no importance.

$$\begin{array}{l} A_1 : B_1 \\ A_2 : B_2 \\ \vdots \\ A_n : B_n \end{array} \quad \overset{\Delta}{\Longleftrightarrow} \quad \begin{array}{c} \forall (i, j) \in \{1, \ldots, n\}^2, \\ A_i : B_i :: A_j : B_j \end{array} \qquad (3)$$

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 the method proposed in [5]. Figure 2 shows examples of analogical clusters.

3.3 Explaining unseen words

For each unseen word, we extract all possible analogical clusters which include it using the words contained in the training set. If there is at least one analogical cluster extracted, it means that the unseen word can be explained on the level of form.

4 Levels of morphology and distributional semantics

4.1 Morphological representation

For the morphological representation, we use a stemmer [8] and HMM-based part-of-speech tagger [12] for Indonesian. Each word is represented by its lexeme and exponent(s), accompanied with its part-of-speech tag (see Figure 3, second line).

We verify analogies on the level of morphological representation by proportional analogy on the strings of the representations themselves using Formula 2.

Form:	makan : makanan :: minum : minuman
Morphological representation:	makan_VB : makan+an_NN :: minum_VB : minum+an_NN
Semantic representation:	$\overrightarrow{\text{makanan}} - \overrightarrow{\text{makan}} + \overrightarrow{\text{minum}} \approx \overrightarrow{\text{minuman}}$

Figure 3: Confirming an analogy on different levels of representation for the word *minuman*.

For each analogical cluster that includes an unseen word, we verify at most 30 ratios. If 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.

4.2 Semantic representation

Linguistic regularities can be captured by representing words in a vector space [11, 7]. Some tasks can be performed using this continuous word representations, such as solving semantic analogical equations. As a famous example [7], the vector for *queen* can be approximated by summing the vectors for *king* and *woman* and subtracting the vector for *man*.

We train a model for all the words contained in the BPPT corpus [6]. The vector dimension is 300 with a window size of 5. 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. If the unseen word comes out in the top 100 answers at least 50 % of the times for at least 30 analogical equations, we consider that the analogy holds on the level of semantic representation.

5 Experiments

In this section, we present our experimental protocol which uses the morphological and semantic representations and the criteria to check for analogy on these levels as explained in the previous sections. We also present the results obtained on the data introduced in Section 2.

5.1 Experimental protocol

For each unseen word in the test set, we first explain it by analogy on the level of form by extracting all possible analogical clusters which include it (see Section 3). These analogies are then confirmed on two other levels of representation: morphological and semantic (see Section 4).

We count how many unseen words can be explained on the three levels at the same time: form, morphological representation and semantic representation.

5.2 Experimental results

Table 2 (next page) shows the results obtained in a ten-fold cross-validation experiment. It shows the exact number of how many unseen words explained on the level of form were also explained on the levels of morphological and semantic representation. Overall, 49 % of the unseen words explained on the level of form could be explained on these two additional representation levels.

On the level of form only, 97 % of the unseen words can always be explained. A manual inspection of the data showed that the remaining unseen words are proper nouns and marked plurals, which confirms our observations (see Section 2) and our theoretical considerations (see Section 3.1). Around 80 % of the unseen words explained on the level of form can also be explained on the level of morphological representation. More than 55 % of the unseen words explained on the level of form can also be explained on the level of semantic representation.

6 Conclusion

We proposed a method to explain unseen words contained in a test set by taking different levels of interpretation into account. By using analogical clusters extracted from a training set, we explained unseen words on the level of form. The method relies on a previously reported formalisation of proportional analogy. Results from a ten-fold cross-validation experiment show that more than 97 % of the unseen words can be explained on the level of form. We further checked the analogy on the two additional levels of morphological and semantic representation. As a final result, 49 % of the unseen words were explained on the three levels at the same time.

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.’

exp	# types		# unseen words				
	training	test	total	explained			
				F	F & M	F & S	F & M & S
1	26,039	8,629	1,276	1,249	1,010	787	721
2	26,110	8,533	1,205	1,186	946	612	540
3	26,030	8,654	1,285	1,255	1,017	685	625
4	26,029	8,732	1,286	1,262	1,031	712	637
5	26,063	8,832	1,252	1,234	1,012	674	599
6	26,163	8,532	1,152	1,131	910	587	536
7	25,948	8,823	1,367	1,343	1,098	791	712
8	26,020	8,712	1,295	1,269	1,031	673	616
9	26,089	8,646	1,226	1,207	992	664	603
10	26,025	8,667	1,290	1,268	1,000	662	587

Table 2: Number of types in training and test set for each experiment batch (*left*). Number of unseen words (*right*) explained on the level of: form (F), morphological representation (M); semantic representation (S).

References

[1] 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.

[2] Philippe Langlais. Efficient identification of formal analogies. In *Proceedings of the Computational Analogy Workshop at the 24th International Conference on Case-Based Reasoning (ICCBR-CA-16)*, pages 77–86, Atlanta, Georgia, October 2016.

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

[4] Yves Lepage. 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, Genève, August 2004.

[5] Yves Lepage. Analogies between binary images: Application to Chinese characters. In Henri Prade and Gilles Richard, editors, *Computational Approaches to Analogical Reasoning: Current Trends*, pages 25–57. Springer, Berlin, Heidelberg, 2014.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[7] Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. 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, June 2013. Association for Computational Linguistics.

[8] Ayu Purwarianti. A non deterministic Indonesian stemmer. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–5, July 2011.

[9] 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. Association for Computational Linguistics.

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

[11] Peter D. Turney and Michael L. Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3):251–278, 2005.

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