Checking the validity of word forms generated to fill empty cells in analogical grids

Rashel Fam, Pengbo Liu and 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, liupengbo@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract

Analogical grids are constructed from a set of words contained in a text. They tend to look like paradigm tables. Obviously, not all word forms in a language are found in a given text or corpus. This leaves empty cells inside analogical grids. Analogy is a possible way to fill in these empty cells. However, there is an issue of creating invalid word forms by analogy. This paper proposes a method to assess the validity of newly generated word forms in such empty cells. Several features are extracted from analogical grids, empty cell, and the word form itself to classify whether newly generated word form is valid or not. We carry experiments in different languages with different morphological richness. Experimental results shows that our model is able to achieve very high accuracies on invalid samples. It also delivers satisfying performance on valid samples.

1 Introduction

Paradigm tables organise word forms and their lemmas according to their morpho-syntactic description. They are usually complete. The task in the SIGMORPHON campaigns [3, 4] consists in guessing a word form from a lemma and its morpho-syntactic description.

play :	*playing* :	*plays* :	*played*
walk :	*walking* :	*walks* :	*walked*
talk :		*talks* :	
fill :	*filling* :		*filled*

Figure 1: An analogical grid in English

Analogical grids have been proposed [6] to automatically organise the lexicon of a corpus without the help of morpho-syntactic description. Empty cells may appear in such analogical grids. Filling empty cells in analogical grids may lead to correct or erroneous forms. On sentences, [11] explored the way of improving the average grammaticality of sentences produced in similar analogical grids by densifying the grids themselves. On words, [7] assessed by hand the validity of newly generated words in one language, Indonesian.

This paper explores the use of various features, to automatically assess the validity of newly generated word forms. The performance is analysed for languages with different richness in morphology: from English to Finnish through German.

2 Filling empty cells in analogical grids

2.1 Analogical grids

Analogical grids are matrices of strings where any four strings from any two columns and two rows make a proportional analogy. A proportional analogy is a relation between four terms, usually noted as $A : B :: C : D$, which states that A is to B as C is to D. Analogical grids can be constructed by merging correlated analogical clusters [6]. Figure 1 shows an example of an analogical grid in English.

The size of an analogical grid is simply its total number of cells. As can be seen, there may be some empty cells. These empty cells represent word forms that do not appear in the corpus used to construct the analogical grid. We define the saturation of an analogical grid as the number non-empty cells over the total number of cells.

$$\text{Saturation} = \frac{\text{Number of non-empty cells}}{\text{Total number of cells}} \times 100\% \quad (1)$$

2.2 Filling empty cells by solving analogical equations

Empty cells can be filled with word forms generated by solving analogical equations [12]. These word forms are unseen words [5, 9, 10, 13]. In this paper, we only consider solving analogical equation on the level of form, not on semantic level. For instance, the word form *talked* can be generated by solving the following analogical equation.

$$walk : walked :: talk : x \quad \Rightarrow \quad x = talked$$

Each empty cell may be filled in by several word forms using different analogical equations made out of different word forms from the analogical grid. Let us consider an

analogical grid G with size $M \times N$. We fill an empty cell W_y^x as follows.

$$\begin{array}{ccccc} W_0^0 & : & W_0^1 & : \cdots : & W_0^m \\ W_1^0 & : & W_1^1 & : \cdots : & W_1^m \\ \vdots & \vdots & & W_y^x & \vdots \\ W_n^0 & : & W_n^1 & : \cdots : & W_n^m \end{array} \qquad W_j^i : W_y^i :: W_j^x : x$$

(a) (b)

Figure 2: A grid (a) and the analogical equation used to fill in cell W_y^x (b)

3 Validity of word forms

3.1 Assessing newly generated word forms

When solving analogical equations on the level of form, we may create erroneous word forms.

(en) $boy : boys :: child : x \quad \Rightarrow \quad x = childs$

Obviously, *childs* is not a correct word form in English. *Children* is. The rule of form transformation for one word may not be appropriate for all word forms. This situation not only exists in English, in any language. Here are also examples in Japanese and Chinese.

(ja) 政治 : 政治家 :: 数学 : $x \quad \Rightarrow \quad x =$ 数学家
(zh) 指挥 : 指挥家 :: 调音 : $x \quad \Rightarrow \quad x =$ 调音家

The correct word form for *mathematician* in Japanese is '数学者', not '数学家', and the correct form for *tuner* in Chinese is '调音师', not '调音家'.

Traditional methods assess the validity of word forms using dictionaries. However, it may be inadequate for newly generated word forms. In particular, dictionaries usually mention lemma as their entries and do not list up all possible word forms for an entry. Because of that, we propose a method to assess the validity of newly generated word forms from analogical grids. We consider assessing the problem as a classification task. We use SVM to perform a binary classification on word forms: valid or invalid. In comparison to traditional methods, we use features extracted from analogical grids and the word forms themselves instead of consulting to dictionaries.

3.2 Features used

We extracted several features for our task. These features can be categorised into three groups based on where they are extracted from: the analogical grid which the word form belongs to, the position of the empty cell, and the word form itself.

- Features extracted from the **analogical grid**:
  - **Saturation**: calculated using Formula 1
  - **Size**: total number of cells in an analogical grid
- Features extracted from the position of the **empty cell**:
  - **Number of unique possible word forms (type)**: number of all possible word form generated for this cell
  - **Number of possible word forms (token)**: same as previous feature but in terms of tokens
  - **Percentage of *None***: percentage of getting *None* (no solution) when solving analogical equations in this cell
- Features extracted from the **word form** itself:
  - **Frequency in empty cell**: how many times the word form is generated as solution for a particular cell
  - **Percentage of frequency**: same as previous feature but in percentage
  - **Char-LM**: score of character-based language model
  - **Morph-LM**: score of morpheme-based language model

Table 1: Distribution of n-grams in training data of language models

	Level of granularity	1-grams	2-grams	3-grams
en	character	140	1,946	13,529
	morpheme	6,243	63,672	118,699
de	character	145	2,031	16,660
	morpheme	5,593	109,234	262,717
fi	character	139	1,954	14,638
	morpheme	7,061	182,096	470,508

These language models are trained using *kenlm* [8] for unigram, bigram and trigram. To get the morpheme, we use *polyglot* [14]. To avoid the influence of the intersection of data set as much as possible, we train the models on parts of the data set which is not used to construct the analogical grid. Table 1 shows the distribution of n-grams in the data used to train language models.

4 Experiment results and analysis

4.1 Experiment protocol

We carry our experiments with analogical grids produced from the first thousand corresponding lines of the Europarl corpus version 3 in English (en), German (de),

Table 2: Statistics on generated word forms

	Saturation (%)	Time (s)	# of empty cells	Filled cell (%)
en	2	92	67,009	10
	10	1	4,869	27
	50	1	1,231	36
de	2	178	94,654	11
	10	1	4,767	31
	50	1	999	42
fi	2	149	96,042	7
	10	3	12,832	21
	50	3t	2,900	28

Table 3: Results of filling back randomly deleted cells in analogical grids

	Saturation (%)	Precision	Recall	F-score
en	10	17.31	5.87	8.77
	50	15.94	4.73	7.29
	90	14.47	3.37	5.45
de	10	23.22	9.25	13.23
	50	23.83	8.11	12.08
	90	21.40	6.12	9.47
fi	10	16.09	3.87	6.24
	50	22.11	6.83	10.39
	90	14.18	3.03	4.96

Finnish (fi). We filter the analogical grids using a saturation threshold and randomly choose cells as pseudo-empty cells (see Section 4.2). For a pseudo-empty cell, we generate word forms by solving all possible analogical equations, as mentioned in Section 2.2. For the classification tool, we use *libsvm* [1] with *fselect* [2] to optimize the selection of features listed in Section 3.2.

4.2 Filling empty cells

Table 2 shows the results of generating word forms from *real* empty cells in analogical grids for all the three languages. Only around 10 % of the empty cells can be filled by solving analogical equations. For computation time reasons, we retain analogical grids above a certain saturation threshold in order to avoid analogical grids with very low saturation. Analogical grids with very low saturation are obviously less reliable.

In contrast to Table 2, Table 3 shows the results of filling *pseudo-empty* cells. We randomly choose 10 % of filled cells in analogical grids. We fill back these cells and verify whether the generated word forms are equal to the actual word forms. From these results, no significant difference between morphologically poor languages and morphologically rich languages is observed.

4.3 Validity of generated word forms

Table 4 shows the accuracy of classification of newly generated word forms. Because the number of valid and invalid word forms in the data set are unbalanced, the table shows the accuracy for both valid and invalid samples. We used saturation and both scores of language model (character-based and morpheme-based) only as features to build baseline systems.

Table 4: Classification accuracy of newly generated word forms on invalid (×) and valid (✓) samples

	en		de		fi	
	×	✓	×	✓	×	✓
Saturation	100 .00	0.00	100.00	0.00	100.00	0.00
Char-LM	98.98	21.43	95.60	30.12	99.05	9.18
Morph-LM	99.32	**80.36**	98.80	**75.90**	**99.43**	**84.69**
Proposal	**99.66**	**80.36**	**99.20**	73.49	**99.43**	83.67

Table 5: F-score of features used in classification

Features	en	de	fi
Saturation	0.000207	0.000092	0.000135
Size	0.000785	0.002371	0.000238
Type	0.016500	0.020129	0.002405
Token	0.000650	0.000671	0.000001
None	0.001904	0.000259	0.004466
Frequency	**0.121445**	0.093534	**0.159033**
% of frequency	0.040672	0.068653	0.008720
Char-LM	**0.216378**	**0.141687**	0.024339
Morph-LM	**0.216614**	**0.181490**	**0.150655**

We observe that the saturation baseline system is unable to assess valid samples. The character-based language model performs poorly on valid samples in comparison to the morpheme-based language model, which shows high accuracy on both valid and invalid samples. Our proposal performs best in English for both valid and invalid samples. It also achieves the best performance for invalid samples in German and Finnish and comparably well on valid samples .

Table 5 shows the individual F-score [2] of each feature we used in the experiments. Given training vectors $x_k, k = 1, ..., m$, the number of positive and negative instances are n_+ and n_- respectively, the individual F-score of the ith feature is defined as:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(+)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (2)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average of the ith feature of the whole, positive, and negative data sets, respectively, $x_{k,i}^{(+)}$ is the ith feature of the kth positive instance, and $x_{k,i}^{(-)}$ is the ith feature of the kth negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The higher the individual F-score, the more discriminative the feature.

4.4 Analysis on the features used

The language models, both character-based and morpheme-based, are discriminative features for the classification of word forms into valid or invalid word forms especially for morphologically rich languages like Finnish. However, character-based language model seems to be not very discriminative for languages use repeated characters in word forms, as is the case of Finnish. The word form's frequency of being generated in an empty cell is also a discriminative feature for classification. The intuition is that a frequently produced word form is more likely to be valid. This feature contributes a lot in valid sample classification.

Ablation experiments were performed to confirm the results on individual F-scores. Due to space limitation, we cannot show detailed results. Character-based language model has a negative influence on valid samples in all three languages while the frequency of the word form being generated has almost no influence. We also found that morpheme-based language model has positive influence on valid samples in German and Finnish. For features with low F-score, we observed that there is a very limited influence, except for the size of the grid which has a negative influence on valid samples.

5 Conclusions

We proposed a method to fill in empty cells in analogical grids. Useful features are used to assess the validity of newly generated word forms. Ablation experiments allowed us to investigate the influence of each feature. Experimental results showed that our model is able to achieve satisfying performance even in the absence of dictionaries. Experiments in languages with more varying morphological richness and on larger data set are necessary to get a more complete view.

References

[1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 27:1–27:27, 2011.

[2] Yi-Wei Chen and Chih-Jen Lin. Combining svms with various feature selection strategies. *Feature extraction*, pages 315–324, 2006.

[3] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 14th SIGMORPHON Workshop*, pages 10–22, 2016.

[4] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task*, pages 1–30, Vancouver, August 2017.

[5] Etienne Denoual. Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of Machine Translation Summit XI*, Copenhagen, September 2007.

[6] Rashel Fam and Yves Lepage. Morphological predictability of unseen words using computational analogy. In *Proceedings of ICCBR-CA-16*, pages 51–60, Atlanta, Georgia, 2016.

[7] Rashel Fam, Yves Lepage, Susanti Gojali, and Ayu Purwarianti. A study of explaining unseen words in Indonesian using analogical clusters. In *Proceedings of ICCA-17*, pages 416–421, Yangon, Myanmar, February 2017.

[8] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st ACL*, Sofia, Bulgaria, 8 2013.

[9] Philippe Langlais and Alexandre Patry. Translating unknown words by analogical learning. In *Proceedings of the 2007 EMNLP-CoNLL*, pages 877–886, 2007.

[10] Philippe Langlais, François Yvon, and Pierre Zweigenbaum. Analogical translation of medical words in different languages. In A. Ranta and N. Nordström, editors, *Gotal'08*, volume 5221 of *Lecture Notes in Artificial Intelligence*, pages 284–295, Berlin, Heidelberg, 2008. Springer Verlag.

[11] Yves Lepage and Guilhem Peralta. Using paradigm tables to generate new utterances similar to those existing in linguistic resources. In *Proceedings of LREC 2004*, volume 1, pages 243–246, Lisbon, May 2004.

[12] Rafik Rhouma and Phillippe Langlais. Fourteen light tasks for comparing analogical and phrase-based machine translation. In *Proceedings of COLING 2014*, pages 444–454, Dublin, August 2014.

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

[14] Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Python implementation and extensions for morfessor baseline. In *Aalto University publication series*, 2013.