First Glance on Pattern-based Language Modeling

Michal Ptaszynski † Fumito Masui † Rafal Rzepka ‡ Kenji Araki ‡
† Department of Computer Science, Kitami Institute of Technology
{ptaszynski,f-masui}@cs.kitami-it.ac.jp
‡ Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

Abstract

In this paper we introduce a novel pattern-based language modeling method. The method builds upon present standards for language modeling and extends them by lifting all of the limitations presupposed by strict definition of a pattern as language entity. In particular, pattern in the proposed method is loosely defined as any frequently appearing ordered combination of sentence elements. By defining a pattern this way we allow extraction of all possible language patterns, beginning with single words, through phrases, ending on sophisticated patterns representing a specific writing style. We discuss the advantages and the constraints for the use of this method and present the so far applications of the method as well as those planned in the near future.

1 Introduction

Language modeling refers to a set of basic techniques in Natural Language Processing (NLP). It is crucial to most of NLP applications, including final word prediction [1], language identification [2], information retrieval [3], speech recognition [4], machine translation [5], part-of-speech (POS) tagging [6], spelling correction [7], or more recently sentiment analysis [8].

However, despite such a wide applicability, there has been little progress within the language modeling techniques themselves. There have been only two to three general methods for language modeling, while most research applies the most basic ones, such as bag-of-words (BOW) model or n-gram model. Although some more sophisticated models have been proposed, such as the skip-gram model, they too are bound with major constraints hindering the thorough analysis of language phenomena. Moreover, none of the more sophisticated language modeling methods has been widely recognized or frequently applied to real world tasks.

In the language modeling method proposed in this research we lift all of the limitations presupposed by previous models, which assume strict and limited definition of a pattern as a language entity. In particular, we define "language pattern" loosely as any frequently appearing ordered non-repeated combination of sentence elements. By applying this flexible definition of a language pattern we allow extraction of all possible language patterns, including single words, as in the BOW model, through phrases, as in the n-gram model, ending on sophisticated patterns with disjoint elements. Moreover, to prove the advantage of our model we have already applied it to various tasks and compared with other language models.

The outline of the paper is as follows. Firstly, section 2 describes other research related to ours, both in the area of language modeling as well as pattern extraction. Section 3 describes the language model proposed in this research. Section 4 describes the tasks in which the model has already been applied. Finally the paper is concluded in section 5 with comments on applications planned in the near future.

2 Related Research

2.1 Language Modeling

When it comes to language modeling, there have been a small number of formulated and well established methods.

For example, the bag-of-words (BOW) model [9], a piece of text or document is perceived as an unordered set of words. BOW thus disregards grammar and word order. Recently there has been proposed a more generalized BOW model with semantic concepts instead of words (bag-of-concepts) [10]. The general rule however remains the same, namely, disregarding the order of elements within input (e.g., order of concepts within sentence), and longer strings of elements (e.g., phrases).

An approach in which word order is retained is called the n-gram approach, first proposed by Shannon [11] over half a century ago, while basis for which, in terms of probabilistic theory, was formulated by Markov [12] over one hundred years ago. This approach perceives a given input (sentence) as a set of n-long ordered sub-sequences of words. This allows matching the words while retaining the sentence word order. However, the n-gram approach when applied to language, allows only for simple sequence matching, while disregarding the structure of the sentence. Although instead of words one could represent a sentence with parts of speech (POS), dependency structure, or semantic relations between concepts, the n-gram approach still does not allow extraction or matching of more sophisticated patterns than the subsequent strings of elements.

An example of such pattern, more sophisticated than n-gram, can be explained as follows. A sentence in Japanese *Kyō wa nante kimochi ii hi nanda !* (What a pleasant day it is today!) contains a pattern *nante * nanda !*[1]. Similar cases can be easily found in other languages, for instance, in English or Colombian Spanish. An exclamative sentence "Oh, she is so pretty, isn't she?", contains a pattern "Oh * is so * isn't *?". In Colombian Spanish, sentences "*¡Qué majo está carro!*" (What a nice car!) and "*¡Que majó está chica!*" (What a nice girl!) contain a common pattern "*¡Que majó está * !*" (What a

[1] Equivalent of *wh*-exclamatives in English [13, 14]; asterisk "*" used as a marker of disjoint elements.

nice * !). With another sentence, like "*¡Qué porquería de película!*" (What a crappy movie!) we can obtain a higher level generalization of this pattern, namely "*¡Que * !*" (What a * !), which is a typical example of a *wh*-exclamative sentence pattern [14, 15]. The existence of such patterns in language is common and well recognized. However, it is not possible to discover such subtle patterns using only n-gram approach.

This disadvantage of standard language models has been recognized and some modifications have been proposed.

For example, an interesting improvement of the BOW model, the positional language model, was proposed by Lv and Zhai [16]. While standard BOW model uses word occurrence frequencies, positional language model takes advantage of word positions within a document. This allows for example comparing documents more effectively. The method is interesting in the sense that it proposes a novel way the statistics of single words are calculated, introducing a notion of word position within a document. Unfortunately, this model still deals with single words, although it retains information on a general word order.

A language model which was aimed to go beyond BOW and n-grams is called the skip-gram model (sometimes also called skipped n-gram or distanced n-gram). It assumes that some words within an n-gram do not necessarily have to refer to adjacent words within a sentence, but some elements can be skipped over. The general definition of skip-gram model is very promising. In theory it should allow extraction of most of frequent language patterns from a corpus. Unfortunately the model has not been examined thoroughly enough to assume all possible variations. The major drawbacks in research studying skip-gram modeling include for example, assuming that the **skip can appear only in one place** [17]. The above examples in different languages clearly indicate that frequent and easily recognizable language patterns can consist of elements, some of which appear on the beginning of a sentence, some in the middle, and some on the end, with multiple gaps between them. Even if the model is artificially improved, and multiple places of skips are allowed, **the same number of skips needs to be retained for each gap**. In particular, a 2-skip-3-gram can only allow 2 skips between the elements, which means that the model would necessarily loose all such patterns for which, for example, first gap has 2-skips, and second has 5-skips. The final and the most influential drawback of the model lies within the generic definition of skip-gram itself. The model assumes the **full control of the skip-length** (or that the length of skip is always predetermined). Thus 2-skip-3-gram and 3-skip-3-gram consisting of the same elements (words) are represented as different entities and can never refer to the same pattern in a corpus. This assumption is non-instinctive, since one can easily imagine that the same pattern appearing in two sentences of different length will be separated by gaps of different sizes. To illustrate this problem in Table 1 we compare which of the above-mentioned language models is capable of discovering particular patterns present in the two sentences below. The last column on the right represents capability of the method proposed in this paper, based on the idea of Language Combinatorics (LC).

(1) *John went to school today.*

(2) *John went to this awful place many people tend to generously call school today.*

Finally, in all previous research on skip-grams the model was not studied for entities longer than 4-elements [18]. Only recently a research on 5-element-long skip-grams has been proposed [19].

Table 1: Comparison of capabilities of different language models to capture certain patterns from the corpus containing two sentences, (1) and (2) (◯ = capable, × = incapable).

pattern	model			
	BOW	n-gram	skip-gram	LC
John	◯	◯	◯	◯
John went	×	◯	◯	◯
John * to	×	×	◯	◯
John * school	×	×	×	◯
John * to * today	×	×	×	◯

The language modeling method presented in this paper is capable of dealing with any of the sophisticated patterns. This is due to the fact that we define the **sentence pattern** as any **ordered non-repeated frequently occurring combination of sentence elements**. This way we can extract all frequent meaningful linguistic patterns from unrestricted text.

Moreover, differently to previous research, we focus more on the possible applications and apply the method to various tasks from the areas of automatic pattern extraction and text classification.

2.2 Pattern Extraction

Generating a model of language can be interpreted as automatic extraction of frequent patterns appearing within a specified corpus (text collection). With this regard, some of the research related the most to ours include Riloff 1996 [20], Uchino et al. 1996 [21], Talukdar et al. 2006 [22], Pantel and Pennacchiotti 2006 [23] or Guthrie et al. [18]. Riloff [20] proposed AutoSlog-TS system, which automatically generates extraction patterns from corpora. However, their system was created using manually annotated corpus and a set of heuristic rules, while "patterns" in their approach were still equivalent to n-grams. Uchino et al. [21] used basic phrase templates to automatically expand the number of template patterns with application to machine translation. They also focused only on n-gram based patterns. Talukdar et al. [22] proposed a context pattern induction method for entity extraction with patterns more sophisticated than n-grams. In their research the seed word set was provided manually with the extraction limited to the patterns neighboring the seed words. Therefore the patterns in their research were separated with one word inside the pattern. Espresso, a system using grammatical information in pattern extraction was reported by Pantel and Pennacchiotti [23]. Espresso used generic patterns (e.g. "is-a" or "part-of") to automatically obtain semantic relations between entities. Although the patterns used by Espresso were not limited to n-grams, they were very generic and were provided to the system manually.

In comparison with the above mentioned methods, our method is advantageous in several ways. Firstly, we aimed to fully formalize and automatize the process of generation of patterns and extraction of frequent patterns. Secondly, we dealt with patterns more sophisticated than n-grams, or generic separated patterns.

3 Pattern Based Language Modeling Method

We assumed that applying sophisticated patterns with disjoint elements should provide better results than the usual BOW, n-gram, or skip-gram approach. We defined such patterns as ordered non-repeated combinations of sentence elements. Thus frequent patterns of this kind could be extracted automatically by firstly generating all possible ordered non-repeated combinations of sentence elements and verifying their occurrences within a corpus. Algorithms using combinatorial approach, sometimes called brute-force search algorithms, generate a massive number of combinations - potential answers to a given problem. Brute-force approach often faces the problem of exponential and rapid growth of function values during combinatorial manipulations. This phenomenon is known as combinatorial explosion [24]. Since this phenomenon often results in very long processing time, combinatorial approaches have often been disregarded. We assumed however, that combinatorial explosion can be dealt with on modern hardware to the extent needed in our research. Moreover, optimizing the combinatorial approach algorithm to the specific requirements of a given problem should shorten the processing time making the method advantageous in language processing task.

Based on the above assumptions we propose a method for pattern-based language modeling, realize the method within the **S**entence **P**attern **E**xtraction ar**C**hitecturte (SPEC) [25] and apply it in the tasks of automatic extraction of frequent sentence patterns and text classification.

In the method, firstly, ordered non-repeated combinations are generated from all elements of a sentence. In every n-element sentence there is k-number of combination clusters, such as that $1 \leq k \leq n$, where k represents all k-element combinations being a subset of n. The number of combinations generated for one k-element cluster of combinations is equal to binomial coefficient, represented in equation 1. In this procedure the system creates all combinations for all values of k from the range of $\{1, ..., n\}$. Therefore the number of all combinations is equal to the sum of all combinations from all k-element clusters of combinations, like in equation 2. Furthermore, all non-subsequent elements are separated with a wildcard (asterisk "*").

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + ... + \frac{n!}{n!(n-n)!} = 2^n - 1 \tag{2}$$

To apply the method to text classification, for all patterns generated this way the number of repetitions, or their occurrences O are used to calculate their normalized weight w_j according to equation 3, as a ratio of all occurrences of a pattern in one corpus O_{pos} to the sum of occurrences in two compared corpora $O_{pos} + O_{neg}$.

The weights are normalized to fit in range from +1 (representing patterns found uniquely in the first of the corpora) to -1 (patterns found uniquely in the second corpus). The normalization is achieved by subtracting 0.5 from the initial score and multiplying this intermediate product by 2. The score of one sentence is calculated as a sum of weights of patterns found in the sentence, like in equation 4.

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5\right) * 2 \tag{3}$$

$$score = \sum w_j, (1 \geq w_j \geq -1) \tag{4}$$

4 Applications

This section describes the so far applications of the model to various text classification tasks.

At the time of writing, there have been five papers presented at different conferences written with the use of SPEC. One of the research analyzes emotive (emotionally loaded) and non-emotive sentences. Firstly Ptaszynski et al. [26] performed preliminary investigation of such sentences with the use of SPEC. The extended analysis including was performed by Ptaszynski et al. [27]. Finally, the most thorough and detailed analysis with additional experiments was performed by Ptaszynski et al. [28]. In this research the pattern-based modeling method helped confirm that completely automatic approach to extraction of emotional patterns from sentences can give similarly good results to state-of-the-art tools developed manually.

In a different research Ptaszynski et al. [29], the SPEC system was applied in a conversation analysis task to find similarities in conversations (in Japanese) between interlocutors of different age, gender, social distance and status. The system extracted several linguistic rules (confirmed with statistical significance) some of which were previously unknown. In particular, SPEC extracted patterns characteristic for specific social distance between the interlocutors (friends or unrelated). For example, a pattern *sō sō sō!* (affirmative interjectional expression meaning "yes, yes that's right!"), although not containing any social distance-specific vocabulary *per se*, in actual language use was used in friend-friend conversations, and did not appear at all in conversations between people who first-met. On the other hand a pattern similar in meaning *hai hai hai* ("yes, yes, yes") was used in first-met conversations, but did not appear in friend-friend conversations.

In another research Nakajima et al. [30] the system was applied in analysis of future related expressions for the task of future prediction from trend information. The experiments performed with the use of the system helped prove that sentences referring to the future contain frequent patterns, while patterns in other sentences (non-future related, such as present, past or not time related) are sparse and scattered. This proved that "future-referring sentences" can be treated and analyzed as one separate kind of sentences. This discovery helped Nakajima et al. choose appropriate methods for further analysis of their data (e.g., grounded in linguistics rather than in information extraction, or data mining).

5 Conclusions and Future Work

We introduced a novel pattern-based language modeling method. By allowing a loose definition of a language pattern as any frequently appearing ordered non-repeated combination of sentence elements the method lifts all of the limitations in previous language modeling methods. We presented a general formulation of the method, discussed its advantages when compared to previous language modeling methods and presented all so far applications.

In the near future we plan to further apply the method to other text classification tasks, not limited to binary classification. We also plan to analyze the behavior of different classifiers when trained on patterns extracted with the proposed method.

References

[1] Steffen Bickel, Peter Haider, and Tobias Scheffer. 2005. Predicting sentences using n-gram language models. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 05)*, pp. 193200, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[2] Li, Haizhou, and Bin Ma. 2005. A phonotactic language model for spoken language identification. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, pp. 515-522.

[3] Ponte, J. M., & Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275-281, ACM.

[4] Peter F. Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):7985

[5] Eric Mays, Fred J Damerau, and Robert L Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5), pp. 517522.

[6] Kupiec, Julian. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, Vol. 6, No.3, pp. 225-242.

[7] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

[8] Hu, Y., Lu, R., Li, X., Chen, Y., & Duan, J. 2007. A language modeling approach to sentiment analysis. In *Computational ScienceICCS 2007*, pp. 1186-1193, Springer Berlin Heidelberg.

[9] Harris, Zellig. 1954. Distributional Structure. *Word*, 10 (2/3), pp. 146162.

[10] E. Cambria and A. Hussain. 2012. *Sentic Computing: Techniques, Tools, and Applications*. Dordrecht, Netherlands: Springer.

[11] C. E. Shannon. 1948. A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol. 27, pp. 379-423 (623-656), 1948.

[12] A.A. Markov. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Reprinted in Appendix B of: R. Howard. 1971. *Dynamic Probabilistic Systems*, Vol. 1: Markov Chains. John Wiley and Sons.

[13] Kaori Sasai. 2006. The Structure of Modern Japanese Exclamatory Sentences: On the Structure of the *Nanto*-Type Sentence. *Studies in the Japanese Language*, Vol, 2, No. 1, pp. 16-31.

[14] Fabian Beijer. 2002. The syntax and pragmatics of exclamations and other expressive/emotional utterances. *Working Papers in Linguistics 2*, The Dept. of English in Lund.

[15] C. Potts and F. Schwarz. 2008. Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora. Ms., UMass Amherst.

[16] Yuanhua Lv and ChengXiang Zhai. 2009. Positional Language Models for Information Retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pp. 299-306.

[17] Xuedong Huang, Fileno Alleva, Hsiao-wuen Hon, Mei-yuh Hwang, Ronald Rosenfeld. 1992. The SPHINX-II Speech Recognition System: An Overview, *Computer, Speech and Language*, Vol. 7, pp. 137–148.

[18] Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pp. 1-4.

[19] Rene Pickhardt, Thomas Gottron, Martin Kōrner, Paul Georg Wagner, Till Speicher, Steffen Staab. 2014. A Generalized Language Model as the Combination of Skipped n-grams and Modified Kneser Ney Smoothing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pp. 1145-1154.

[20] E. Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text, In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 1044-1049, 1996.

[21] H. Uchino, S. Shirai, S. Ikehara, M. Shintami. 1996. Automatic Extraction of Template Patterns Using n-gram with Tokens [in Japanese], *IEICE Technical Report on Natural Language Understanding and Models of Communication, 96(157)*, pp. 63-68, 1996.

[22] P. P. Talukdar, T. Brants, M. Liberman and F. Pereira. 2006. A Context Pattern Induction Method for Named Entity Extraction, In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, pp. 141-148, 2006.

[23] P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 113-120, 2006.

[24] Klaus Krippendorff. 1986. Combinatorial Explosion, In: Web Dictionary of Cybernetics and Systems. Princia Cybernetica Web.

[25] Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, Issue 1, pp. 24-36.

[26] Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki. 2014. Automatic Extraction of Emotive and Non-emotive Sentence Patterns, In *Proceedings of The Twentieth Annual Meeting of The Association for Natural Language Processing (NLP2014)*, pp. 868-871, Sapporo, Japan, March 17-21.

[27] Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki. 2014. Emotive or Non-emotive: That is The Question, In *Proceedings of 5th Workhsop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2014)*, pp. 59-65, held in conjunction with *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June 22-27.

[28] Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki. 2014. Detecting emotive sentences with pattern-based language modelling. In *Proceedings of 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2014)*, Gdynia, Poland (to appear).

[29] Michal Ptaszynski, Dai Hasegawa, Fumito Masui, Hiroshi Sakuta, Eijiro Adachi. 2014. How Differently Do We Talk? A Study of Sentence Patterns in Groups of Different Age, Gender and Social Status. In *Proceedings of The Twentieth Annual Meeting of The Association for Natural Language Processing (NLP2014)*, pp. 3-6, Sapporo, Japan, March 17-21.

[30] Yoko Nakajima, Michal Ptaszynski, Hirotoshi Honma, Fumito Masui. 2014. Investigation of Future Reference Expressions in Trend Information. In *Proceedings of the 2014 AAAI Spring Symposium Series*, "Big data becomes personal: knowledge into meaning – For better health, wellness and well-being –", pp. 31-38, Stanford, USA, March 24-26, 2014.