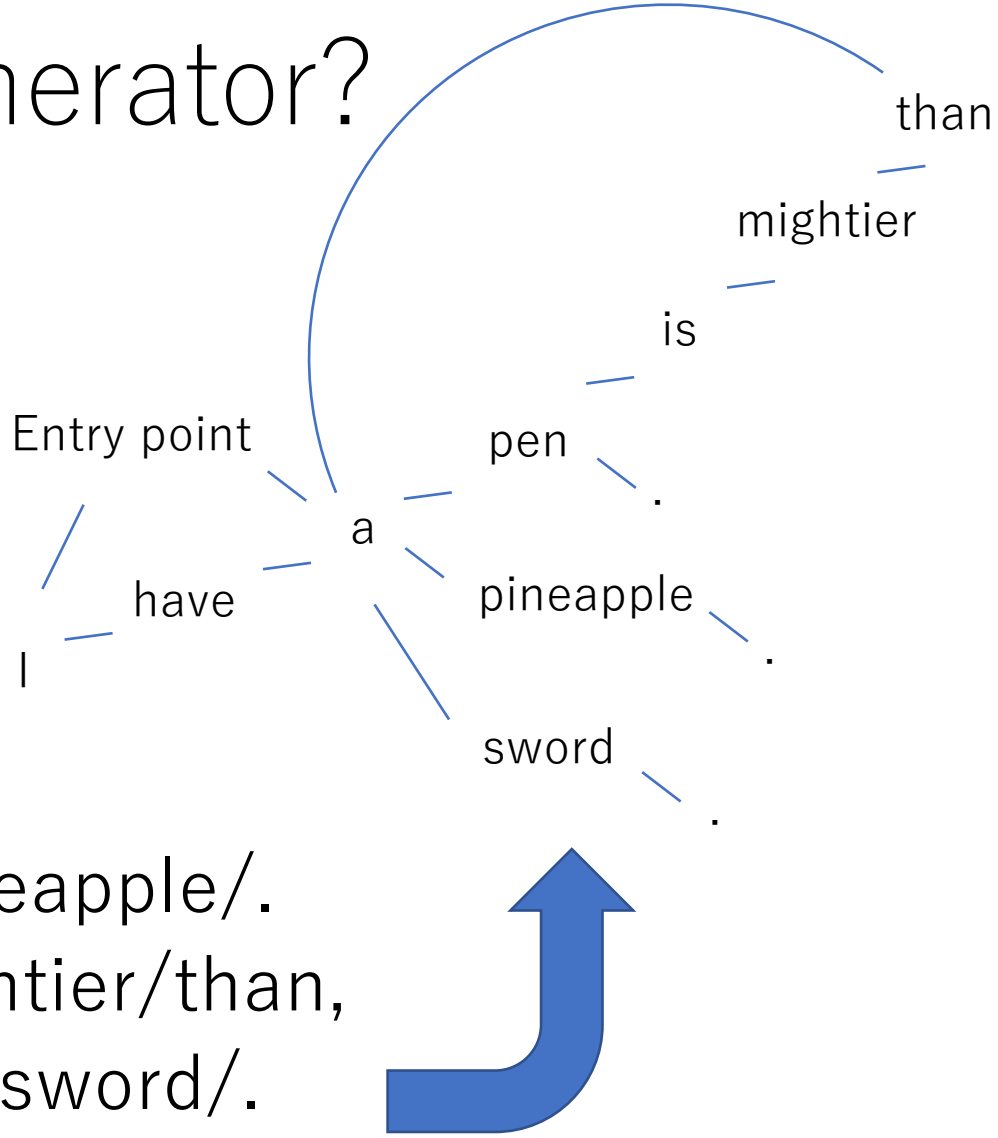
Extension of Markov text generator using Word2Vec and Machine Learning

Go Nishimura

What is Markov text generator?

Data:
I have a pen.
I have a pineapple.
A pen is mightier than a sword.
↓
I/have/a, have/a/pen, a/pen/.
I/have/a, have/a/pineapple, a/pineapple/.
A/pen/is, pen/is/mightier, is/mightier/than,
mightier/than/a, than/a/sword, a/sword/.



What's the problem?

- Can't handle words that are not in the dictionary
  - If the word is in a Word2Vec model, we can infer the dictionary for unknown words from the dictionary of other words

- Can't handle contexts more than two words
  - If target words are given, we can handle them just by joining the target words with appropriate word(s)

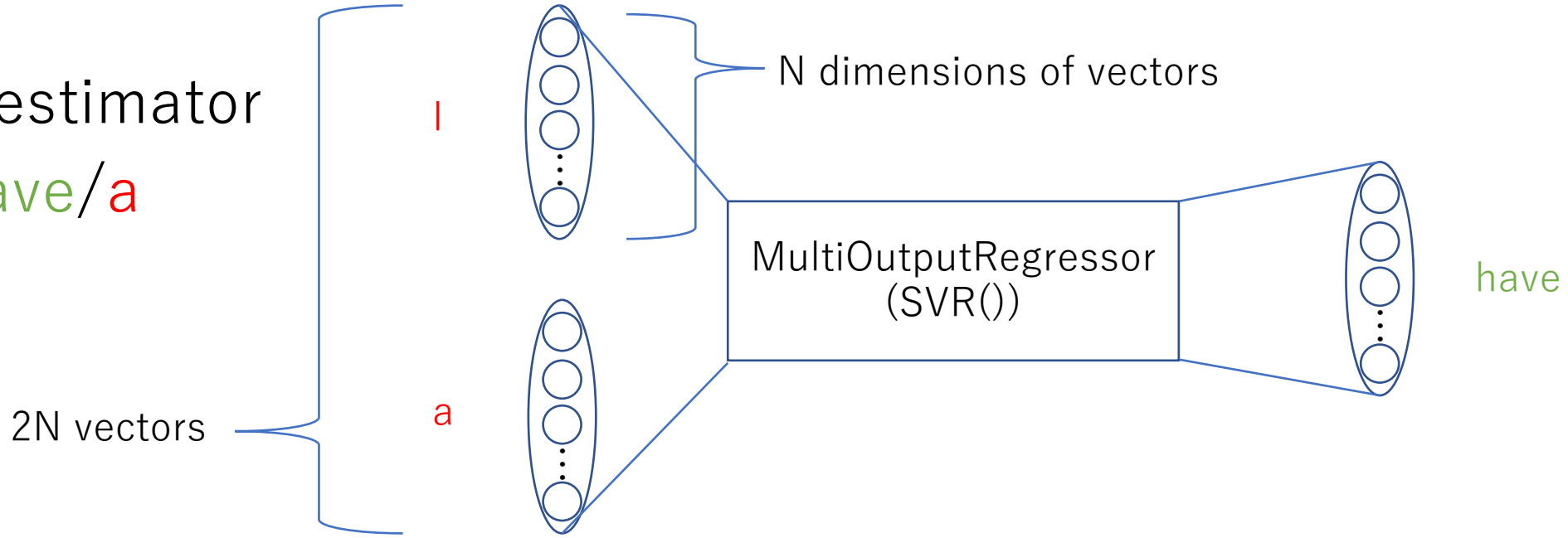
→ Let's do it in ML!

Structure

- Inputs: the vectors of the two words of the set in the dictionary (2N dimensions of vectors in total)
- Output: the vectors of the other word

e.g. 13_2 estimator

Data: I/have/a



Result

- Score: 0.18911133381
- Markov dic: 『銀河鉄道の夜』
- Word2Vec model: Japanese articles of Wikipedia
- Target words: 「野球」「する」

3 words, 13_2 only
野球 も する 0.6849915981292725
野球 が する 0.6833537817001343
野球 でも する 0.6356995105743408
野球 は する 0.566870689392089 8
野球 すら する 0.5552875995635986
野球 で する 0.551555275917053 2
野球 さえ する 0.5476392507553101
野球 に する 0.5437576174736023
野球 にも する 0.5433058738708496

4 words, 12_3 and 13_2
野球 も 気 する 0.777845025062561
野球 さえ じっと する 0.7724642753601074
野球 さえ 人目 する 0.761706471443176 3
野球 も 迷い する 0.759661793708801 3
野球 さえ 迷い する 0.756269931793212 9
野球 も じっと する 0.7494518756866455
野球 ときには 人目 する 0.654132604598999
野球 ときには じっと する 0.6183290481567383
野球 ときには 自然に する 0.601850092411041 3

4 words, 13_2 and 23_1
野球 も さえ する 0.7344176769256592
野球 でも さえ する 0.725676953792572
野球 は さえ する 0.721289396286010 7
野球 人目 も する 0.684594988822937
野球 ときには も する 0.665601849555969 2
野球 見境 た人 する 0.6409870386123657
野球 気配 も する 0.6401967406272888
野球 さして た人 する 0.6041406393051147
野球 でも た人 する 0.589533746242523 2

Filling holes of famous phrases (13_2 only)

- 吾輩 **の** 猫 **の** ある 。
- 国境 **も** 長い **ロウソク** を **かえる** と **微笑ん** で **怖がっ** た 。
- 天 **、** 人 **は** 上 **の** 人 **が** 作ら **難し** 、 人 **は** 下 **の** 人 **も** 作ら **づら** 。

Random generation (12_3 only) start = 米

米何だか何とのに用事できるだけ黙っつつ黙っておくれて怖くいつの間にか皆独り独り何かとじっと何度も面白半分それから独り人目迷いひたすらじっとて走り去って来るとかけれどなにか気持がいつもいつの間にかこっそりいつの間にか独り独りからかいた人悲しくじっとた人さしたらああ独り迷い無言ところその家々、明かり、暗闇自然に掬っ底無しの傘を垂らしたのでいつか何もかも見栄ひたすらじっとたりひたすら独りひたすら迷いそれで何もかも優しさ悪戯何かと何かと独り人目何度も枕元を放り出してくるのにどうしても何と何かとからかい独り迷い迷い独り人目人目何度もっぱなし何度も枕元も無言何度も面白半分それから無言も迷い迷いひたすらそれで何もかも楊過優しさ何かと優しさ独り迷い独りひたすらそっとたり無言何度もたらついつい独り独りからかい独り人目ひたすら迷い無言もいつも人目ひたすら迷い迷い迷い独り独り独り迷い迷い迷い無言直にひたすら独り皆独り独り何かと迷い何かとじっとて逃げだしたとかふとオーヴァーベック丁寧にもさすが気持ち、何もかもふと迷い独り人目何度もたら怖く何かとからかい迷いひたすらじっとた人突き合わせた人こもった無言。

Why not RNN or LSTM?

- Much faster to train and easier to program

- Can control sentence generation strictly
  - Haiku using specific words
  - Lyrics with rhymes
  - Make sentences under the context already given