Learning Mathematical Properties of Integers

Maria Ryskina
mryskina@cs.cmu.edu

Kevin Knight
kevinknight@didiglobal.com

Main idea

Formulating mathematical hypotheses often involves noticing useful patterns in sets of numbers. Can we help automate it?

We learn NLP-style **number embeddings** from a corpus of **mathematically interesting integer sequences** and **probe them for number-theoretic knowledge**.

Online Encyclopedia of Integer Sequences

Corpus of 336K integer sequences that represent interesting mathematical properties of different levels of complexity: oeis.org

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, …	Prime numbers
0, 1, 1, 2, 3, 5, 8, 13, 21, …	Fibonacci numbers
0, 0, 4, 10, 20, 34, 52, 73, 100, …	Greatest possible number of diagonals of a polyhedron having *n* faces

- Sequences are represented by *n* first elements (avg. 43)
- We split the corpus 90/5/5% for train/dev/test
- Integers appearing < 3 times in train replaced with UNK

Embeddings

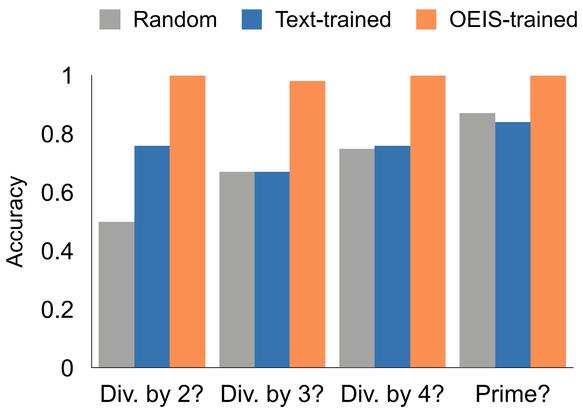
- We train embeddings on OEIS, treating sequences as "sentences" and numbers as "words"
  - LSTM: rows of the weight matrix of the embedding layer
  - LSA: truncated SVD on the number–sequence co-occurrence matrix
  - FastText: skip-gram embeddings with subword information
- Pre-trained embeddings learned from English text:
  - GloVe trained on Common Crawl
  - SkipGram trained on Wikipedia
  - FastText trained on Wikipedia + UMBC + statmt.org news

What's in an integer embedding?

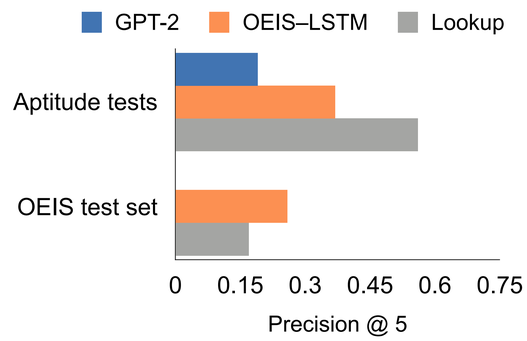
- **'Evenness neuron':** element 156 of OEIS–LSTM embeddings is positive for even values and negative for odd values, holds up to 50 with a few exceptions

Integer	1	2	3	4	5	6	7	8	9	10
Neuron 156	0.15	0.29	-0.04	0.26	-0.08	0.38	-0.31	0.39	-0.02	0.43

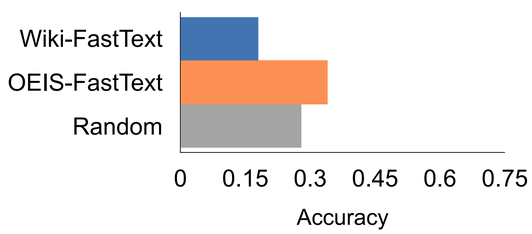
- **Probing classifiers:**
  - Train logistic regression on integers 1–1000, test on integers 1001–2000
  - Probe embeddings for divisibility by 2, 3, 4, and primality
  - More experiments with probing for value and magnitude in the paper



- **Sequence completion:**
  - Testing on OEIS test set and human aptitude test questions, e.g. *65536, 256, 16, ? → 4*
  - Lookup baseline: search for the sequence in OEIS and return the most frequent continuation



- **Mathematical analogies:**
  - Human aptitude multiple-choice tests, e.g. *5 : 36 :: 6 : ? → 49*
  - Predicting the answer with vector arithmetic



- **Expanding integer seed sets:**
  - Given a small set of integers, predict possible expansions
  - Predicting candidates by distance to embedding set centroid

*5, 13, 29 → **19**, **7**, **17**, **3**, 9, **23** (primes)*

*73, 97, 83 → **79**, **71**, **67**, **89**, 77, **103** (primes)*

*729, 1024, 243 → **2187**, **81**, **256**, **64**, **27**, **512** (powers of 2 and 3)*