Scaling up to 10 billion+ corpora

Miloš Jakubíček

milos.jakubicek@sketchengine.co.uk

4th Sketch Engine Workshop
Tallinn, October 16, 2013

Premise

Corpora: the bigger the better.

TenTen series – 10^{10} words.

Premise

Corpora: the bigger the better.

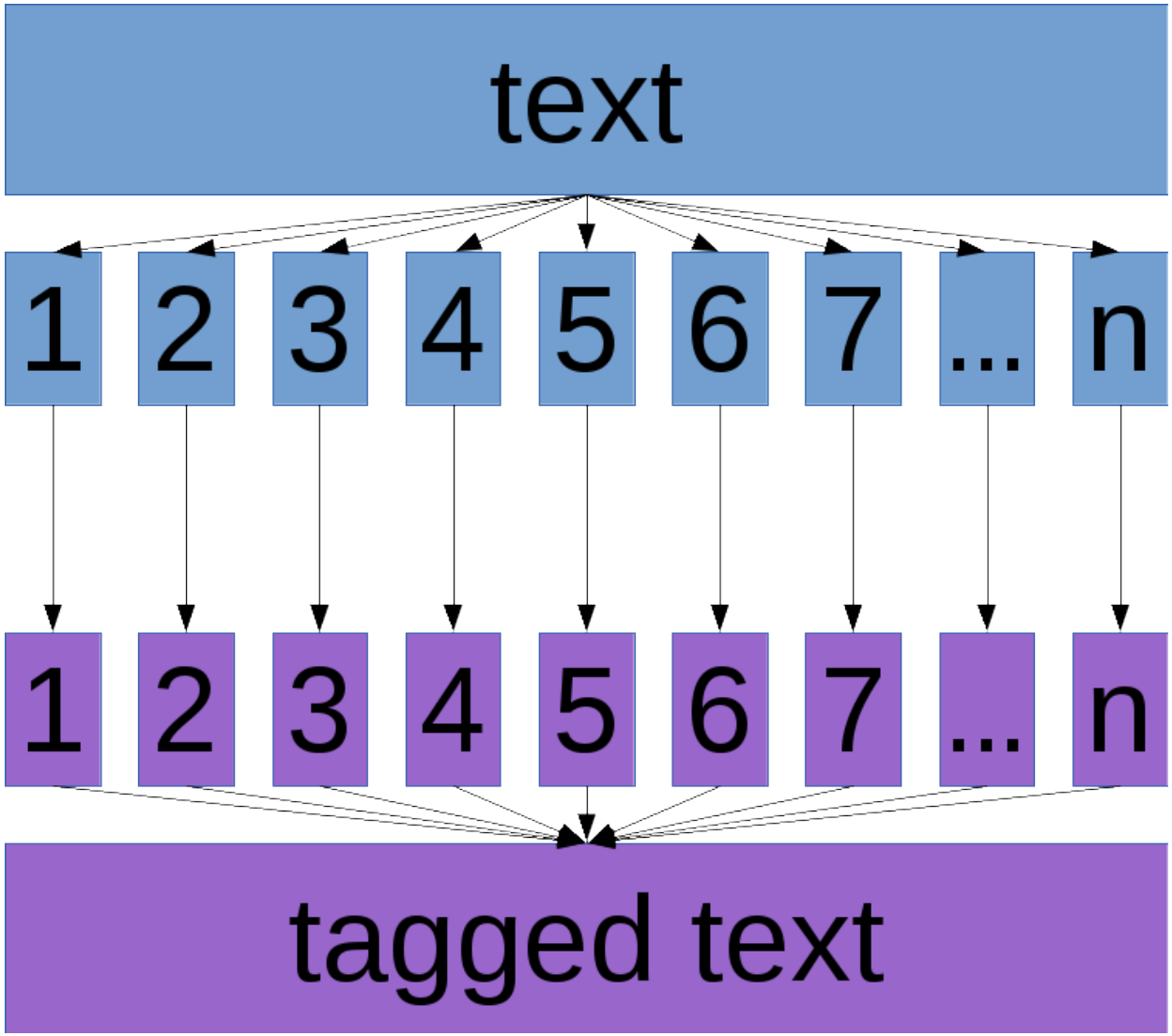
TenTen series – 10^{10} words.

Too big?

- to compile
- to store
- to search

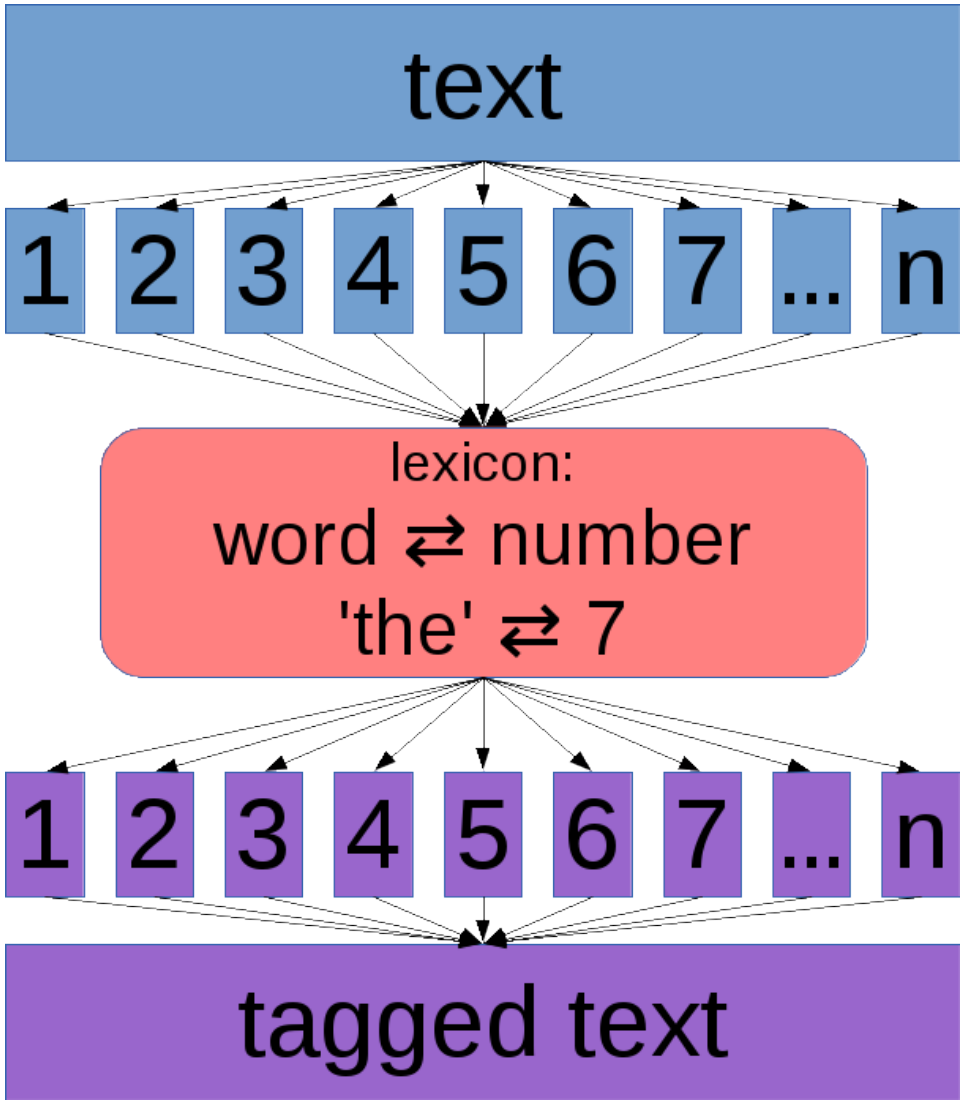
Compilation

- "Go through 2 TB of data and create all the indices."
- English ClueWeb (80 billion tokens) – 14 days
- For many NLP tasks
  - if too slow $\Rightarrow$ trivial parallelization (split and run n times, concatenate results).


text
1
2
3
4
5
6
7
...
n
1
2
3
4
5
6
7
...
n
tagged text

Compilation

- "Go through 2 TB of data and create all the indices."
- English ClueWeb (80 billion tokens) – 14 days
- For many NLP tasks
  - if too slow $\Rightarrow$ trivial parallelization (split and run n times, concatenate results).
  - not possible for corpus indexing – indices operate on word numbers, lexicon must be consistent


text
1
2
3
4
5
6
7
...
n
lexicon:
word ⇄ number
'the' ⇄ 7
1
2
3
4
5
6
7
...
n
tagged text

Parallel Corpus Compilation

Parallelization through virtualization

1. split into n parts
2. compile n independent corpora concurrently
3. create a virtual corpus ($\rightarrow$ renumber indices)
4. devirtualize the virtual corpus
5. remove former corpus parts

Effective speedup close to 80% of n.

Parallel Word Sketch Compilation

- a similar approach, shared resource = lexicon of grammatical relations
- different handling per grammatical relation type (TRINARY, COLLOC)
- speedup up to very close to n for n parts

Subcorpus Compilation

- subcorpus = set of corpus parts
- for some tasks additional statistics need to be computed

- previous approach not really scalable for heavily scattered subcorpora
- reimplementation $\Rightarrow$ speedup 10x-10000x

Storing

By September 2013 more than **400 corpora** for **70 languages**:

- 100+ corpora having more than 100 million tokens
- 30+ corpora having more than 1 billion tokens
  - In 2010 a series of TenTen (10^{10}) corpora started
- 56 languages with a PoS-tagged corpus
- 36 languages with word sketches
- 21 languages with integrated tagger for tagging user corpora
- ~50 TB of data

Storing

- how to store text (corpus source)
- how to store numbers (corpus indices)

Storing

- how to store text (corpus source)
- how to store numbers (corpus indices)

- naïve approach very costly: fixed-length encoding, e.g. each number 4 bytes, 10 billion positions = 40 GB of data ($\times$ 2 $\times$ number of attributes)
- much better: variable-length encoding of numbers
- but: more complex algorithms and data structures needed for bit operations
- 10 billion positions = 15 GB of data

Searching

Asynchronous Query Processing (AQP)

- 10 billion positions: 2 billion nouns
- let's do a concordance!
- what for?
  - *not* to read it
- sampling enough – but number of hits needed
- immediate user response – partial results

Conclusions

- Parallel and distributed processing required for large corpora
- Not always trivial in case of corpus processing
- But solutions are ready and in place