TID – lecture

Incremental construction of minimal finite automata and its utilization in natural language processing

Jan Kouřil

13.12.2011

Finite automata

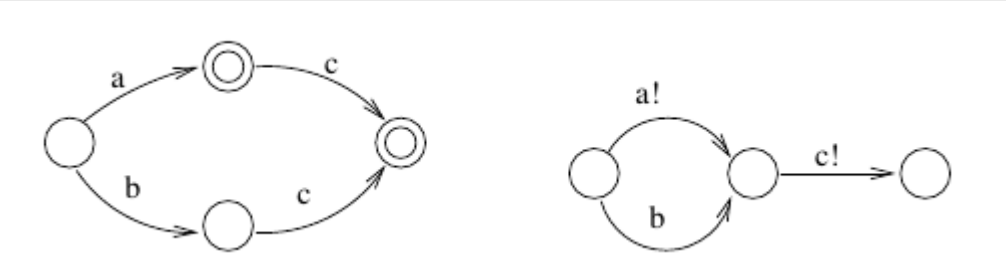
- Criteria for usage
  - Low memory requirements
  - Speed
- Types used in NLP
  - Deterministic
  - Acyclic
  - Minimal

FA Attributes

- Reachability and Usefulness
  - $Useful_s(M) = (\forall_{q \in Q} \exists_{x \in \Sigma^*} q \in \delta^*(q_0, x))$
- Minimality
  - $Minimal(M) = (\forall_{q1,q2 : q1 \in Q \wedge q2 \in Q \wedge q1 \neq q2} \vec{L}(q_1) \neq \vec{L}(q_2)) \wedge Useful_s(M)$

Final transitions

- Equivalence with classical FA
  - $Minimal(FA) \geq Minimal(FA\ with\ final\ transitions)$



Transducers

- Translate one string into another
  - Used mostly in morphology
  - Recognition tasks

Incremental construction

- Creating dictionaries
  - From sorted data
  - From unsorted data
- Dictionary changed after every new word
- Isomorphic trees avoided

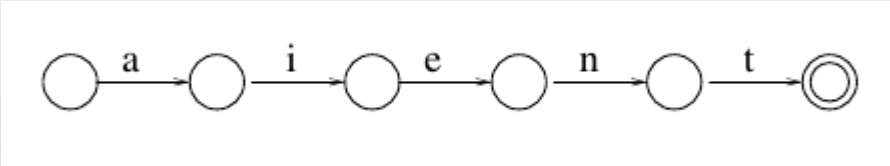
From sorted data

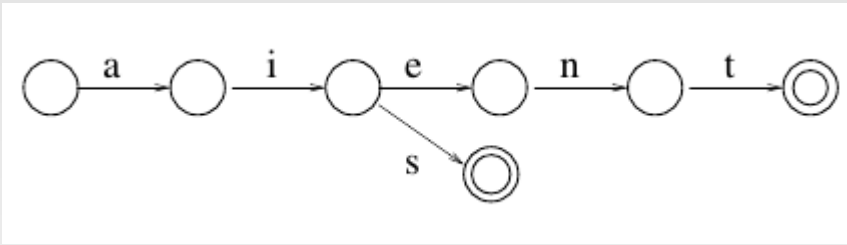
- States belong to same class if:
  - They are either both final, or both non-final
  - They have the same number of outgoing transitions
  - Corresponding transitions have the same labels
  - Corresponding transitions lead to the same states
  - States reachable via outgoing transitions are the sole representatives of their classes

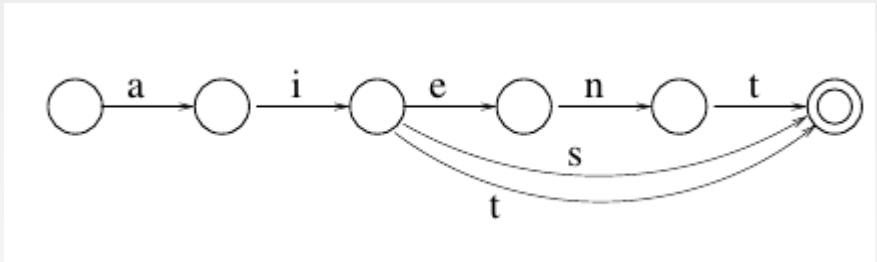
Incremental algorithm

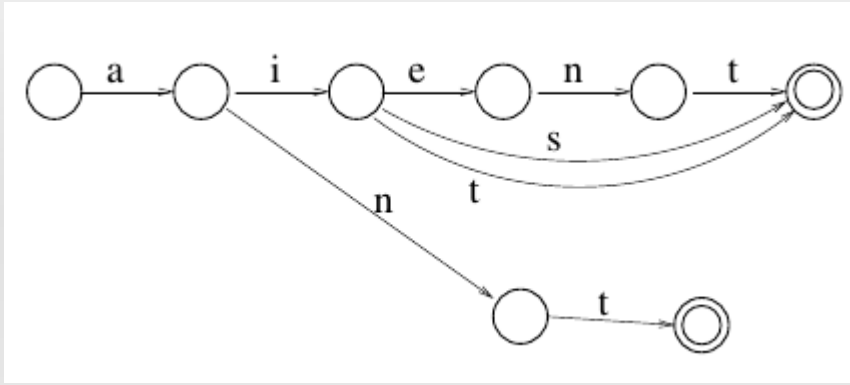
- Compute common prefix
- Append the rest of the word
- Merge common word endings

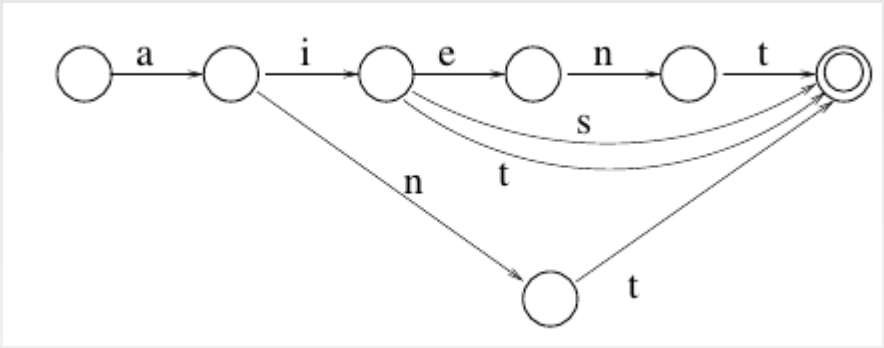
Example

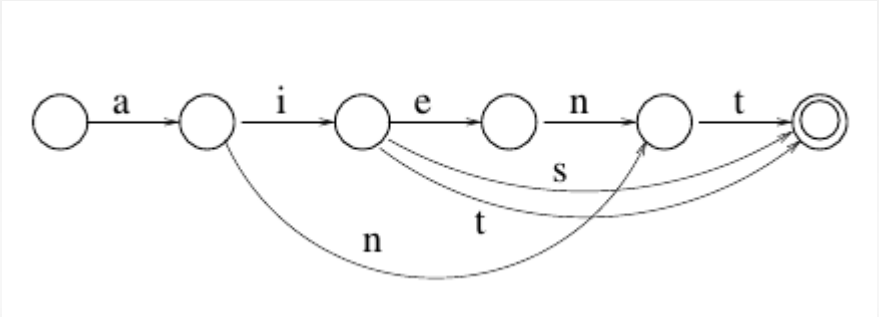






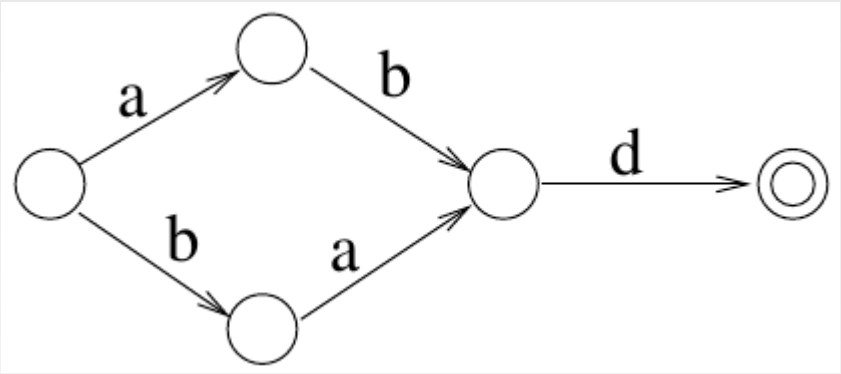


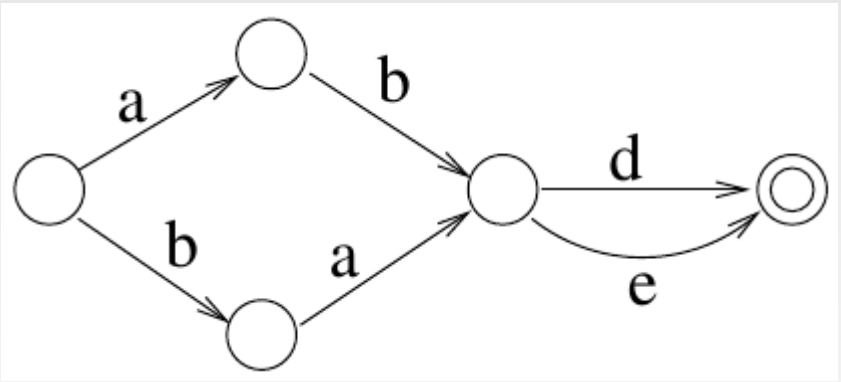




From unsorted data

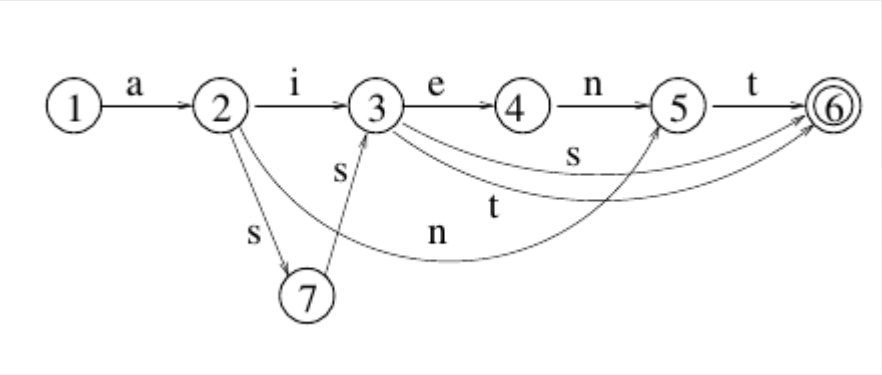
- Why is this different?

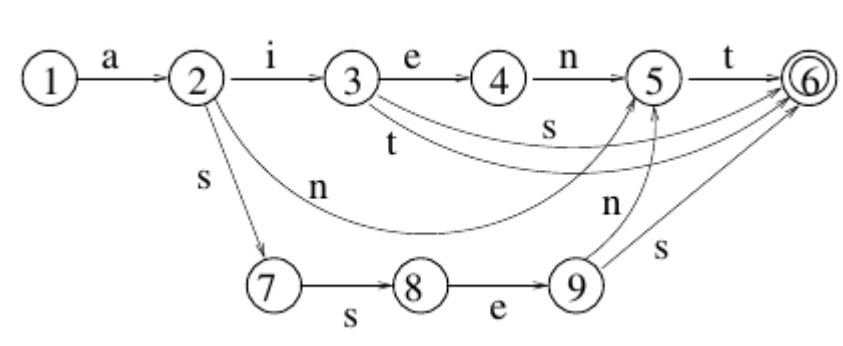




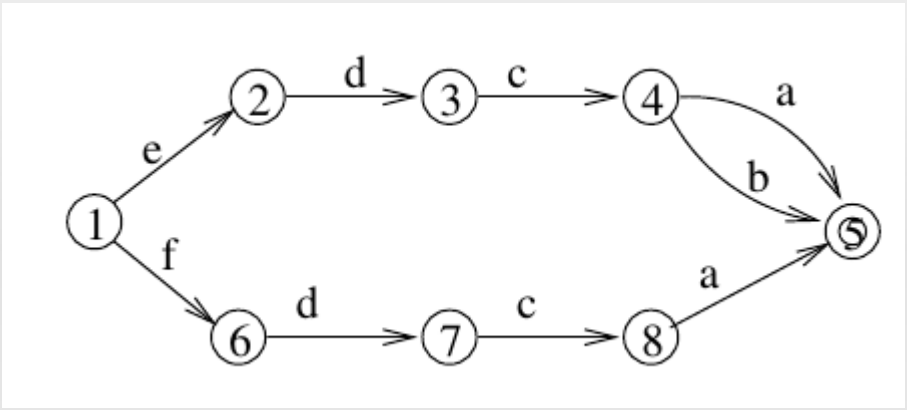
Amended algorithm

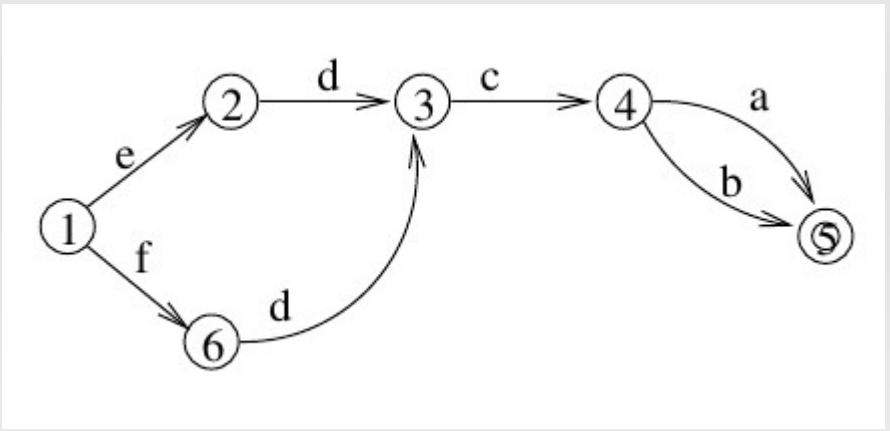
- Compute common prefix
- Clone confluent states
- Add the rest of the word

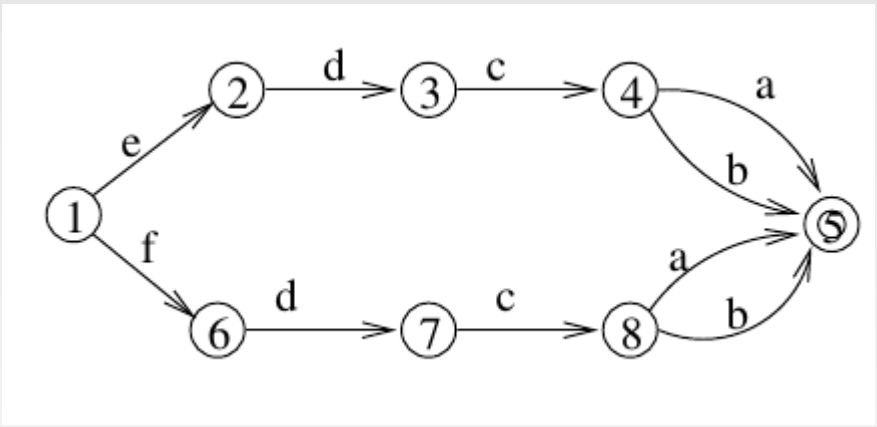


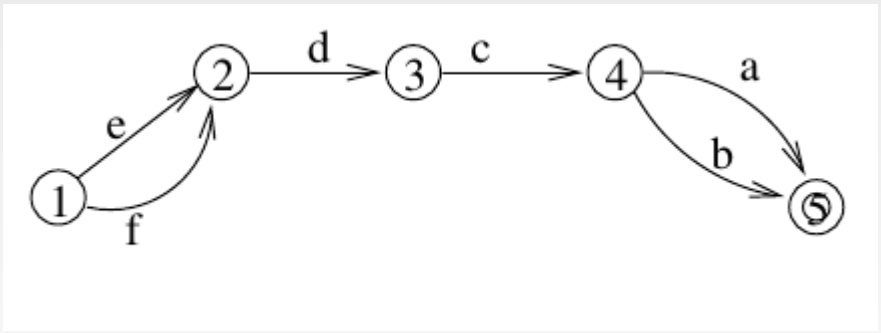


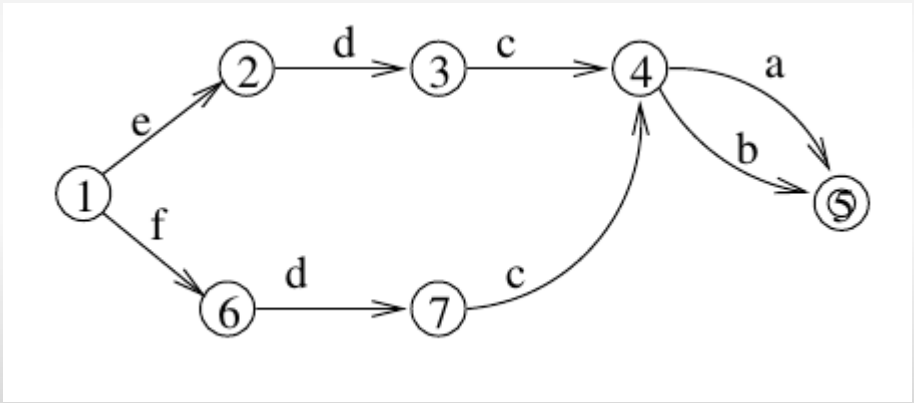
Another example











Usage

- Morphology
  - Lookup & guessing
- Spelling correction
  - Restoration of diacritics
  - Edit distance
  - Searching for similar words

Error correction

- Errors can be created by
  - Misprints
  - Unsufficient morphology knowledge
  - Lack of knowledge of proper spelling
- Correction methodology
  - Short edit distance
  - Morphology analysis of word's stem
  - Analysis of words with similar pronunciation

Thank you for your attention