Excursion: Machine Translation

Andreas Maletti

April 24, 2007

Natural Language Processing

Subfields

- Speech recognition and synthesis
- Machine translation
- Language modelling
- Text summarization
- ...

Machine Translation

- Rule-based (e.g. SysTran)
- Statistical (e.g. Google Translator)

Statistical Machine Translation System

Overview

We like you

⇓

Machine translation

⇓

0.1	Wir mögen dich	0.1	Wir mögen euch
0.05	Wir mögen sie	0.02	Wir haben dich gern
...			
0.01	Wir und du	0.005	Wir wie du
...			
≈ 0	Man auf deine Weise		

Implicit vs. explicit knowledge

Implicit knowledge

- model restrictions; hard-wired information
- task-specific; difficult to adapt
- supplied by programmer

Explicit knowledge

- model parameters; inputs
- task-specific; easy to change
- supplied by programmer or training

Implicit vs. explicit knowledge

Implicit knowledge

- model restrictions; hard-wired information
- task-specific; difficult to adapt
- supplied by programmer

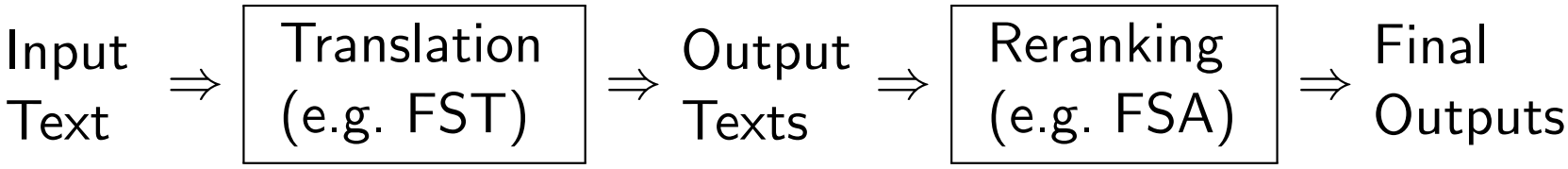
Explicit knowledge

- model parameters; inputs
- task-specific; easy to change
- supplied by programmer or training

$\Longrightarrow$ Framework approach

A closer look — string-based (e.g. CARMEL)

Overview

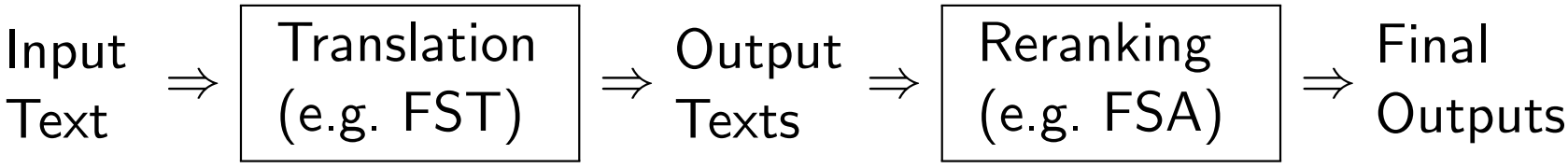


Abbreviations

- FST = Finite State Transducer
- FSA = Finite State Automaton

A closer look — syntax-based (e.g. TIBURON)

Overview

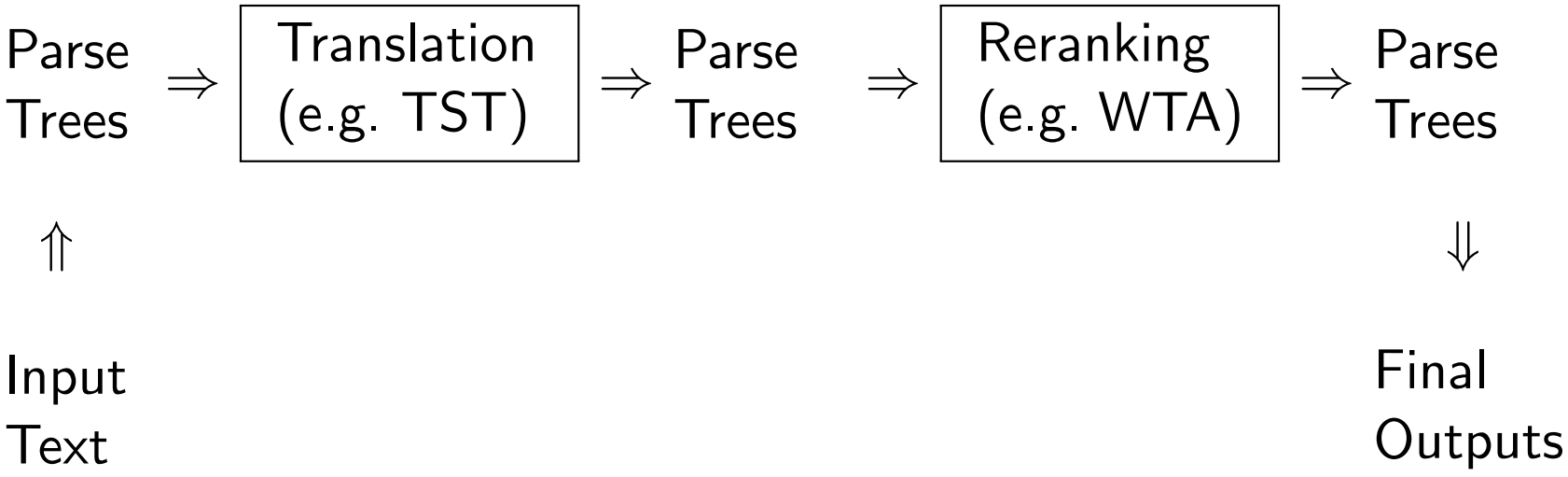


Abbreviations

- FST = Finite State Transducer
- FSA = Finite State Automaton

A closer look — syntax-based (e.g. TIBURON)

Overview



Abbreviations

- TST = Tree Series Transducer
- WTA = Weighted Tree Automaton

Formalization and Representation

Parsing

- In general: $p\colon \Sigma^* \to (T_\Delta \to [0,1])$
- Representation of $p(w)\colon T_\Delta \to [0,1]$ by WTA

Formalization and Representation

Parsing

- In general: $p\colon \Sigma^* \to (T_\Delta \to [0,1])$
- Representation of $p(w)\colon T_\Delta \to [0,1]$ by WTA

Translation

- In general: $\tau\colon (T_\Delta \to [0,1]) \to (T_\Gamma \to [0,1])$
- Implementation by TST
- but τ should yield WTA; i.e., $\tau(\mathrm{WTA}) \subseteq \mathrm{WTA}$

Formalization and Representation

Parsing

- In general: $p\colon \Sigma^* \to (T_\Delta \to [0,1])$
- Representation of $p(w)\colon T_\Delta \to [0,1]$ by WTA

Translation

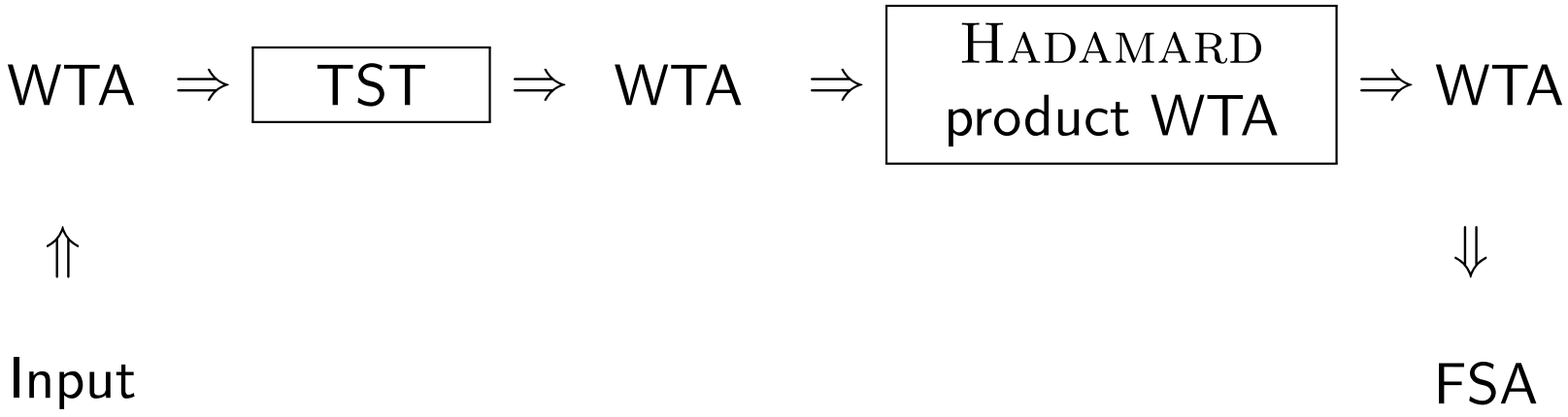
- In general: $\tau\colon (T_\Delta \to [0,1]) \to (T_\Gamma \to [0,1])$
- Implementation by TST
- but τ should yield WTA; i.e., $\tau(\mathrm{WTA}) \subseteq \mathrm{WTA}$

Reranking

- In general: $r\colon (T_\Gamma \to [0,1]) \to (T_\Gamma \to [0,1])$
- Implementation by $\textsc{Hadamard}$ product with WTA
- can be made such that r delivers WTA

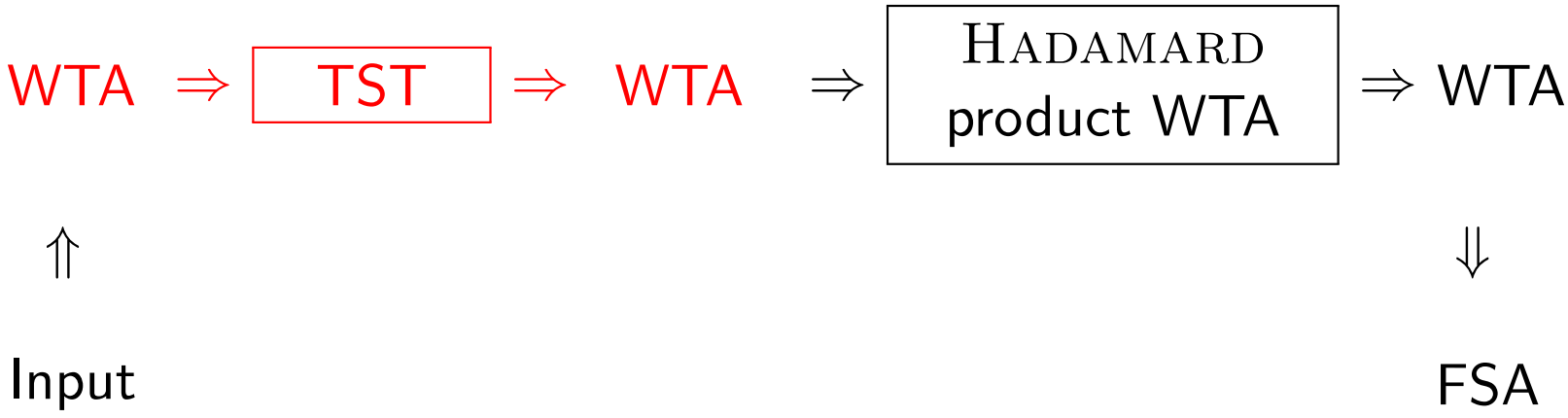
The "full" picture — syntax-based

Overview



The "full" picture — syntax-based

Overview

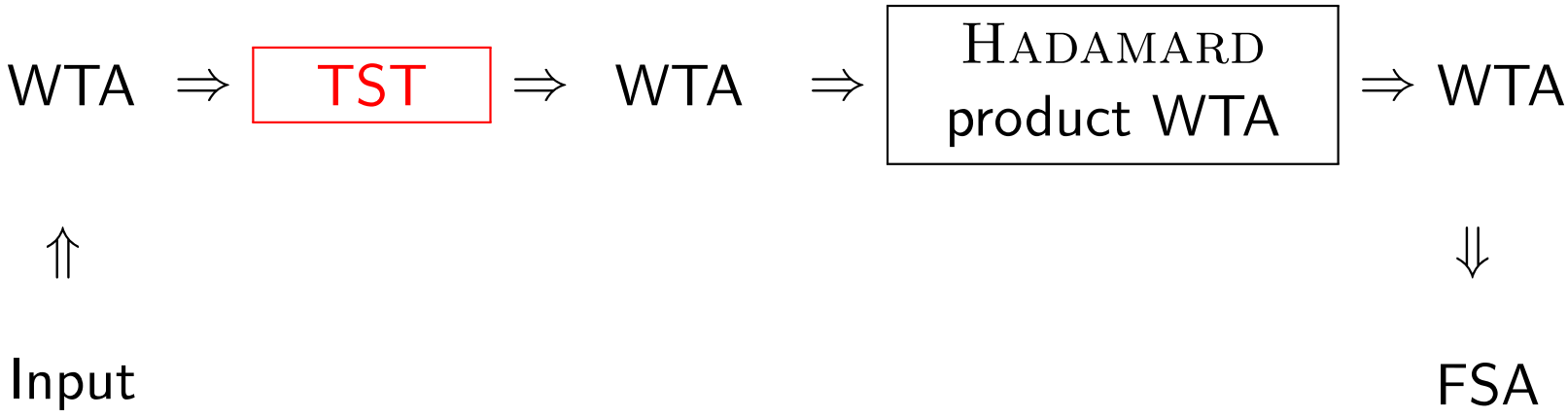


Algorithms

- Application of TST to WTA
- Composition of TST
- HADAMARD product for WTA
- Determinization and Minimization for WTA

The "full" picture — syntax-based

Overview

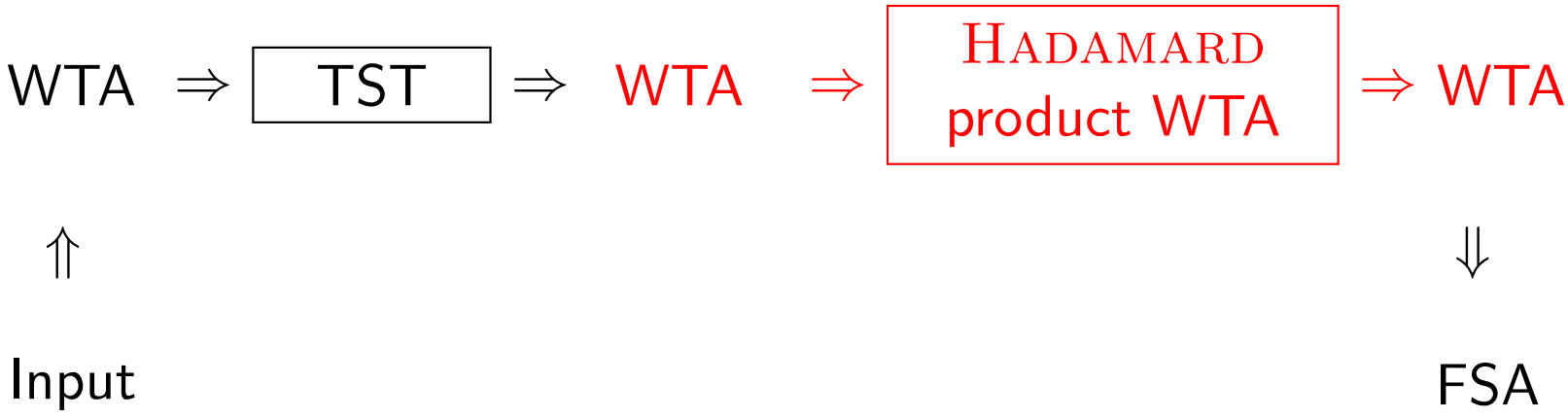


Algorithms

- Application of TST to WTA
- Composition of TST
- HADAMARD product for WTA
- Determinization and Minimization for WTA

The "full" picture — syntax-based

Overview

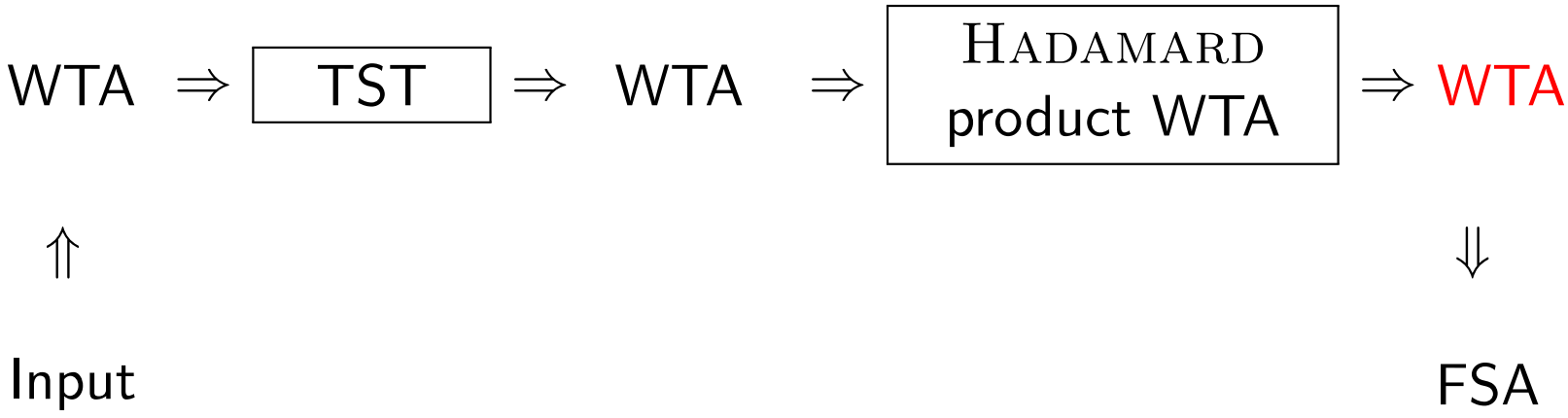


Algorithms

- Application of TST to WTA
- Composition of TST
- HADAMARD product for WTA
- Determinization and Minimization for WTA

The "full" picture — syntax-based

Overview



Algorithms

- Application of TST to WTA
- Composition of TST
- HADAMARD product for WTA
- Determinization and Minimization for WTA