Statistical Machine Translation of Natural Languages

– Rule Extraction and Training Probabilities –

Matthias Büchse, Toni Dietze, Johannes Osterholzer,
Torsten Stüber, Heiko Vogler

Technische Universität Dresden
Germany

Graduiertenkolleg
"Quantitative Logics and Automata"
Gohrisch, March, 2013

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities

given:

- source language SL
- target language TL

find:

$$\text{translation} \quad h : \mathrm{SL} \to \mathrm{TL}$$

e.g.

SL = English $\qquad$ s = I saw the man with the telescope

TL = German $\qquad$ h(s) = Ich sah den Mann durch das Tel.

given:

- source language SL
- target language TL

find:

machine translation $h : \mathrm{SL} \rightarrow \mathrm{TL}$

e.g.

SL = English   s = I saw the man with the telescope
TL = German h(s) = Ich sah den Mann durch das Tel.

given:

- source language SL
- target language TL

find:

machine translation $\quad h : \mathrm{SL} \to \mathrm{TL}$

e.g.

$\mathrm{SL} =$ English $\qquad$ s $=$ I saw the man with the telescope

$\mathrm{TL} =$ German $\qquad$ h(s) $=$ Ich sah den Mann durch das Tel.

machine translation $\rightsquigarrow$ statistical machine translation (SMT)

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

hypothesis space: $\mathcal{H} \subseteq \{h \mid h : \mathrm{SL} \to \mathrm{TL}\}$

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

  hypothesis space: $\mathcal{H} \subseteq \{h \mid h : \mathrm{SL} \rightarrow \mathrm{TL}\}$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

  [Lopez 08]: *"By examining many samples
  of human-produced translations,
  SMT algorithms automatically
  learn how to translate."*

- assumptions $\longrightarrow$ modeling $\longrightarrow$ $\mathcal{H}$ hypothesis space

  hypothesis space: $\mathcal{H} \subseteq \{h \mid h : \mathrm{SL} \to \mathrm{TL}\}$

- $\mathcal{H}$ and training data $\longrightarrow$ training $\longrightarrow$ $\hat{h} \in \mathcal{H}$

  [Lopez 08]: *"By examining many samples of human-produced translations, SMT algorithms automatically learn how to translate."*

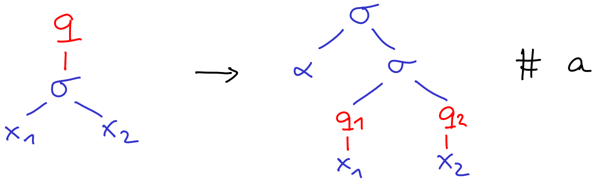
- $\hat{h}$ and test data $\longrightarrow$ evaluation $\longrightarrow$ score

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities

weighted tree transducer (wtt) $\mathcal{M} = (Q, \Sigma, \Delta, q_0, R)$

- Q finite set (states)
- Σ, Δ finite sets (input- / output-symbols)
- $q_0 \in Q$ (initial state)
- R finite set of particular term rewrite rules with weights



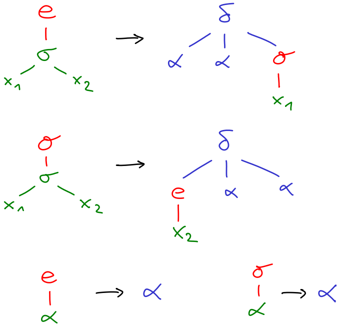
linear, nondeleting in $x_1, \ldots, x_k$

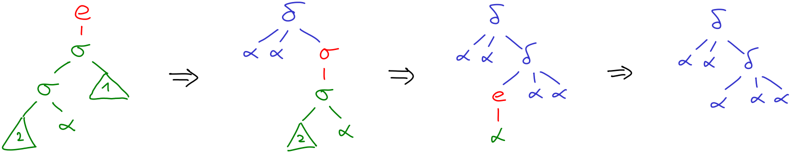
$$\mathcal{M} = (Q, \Sigma, \Delta, e, R)$$

$$Q = \{ e, \sigma \}$$

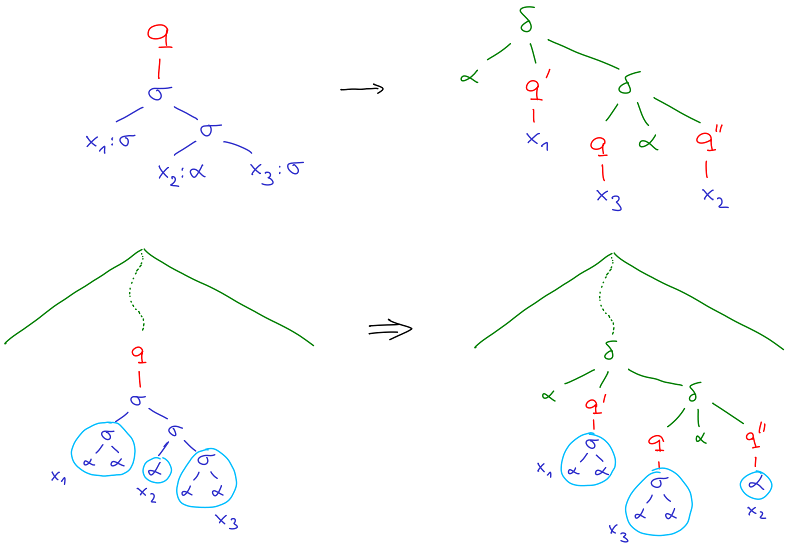
$$\Sigma = \{ \sigma^{(2)}, \alpha^{(0)} \}$$

$$\Delta = \{ \delta^{(3)}, \alpha^{(0)} \}$$





extended left-hand sides, one-symbol look-ahead:



- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname*{argmax}_{\substack{(d,r)\in Y:\\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

- $\hat{h}$ and test data $\longrightarrow$ [evaluation] $\longrightarrow$ score

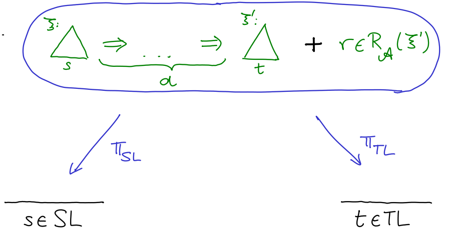
- assumptions $\longrightarrow$ modeling $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y:\\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data

recall:



- $\hat{h}$ and test data $\longrightarrow$ evaluation $\longrightarrow$ score

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y:\\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$
$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y: \\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

training data (corpora):

- Hansards of Parliament of Canada, English-French
  2 x 1,278,000
- EUROPARL, Danish, German, English, French, ...,
  1,500,000 / language pair, 50,000,000 words / language
- TIGER (v2.1), German, 50,000 sentences

- assumptions $\longrightarrow$ $\boxed{\text{modeling}}$ $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y: \\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ $\boxed{\text{training}}$ $\longrightarrow$ $\hat{h} \in \mathcal{H}$

- assumptions $\longrightarrow$ $\boxed{\text{modeling}}$ $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\Big(\operatorname{argmax}_{\substack{(d,r)\in Y: \\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\Big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ $\boxed{\text{training}}$ $\longrightarrow$ $\hat{h} \in \mathcal{H}$

$$\mathcal{H} \rightsquigarrow \mathcal{H}' = \{h_{\mathcal{M}} \mid \text{wtt } \mathcal{M}\}$$

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$
$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y:\\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

$$\begin{aligned} \mathcal{H} \rightsquigarrow \mathcal{H}' &= \{h_{\mathcal{M}} \mid \text{wtt } \mathcal{M}\} \\ &= \{h_{(\mathcal{N},p)} \mid \text{tt } \mathcal{N}, \text{ probability assignment } p\} \end{aligned}$$

- assumptions $\longrightarrow$ [modeling] $\longrightarrow$ $\mathcal{H}$ hypothesis space

$$\mathcal{H} = \{h_{\lambda,\mathcal{M},\mathcal{A}} \mid \lambda \in \mathbb{R}^2_{\geq 0}, \quad \text{wtt } \mathcal{M}, \quad \text{wta } \mathcal{A}\}$$

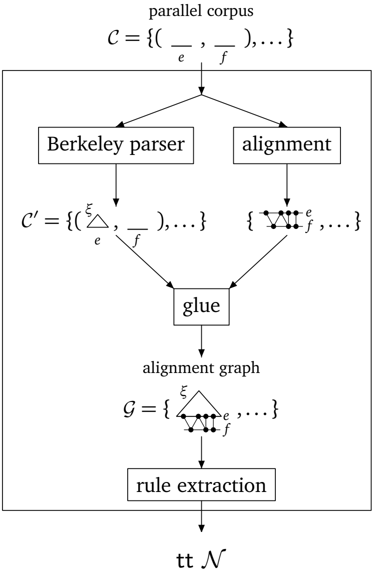
$$h_{\lambda,\mathcal{M},\mathcal{A}}(s) = \pi_{\mathrm{TL}}\big(\operatorname{argmax}_{\substack{(d,r)\in Y: \\ \pi_{\mathrm{SL}}(d,r)=s}} \mathrm{wt}_{\mathcal{M}}(d)^{\lambda_1} \cdot \mathrm{wt}_{\mathcal{A}}(r)^{\lambda_2}\big)$$

- $\mathcal{H}$ and training data $\longrightarrow$ [training] $\longrightarrow$ $\hat{h} \in \mathcal{H}$

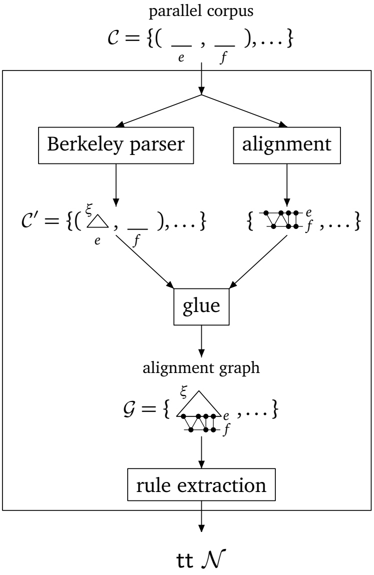
$$\begin{aligned} \mathcal{H} \rightsquigarrow \mathcal{H}' &= \{h_{\mathcal{M}} \mid \text{wtt } \mathcal{M}\} \\ &= \{h_{(\mathcal{N},p)} \mid \text{tt } \mathcal{N}, \text{ probability assignment } p\} \\ &= \bigcup_{\text{tt } \mathcal{N}} \{h_{(\mathcal{N},p)} \mid \text{probability assignment } p\} \end{aligned}$$

rule extraction:	tt $\mathcal{N}$
training probabilities:	probability assignment $p : R \to [0,1]$

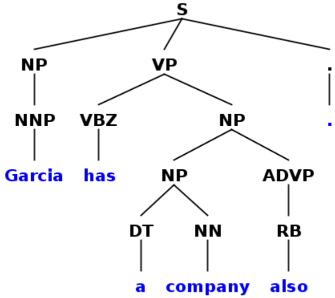
Rule extraction: [Galley, Hopkins, Knight, Marcu 04]



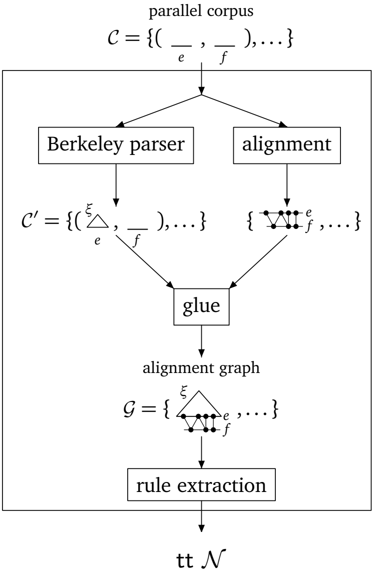
Rule extraction: [Galley, Hopkins, Knight, Marcu 04]



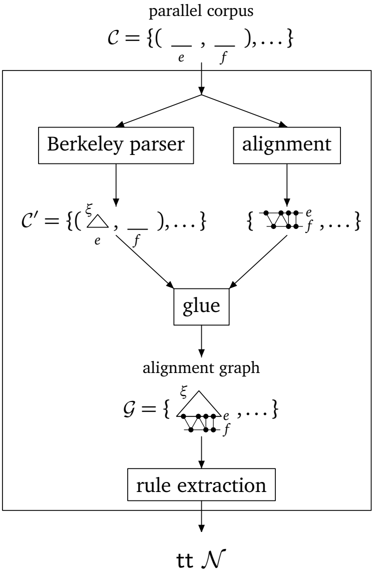
Garcia has a company also .



Rule extraction: [Galley, Hopkins, Knight, Marcu 04]



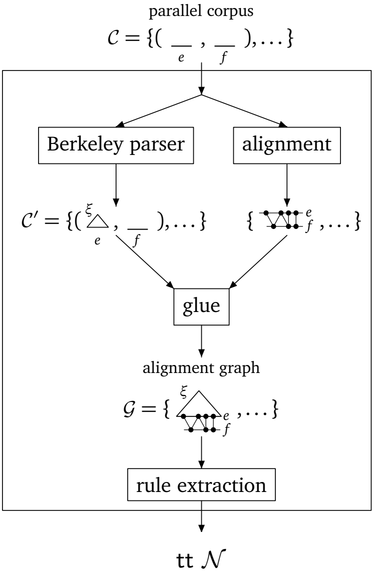
Rule extraction: [Galley, Hopkins, Knight, Marcu 04]

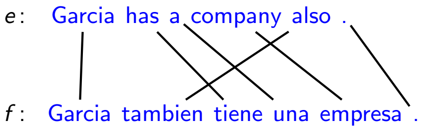


e : Garcia has a company also .

f : Garcia tambien tiene una empresa .

Rule extraction: [Galley, Hopkins, Knight, Marcu 04]

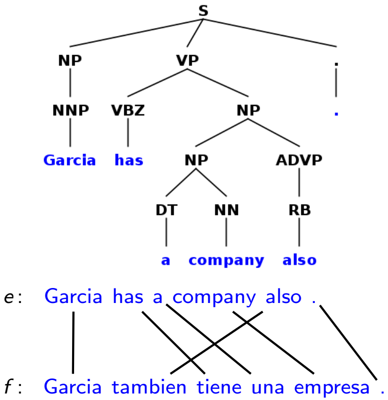




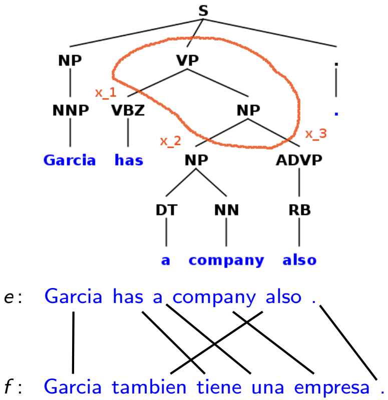
alignment:

0-0 1-2 2-3 3-4 4-1 5-5

alignment graph:



alignment graph:



extracted rule for the tt $\mathcal{N}$ (tree-to-string):

$$q(\mathrm{VP}(x_1 : \mathrm{VBZ}, \mathrm{NP}(x_2 : \mathrm{NP}, x_3 : \mathrm{ADVP}))) \longrightarrow q(x_3)\ q(x_1)\ q(x_2)$$

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities
    - random experiments
    - corpora and maximum-likelihood estimation
    - Expectation-Maximization algorithm
    - instantiation to training probability assignment for tt

random experiment (Y, p):

- finite set Y (events),
- distribution p over Y: $\quad p : Y \to [0, 1]$, $\sum_y p(y) = 1$

random experiment (Y, p):

- finite set Y (events),
- distribution p over Y: $p : Y \to [0, 1]$, $\sum_y p(y) = 1$

set of all distributions over Y: $\mathcal{B}(Y)$
probability model over Y: $\mathcal{B} \subseteq \mathcal{B}(Y)$

random experiment (Y, p):

- finite set Y (events),
- distribution p over Y: $p : Y \to [0, 1]$, $\sum_y p(y) = 1$

set of all distributions over Y: $\mathcal{B}(Y)$
probability model over Y: $\mathcal{B} \subseteq \mathcal{B}(Y)$
unrestricted probability model: $\mathcal{B}(Y)$

random experiment (Y, p):

- finite set Y (events),
- distribution p over Y: $p : Y \to [0,1]$, $\sum_y p(y) = 1$

set of all distributions over Y: $\mathcal{B}(Y)$
probability model over Y: $\mathcal{B} \subseteq \mathcal{B}(Y)$
unrestricted probability model: $\mathcal{B}(Y)$

Let (Y_1, p_1), (Y_2, p_2) two random experiments.

independent product:

$$(Y_1 \times Y_2, p_1 \times p_2)$$

$$(p_1 \times p_2)(y_1, y_2) = p_1(y_1) \cdot p_2(y_2)$$

Examples:

- "tossing a coin"
  - $Y = \{h, t\}$
  - $p(h) = 0.4, \ p(t) = 0.6$

Examples:

- "tossing a coin"
  - $Y = \{h, t\}$
  - $p(h) = 0.4, \; p(t) = 0.6$

- "tossing two coins"
  - $Y = \{h, t\} \times \{h, t\}$
  -

	h	t
p_1	0.4	0.6
p_2	0.5	0.5

  - probability model $\mathcal{B} = \{p_1 \times p_2 \in \mathcal{B}(Y) \mid p_1, p_2 \in \mathcal{B}(\{h, t\})\}$

Examples:

- "tossing a coin"
  - $Y = \{h, t\}$
  - $p(h) = 0.4, \; p(t) = 0.6$

- "tossing two coins"
  - $Y = \{h, t\} \times \{h, t\}$
  -

	h	t
p_1	0.4	0.6
p_2	0.5	0.5

  - probability model $\mathcal{B} = \{p_1 \times p_2 \in \mathcal{B}(Y) \mid p_1, p_2 \in \mathcal{B}(\{h, t\})\}$

- consider distribution $p \in \mathcal{B}(\{h, t\} \times \{h, t\})$:

$$p(h, h) = p(t, t) = 0, \quad p(h, t) = p(t, h) = 0.5$$

Examples:

- "tossing a coin"
  - $Y = \{h, t\}$
  - $p(h) = 0.4, \; p(t) = 0.6$

- "tossing two coins"
  - $Y = \{h, t\} \times \{h, t\}$
  -

	h	t
p_1	0.4	0.6
p_2	0.5	0.5

  - probability model $\mathcal{B} = \{p_1 \times p_2 \in \mathcal{B}(Y) \mid p_1, p_2 \in \mathcal{B}(\{h, t\})\}$

- consider distribution $p \in \mathcal{B}(\{h, t\} \times \{h, t\})$:

  $$p(h, h) = p(t, t) = 0, \quad p(h, t) = p(t, h) = 0.5$$

  observe: $p \notin \mathcal{B}$

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities
    - random experiments
    - corpora and maximum-likelihood estimation
    - Expectation-Maximization algorithm
    - instantiation to training probability assignment for tt

corpus: $c : Y \to \mathbb{R}_{\geq 0}$

$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$

size of c: $|c| = \sum_y c(y)$

corpus: $c : Y \to \mathbb{R}_{\geq 0}$

$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$

size of c: $|c| = \sum_y c(y)$

Let $p \in \mathcal{B}(Y)$ and $c : Y \to \mathbb{R}_{\geq 0}$ corpus

likelihood of c under p: $L(c, p) = \prod_y p(y)^{c(y)}$

corpus: $c : Y \to \mathbb{R}_{\geq 0}$

$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$

size of c: $|c| = \sum_y c(y)$

Let $p \in \mathcal{B}(Y)$ and $c : Y \to \mathbb{R}_{\geq 0}$ corpus

likelihood of c under p: $L(c, p) = \prod_y p(y)^{c(y)}$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus c $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

corpus: $c : Y \to \mathbb{R}_{\geq 0}$
$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$
size of c: $|c| = \sum_y c(y)$

Let $p \in \mathcal{B}(Y)$ and $c : Y \to \mathbb{R}_{\geq 0}$ corpus
likelihood of c under p: $L(c, p) = \prod_y p(y)^{c(y)}$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus c $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

maximum-likelihood estimation of c in $\mathcal{B}$:

$$\hat{p} = \mathrm{argmax}_{p \in \mathcal{B}} L(c, p)$$

corpus: $c : Y \to \mathbb{R}_{\geq 0}$
$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$
size of c: $|c| = \sum_y c(y)$

Let $p \in \mathcal{B}(Y)$ and $c : Y \to \mathbb{R}_{\geq 0}$ corpus
likelihood of c under p: $L(c, p) = \prod_y p(y)^{c(y)}$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus c $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

maximum-likelihood estimation of c in $\mathcal{B}$:

$$\hat{p} = \mathrm{argmax}_{p \in \mathcal{B}} L(c, p) = \mathrm{mle}(c, \mathcal{B})$$

corpus: $c : Y \to \mathbb{R}_{\geq 0}$
$\mathrm{supp}(c)$ finite!

frequency of y: $c(y)$
size of c: $|c| = \sum_y c(y)$

Let $p \in \mathcal{B}(Y)$ and $c : Y \to \mathbb{R}_{\geq 0}$ corpus
likelihood of c under p: $L(c, p) = \prod_y p(y)^{c(y)}$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus c $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

maximum-likelihood estimation of c in $\mathcal{B}$:

$$\hat{p} = \mathrm{argmax}_{p \in \mathcal{B}} L(c, p) = \mathrm{mle}(c, \mathcal{B})$$

note: $\mathrm{mle}(c, \mathcal{B}) \in \mathcal{B}$

Theorem Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus.

- $\mathrm{mle}(c, \mathcal{B}(Y)) = \mathrm{rfe}(c)$

Theorem Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus.

- $\mathrm{mle}(c, \mathcal{B}(Y)) = \mathrm{rfe}(c)$

relative frequency estimation of c:
(empirical distr.)

$$\mathrm{rfe}(c) : Y \to [0,1]$$
$$\mathrm{rfe}(c)(y) = \frac{c(y)}{|c|}$$

$$\mathrm{rfe}(c) \in \mathcal{B}(Y)$$

Example: "tossing of **one** coin"

	h	t
c:	8	12
$\mathrm{rfe}(c)$	0.4	0.6

Theorem Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus.

- $\mathrm{mle}(c, \mathcal{B}(Y)) = \mathrm{rfe}(c)$
- $\forall \mathcal{B} \subseteq \mathcal{B}(Y) \setminus \{\mathrm{rfe}(c)\} : \ \mathrm{mle}(c, \mathcal{B}) \neq \mathrm{rfe}(c)$

relative frequency estimation of c: (empirical distr.)

$$\mathrm{rfe}(c) : Y \to [0, 1]$$
$$\mathrm{rfe}(c)(y) = \frac{c(y)}{|c|}$$

$$\mathrm{rfe}(c) \in \mathcal{B}(Y)$$

Theorem Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus.

- $\mathrm{mle}(c, \mathcal{B}(Y)) = \mathrm{rfe}(c)$
- $\forall \mathcal{B} \subseteq \mathcal{B}(Y) \setminus \{\mathrm{rfe}(c)\} : \ \mathrm{mle}(c, \mathcal{B}) \neq \mathrm{rfe}(c)$

relative frequency estimation of c:
(empirical distr.)

$$\mathrm{rfe}(c) : Y \to [0, 1]$$
$$\mathrm{rfe}(c)(y) = \frac{c(y)}{|c|}$$

$$\mathrm{rfe}(c) \in \mathcal{B}(Y)$$

Example: "tossing **two** coins"

	(h, h)	(h, t)	(t, h)	(t, t)
c:	0	5	5	0
$\mathrm{rfe}(c)$	0	1/2	1/2	0

$$\mathrm{rfe}(c) \notin \{p_1 \times p_2 \mid p_1, p_2 \in \mathcal{B}(\{h, t\})\}$$

Theorem Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus. **but ...**

- $\mathrm{mle}(c, \mathcal{B}(Y)) = \mathrm{rfe}(c)$
- $\forall \mathcal{B} \subseteq \mathcal{B}(Y) \setminus \{\mathrm{rfe}(c)\} : \ \mathrm{mle}(c, \mathcal{B}) \neq \mathrm{rfe}(c)$

relative frequency estimation of c:
(empirical distr.)

$$\mathrm{rfe}(c) : Y \to [0, 1]$$
$$\mathrm{rfe}(c)(y) = \frac{c(y)}{|c|}$$

$$\mathrm{rfe}(c) \in \mathcal{B}(Y)$$

Example: “tossing **two** coins”

	(h,h)	(h,t)	(t,h)	(t,t)
c:	0	5	5	0
$\mathrm{rfe}(c)$	0	1/2	1/2	0

$$\mathrm{rfe}(c) \notin \{p_1 \times p_2 \mid p_1, p_2 \in \mathcal{B}(\{h, t\})\}$$

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and $\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and $\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Then: $\mathrm{mle}(c, \mathcal{B}) = \mathrm{rfe}(c_1) \times \mathrm{rfe}(c_2)$

where $c_1(y_1) = \sum_{y_2} c(y_1, y_2)$
$c_2(y_2) = \sum_{y_1} c(y_1, y_2)$

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and $\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Then: $\mathrm{mle}(c, \mathcal{B}) = \mathrm{rfe}(c_1) \times \mathrm{rfe}(c_2)$

where $c_1(y_1) = \sum_{y_2} c(y_1, y_2)$
$c_2(y_2) = \sum_{y_1} c(y_1, y_2)$

Example: "tossing two coins" $\quad (Y_1 = Y_2 = \{h, t\})$

corpus:

	(h,h)	(h,t)	(t,h)	(t,t)
c	0	5	5	0

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and
$\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Then: $\mathrm{mle}(c, \mathcal{B}) = \mathrm{rfe}(c_1) \times \mathrm{rfe}(c_2)$

where $c_1(y_1) = \sum_{y_2} c(y_1, y_2)$
$c_2(y_2) = \sum_{y_1} c(y_1, y_2)$

Example: "tossing two coins" $(Y_1 = Y_2 = \{h, t\})$

corpus:

	(h,h)	(h,t)	(t,h)	(t,t)
c	0	5	5	0

for $i \in \{1, 2\}$:

	h	t
c_i	5	5
$\mathrm{rfe}(c_i)$	1/2	1/2

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and $\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Then: $\mathrm{mle}(c, \mathcal{B}) = \mathrm{rfe}(c_1) \times \mathrm{rfe}(c_2)$

where $c_1(y_1) = \sum_{y_2} c(y_1, y_2)$

$c_2(y_2) = \sum_{y_1} c(y_1, y_2)$

Example: "tossing two coins" $(Y_1 = Y_2 = \{h, t\})$

corpus:

	(h,h)	(h,t)	(t,h)	(t,t)
c	0	5	5	0

for $i \in \{1, 2\}$:

	h	t
c_i	5	5
$\mathrm{rfe}(c_i)$	1/2	1/2

corpus:

	(h,h)	(t,h)	(h,t)	(t,t)
$\mathrm{mle}(c, \mathcal{B})$	1/4	1/4	1/4	1/5

Observation Let $c : Y \to \mathbb{R}_{\geq 0}$ corpus, $Y = Y_1 \times Y_2$, and $\mathcal{B} = \{p_1 \times p_2 \mid p_1 \in \mathcal{B}(Y_1), p_2 \in \mathcal{B}(Y_2)\}$.

Then: $\mathrm{mle}(c, \mathcal{B}) = \mathrm{rfe}(c_1) \times \mathrm{rfe}(c_2)$

$$\text{where } c_1(y_1) = \sum_{y_2} c(y_1, y_2)$$
$$c_2(y_2) = \sum_{y_1} c(y_1, y_2)$$

Example: "tossing two coins" $(Y_1 = Y_2 = \{h, t\})$

corpus:

	(h,h)	(h,t)	(t,h)	(t,t)
c	0	5	5	0

for $i \in \{1, 2\}$:

	h	t
c_i	5	5
$\mathrm{rfe}(c_i)$	1/2	1/2

corpus:

	(h,h)	(t,h)	(h,t)	(t,t)
$\mathrm{mle}(c, \mathcal{B})$	1/4	1/4	1/4	1/4

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities
    - random experiments
    - corpora and maximum-likelihood estimation
    - Expectation-Maximization algorithm
    - instantiation to training probability assignment for tt

Example: “tossing two coins and hiding information”

player A:

- tosses two coins several times,
- each time she maintains the number of heads, and
- eventually forms the corresponding corpus $c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$

Example: $c(0) = 4, \quad c(1) = 9, \quad c(2) = 2$

Example: "tossing two coins and hiding information"

player A:

- tosses two coins several times,
- each time she maintains the number of heads, and
- eventually forms the corresponding corpus $c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$

Example: $c(0) = 4, \quad c(1) = 9, \quad c(2) = 2$

player B:

- gets a corpus $c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$ of "observations" and
- has to estimate the distribution p_1 and p_2 of the coins.

Example: "tossing two coins and hiding information"

player A:

- tosses two coins several times,
- each time she maintains the number of heads, and
- eventually forms the corresponding corpus $c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$

Example: $c(0) = 4, \quad c(1) = 9, \quad c(2) = 2$

player B:

- gets a corpus $c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$ of "observations" and
- has to estimate the distribution p_1 and p_2 of the coins.

$c : \{0, 1, 2\} \to \mathbb{R}_{\geq 0}$ is <u>incomplete data</u>

set of events:	Y
set of observations:	X
observation mapping:	$\pi : Y \to X$

Example: "tossing two coins and hiding information"

- set of events $Y = \{(h, h),\ (h, t),\ (t, h),\ (t, t)\}$
- set of observations: $X = \{0, 1, 2\}$
- observation mapping:

y	(h, h)	(h, t)	(t, h)	(h, h)
$\pi(y)$	2	1	1	0

set of events: Y

set of observations: X

observation mapping: $\pi : Y \to X$

set of events: Y

set of observations: X

observation mapping: $\pi : Y \to X$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus $c : X \to \mathbb{R}_{\geq 0}$ $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

set of events: Y

set of observations: X

observation mapping: $\pi : Y \to X$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus $c : X \to \mathbb{R}_{\geq 0}$ $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

likelihood of c under $p \in \mathcal{B}$:

$$L(c, p) = \prod_x \Big(\sum_{y:\pi(y)=x} p(y) \Big)^{c(x)}$$

set of events: Y

set of observations: X

observation mapping: $\pi : Y \to X$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and
corpus $c : X \to \mathbb{R}_{\geq 0}$ $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

likelihood of c under $p \in \mathcal{B}$:

$$L(c, p) = \prod_x \Big(\sum_{y:\pi(y)=x} p(y) \Big)^{c(x)}$$

maximum-likelihood estimation:

$$\hat{p} = \mathrm{argmax}_{p \in \mathcal{B}} L(c, p)$$

set of events: Y

set of observations: X

observation mapping: $\pi : Y \to X$

$\mathcal{B} \subseteq \mathcal{B}(Y)$ and corpus $c : X \to \mathbb{R}_{\geq 0}$ $\longrightarrow$ training $\longrightarrow$ $\hat{p} \in \mathcal{B}$

likelihood of c under $p \in \mathcal{B}$:

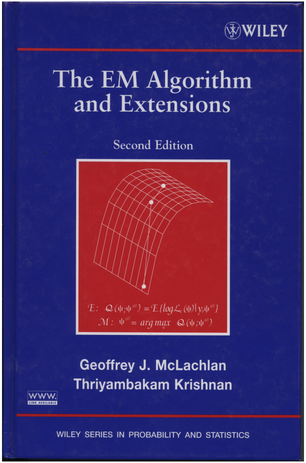
$$L(c, p) = \prod_x \Big(\sum_{y:\pi(y)=x} p(y) \Big)^{c(x)}$$

maximum-likelihood estimation:

$$\hat{p} = \mathrm{argmax}_{p \in \mathcal{B}} L(c, p) = \mathrm{mle}(c, \mathcal{B})$$

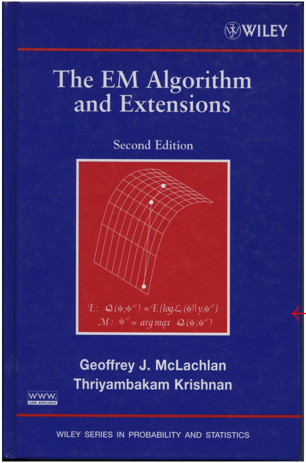
[Dempster, Laird, Rubin 77]
Maximum likelihood from incomplete data via the EM algorithm,
Journal of the Royal Statistical Society, 39:1–38.

[Dempster, Laird, Rubin 77]
Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39:1–38.



359 pages, 2008

[Dempster, Laird, Rubin 77]
Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39:1–38.



$$\mathcal{E}: \quad Q(\psi;\psi^{(k)}) = E_{\psi^{(k)}}\{\log L(c,\psi)\}$$
$$\mathcal{M}: \quad \psi^{(k+1)} = \mathrm{argmax}_{\psi}\, Q(\psi;\psi^{(k)})$$

359 pages, 2008

Expectation-Maximization algorithm [Prescher 05]

input corpus $c : X \to \mathbb{R}_{\geq 0}$;
probability model $\mathcal{B} \subseteq \mathcal{B}(Y)$, starting distribution $p_0 \in \mathcal{B}$.
observation mapping $\pi : Y \to X$

output sequence $p_1, p_2, p_3 \ldots \in \mathcal{B}$

Expectation-Maximization algorithm [Prescher 05]

input corpus $c : X \to \mathbb{R}_{\geq 0}$;
probability model $\mathcal{B} \subseteq \mathcal{B}(Y)$, starting distribution $p_0 \in \mathcal{B}$.
observation mapping $\pi : Y \to X$

output sequence $p_1, p_2, p_3 \ldots \in \mathcal{B}$

for each $i = 1, 2, 3, \ldots$

E-step: compute corpus $c_{p_{i-1}} : Y \to \mathbb{R}_{\geq 0}$ expected by p_{i-1}:

$$c_{p_{i-1}}(y) := c(\pi(y)) \cdot \Big(p_{i-1}(y) \,/ \sum_{y' : \pi(y') = \pi(y)} p_{i-1}(y') \Big)$$

M-step: compute maximum-likelihood estimate:

$$p_i := \mathrm{mle}(c_{p_{i-1}}, \mathcal{B})$$

print p_i

Theorem [Dempster, Laird, Rubin 77]

Let $c : X \to \mathbb{R}_{\geq 0}$ be a corpus,
$\mathcal{B} \subseteq \mathcal{B}(Y)$, $p_0 \in \mathcal{B}$,
$\pi : Y \to X$ observation mapping.

Theorem [Dempster, Laird, Rubin 77]

Let $c : X \to \mathbb{R}_{\geq 0}$ be a corpus,
$\mathcal{B} \subseteq \mathcal{B}(Y)$, $\quad p_0 \in \mathcal{B}$,
$\pi : Y \to X$ observation mapping.

If $p_1, p_2, p_3, \ldots$ is generated by the EM algorithm,

then $L(c, p_0) \leq L(c, p_1) \leq L(c, p_2) \leq \ldots \leq \mathrm{mle}(c, \mathcal{B})$.

Example: “tossing of two coins and hiding information”

$$c(0) = 4 \quad c(1) = 9 \quad c(2) = 2$$

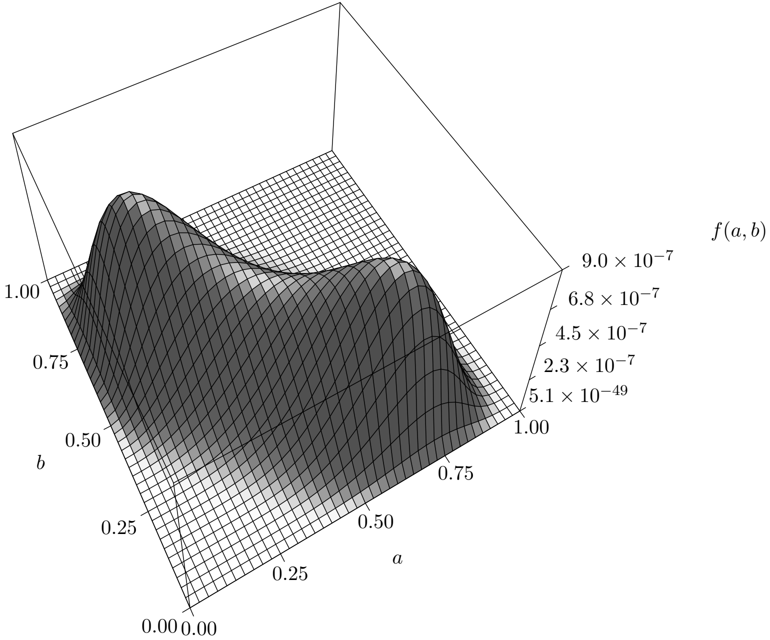
	run 1	run 2	run 3	run 4
$(p_1^0(h), p_2^0(h))$	$(0.200, 0.500)$	$(0.900, 0.600)$	$(0.000, 1.000)$	$(0.400, 0.400)$
$(p_1^1(h), p_2^1(h))$	$(0.253, 0.613)$	$(0.648, 0.219)$	$(0.133, 0.733)$	$(0.433, 0.433)$
$(p_1^2(h), p_2^2(h))$	$(0.239, 0.628)$	$(0.654, 0.213)$	$(0.165, 0.687)$	$(0.433, 0.433)$
$(p_1^3(h), p_2^3(h))$	$(0.228, 0.639)$	$(0.658, 0.208)$	$(0.180, 0.679)$	$(0.433, 0.433)$
$(p_1^4(h), p_2^4(h))$	$(0.219, 0.648)$	$(0.661, 0.205)$	$(0.188, 0.674)$	$(0.433, 0.433)$
$(p_1^5(h), p_2^5(h))$	$(0.213, 0.654)$	$(0.663, 0.204)$	$(0.193, 0.671)$	$(0.433, 0.433)$
...	...	...	...	...
$(p_1^{20}(h), p_2^{20}(h))$	$(0.200, 0.667)$	$(0.667, 0.200)$	$(0.200, 0.667)$	$(0.433, 0.433)$

EM-algorithm converges to either of the three distributions:

	h	t
p_1	$1/5$	$4/5$
p_2	$2/3$	$1/3$

	h	t
p_1	$2/3$	$1/3$
p_2	$1/5$	$4/5$

	h	t
p_1	$13/30$	$17/30$
p_2	$13/30$	$13/30$


$f(a,b)$
9.0×10^{-7}
6.8×10^{-7}
4.5×10^{-7}
2.3×10^{-7}
5.1×10^{-49}
1.00
0.75
0.50
0.25
0.00
b
1.00
0.75
0.50
0.25
0.00
a

outline of the talk:

- Statistical machine translation (recall ...)
- Modeling with wta and wtt (recall ...)
- Training
  - Rule extraction
  - Training probabilities
    - random experiments
    - corpora and maximum-likelihood estimation
    - Expectation-Maximization algorithm
    - instantiation to training probability assignment for tt

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events: $Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)

set of observations: $X = T_\Sigma \times T_\Delta$

observation mapping: $\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$
$\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d

probability assignment: $p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events: $Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)

set of observations: $X = T_\Sigma \times T_\Delta$

observation mapping: $\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$
$\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d

probability assignment: $p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$
$p(q)$

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events: $Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)

set of observations: $X = T_\Sigma \times T_\Delta$

observation mapping: $\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$
$\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d

probability assignment: $p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$
$p(q)(l, r)$

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events:	$Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)
set of observations:	$X = T_\Sigma \times T_\Delta$
observation mapping:	$\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$ $\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d
probability assignment:	$p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$ $p(q)(l, r) \rightsquigarrow p(q(l) \to r)$

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events: $Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)

set of observations: $X = T_\Sigma \times T_\Delta$

observation mapping: $\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$
$\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d

probability assignment: $p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$
$p(q)(l, r) \quad \rightsquigarrow \quad p(q(l) \to r)$

Let p be a prob. assignment. Extend p to $\widetilde{p} \in \mathcal{B}(D_{\mathcal{N}})$:

$$\widetilde{p}\Big((q_1(l_1) \to r_1) \dots (q_n(l_n) \to r_n)\Big) = \prod_{i=1}^{n} p(q_i(l_i) \to r_i)$$

Let $\mathcal{N} = (Q, \Sigma, \Delta, q_0, R)$ extended tree transducer

set of events:	$Y = D_{\mathcal{N}}$ (set of derivations of $\mathcal{N}$)
set of observations:	$X = T_\Sigma \times T_\Delta$
observation mapping:	$\pi : D_{\mathcal{N}} \to T_\Sigma \times T_\Delta$ $\pi(d) = (\xi, \zeta)$: retrieve ξ, ζ from d
probability assignment:	$p : Q \to \mathcal{B}(\mathrm{LHS} \times \mathrm{RHS})$ $p(q)(l, r) \quad \rightsquigarrow \quad p(q(l) \to r)$

Let p be a prob. assignment. Extend p to $\widetilde{p} \in \mathcal{B}(D_{\mathcal{N}})$:

$$\widetilde{p}\Big((q_1(l_1) \to r_1) \dots (q_n(l_n) \to r_n)\Big) = \prod_{i=1}^{n} p(q_i(l_i) \to r_i)$$

probability model:

$$\mathcal{B}_{\mathcal{N}} = \{\widetilde{p} \in \mathcal{B}(D_{\mathcal{N}}) \mid \text{ probability assignment } p\}$$

Expectation-Maximization algorithm, i

input corpus $c : T_\Sigma \times T_\Delta \to \mathbb{R}_{\geq 0}$;
probability model $\mathcal{B}_\mathcal{N} \subseteq \mathcal{B}(D_\mathcal{N})$, starting distribution $p_0 \in \mathcal{B}_\mathcal{N}$.
observation mapping $\pi : D_\mathcal{N} \to T_\Sigma \times T_\Delta$

output sequence $p_1, p_2, p_3 \ldots \in \mathcal{B}_\mathcal{N}$

Expectation-Maximization algorithm, i

input corpus $c : T_\Sigma \times T_\Delta \to \mathbb{R}_{\geq 0}$;
probability model $\mathcal{B}_\mathcal{N} \subseteq \mathcal{B}(D_\mathcal{N})$, starting distribution $p_0 \in \mathcal{B}_\mathcal{N}$.
observation mapping $\pi : D_\mathcal{N} \to T_\Sigma \times T_\Delta$

output sequence $p_1, p_2, p_3 \ldots \in \mathcal{B}_\mathcal{N}$

for each $i = 1, 2, 3, \ldots$

E-step: compute corpus $c_{p_{i-1}} : D_\mathcal{N} \to \mathbb{R}_{\geq 0}$ expected by p_{i-1}:

$$c_{p_{i-1}}(d) := c(\pi(d)) \cdot \Big(p_{i-1}(d) / \sum_{d' : \pi(d') = \pi(d)} p_{i-1}(d')\Big)$$

M-step: compute maximum-likelihood estimate:

$$p_i := \mathrm{mle}(c_{p_{i-1}}, \mathcal{B}_\mathcal{N})$$

print p_i

some more calculations ... yield ...

Require:

corpus $c : T_\Sigma \times T_\Delta \to \mathbb{R}_{\geq 0}$ with $\mathrm{supp}(c) = \{(\xi_1, \zeta_1), \ldots, (\xi_n, \zeta_n)\}$
some initial prob. assignment $p_0 : R \to \mathbb{R}_{\geq 0}$

Ensure:

approximation of a local maximum or saddle point of
$L(c, .) : {\mathbb{R}_{\geq 0}}^R \to \mathbb{R}_{\geq 0}$ with $L(c, p) = \prod_{(\xi,\zeta) \in \mathrm{supp}(c)} P(\xi, \zeta)^{c(\xi,\zeta)}$.

Variables: $p : R \to \mathbb{R}_{\geq 0}$, $\mathrm{count} : R \to \mathbb{R}_{\geq 0}$, $\gamma \in \mathbb{R}_{\geq 0}$

Approach: compute a sequence of probability assignments $p_0, p_1, p_2, \ldots$
such that $L(c, p_i) \leq L(c, p_{i+1})$.

for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
  construct the RTG $\mathrm{Prod}(\xi, \mathcal{N}, \zeta)$

for all $i = 1, 2, 3, \ldots$ **do**
  $\mathrm{count}(\rho) := 0$ for every $\rho \in R$;
  $p := p_{i-1}$

  for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
    let $\mathcal{G} = (\mathrm{Prod}(\xi, \mathcal{N}, \zeta), p)$ and $\mathrm{Prod}(\xi, \mathcal{N}, \zeta) = (N, R, S, R')$;
    compute $out_\mathcal{G}$ and $in_\mathcal{G}$;
    for all $(A \to \tau) \in R'$ **do**
      $\gamma := out_\mathcal{G}(A) \cdot p(\tau(\varepsilon)) \cdot in_\mathcal{G}(\tau)$;
      $\mathrm{count}(\tau(\varepsilon)) := \mathrm{count}(\tau(\varepsilon)) + c(\xi, \zeta) \cdot \frac{\gamma}{in_\mathcal{G}(S)}$

  for all $\rho = (q(l) \to r) \in R$ **do**
    $p_i(\rho) := \mathrm{count}(\rho) \cdot \Big(\sum_{\rho' \in R_q} \mathrm{count}(\rho') \Big)^{-1}$

  $\mathrm{output}(p_i)$

[Graehl, Knight 04]

Require:

corpus $c : T_\Sigma \times T_\Delta \to \mathbb{R}_{\geq 0}$ with $\mathrm{supp}(c) = \{(\xi_1, \zeta_1), \ldots, (\xi_n, \zeta_n)\}$
some initial prob. assignment $p_0 : R \to \mathbb{R}_{\geq 0}$

Ensure:

approximation of a local maximum or saddle point of
$L(c, .) : {\mathbb{R}_{\geq 0}}^R \to \mathbb{R}_{\geq 0}$ with $L(c, p) = \prod_{(\xi,\zeta) \in \mathrm{supp}(c)} P(\xi, \zeta)^{c(\xi,\zeta)}$.

Variables: $p : R \to \mathbb{R}_{\geq 0}$, $\mathrm{count} : R \to \mathbb{R}_{\geq 0}$, $\gamma \in \mathbb{R}_{\geq 0}$

Approach: compute a sequence of probability assignments $p_0, p_1, p_2, \ldots$
such that $L(c, p_i) \leq L(c, p_{i+1})$.

for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
  construct the RTG $\mathrm{Prod}(\xi, \mathcal{N}, \zeta)$

for all $i = 1, 2, 3, \ldots$ **do**
  $\mathrm{count}(\rho) := 0$ for every $\rho \in R$;
  $p := p_{i-1}$

  for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
    let $\mathcal{G} = (\mathrm{Prod}(\xi, \mathcal{N}, \zeta), p)$ and $\mathrm{Prod}(\xi, \mathcal{N}, \zeta) = (N, R, S, R')$;
    compute $out_\mathcal{G}$ and $in_\mathcal{G}$;
    for all $(A \to \tau) \in R'$ **do**
      $\gamma := out_\mathcal{G}(A) \cdot p(\tau(\varepsilon)) \cdot in_\mathcal{G}(\tau)$;
      $\mathrm{count}(\tau(\varepsilon)) := \mathrm{count}(\tau(\varepsilon)) + c(\xi, \zeta) \cdot \frac{\gamma}{in_\mathcal{G}(S)}$

  for all $\rho = (q(l) \to r) \in R$ **do**
    $p_i(\rho) := \mathrm{count}(\rho) \cdot \Big(\sum_{\rho' \in R_q} \mathrm{count}(\rho') \Big)^{-1}$

  output(p_i)

[Graehl, Knight 04]
$\rightsquigarrow$ VANDA

Require:

corpus $c : T_\Sigma \times T_\Delta \to \mathbb{R}_{\geq 0}$ with $\mathrm{supp}(c) = \{(\xi_1, \zeta_1), \ldots, (\xi_n, \zeta_n)\}$
some initial prob. assignment $p_0 : R \to \mathbb{R}_{\geq 0}$

Ensure:

approximation of a local maximum or saddle point of
$L(c, .) : {\mathbb{R}_{\geq 0}}^R \to \mathbb{R}_{\geq 0}$ with $L(c, p) = \prod_{(\xi,\zeta) \in \mathrm{supp}(c)} P(\xi, \zeta)^{c(\xi,\zeta)}$.

Variables: $p : R \to \mathbb{R}_{\geq 0}$, $\mathrm{count} : R \to \mathbb{R}_{\geq 0}$, $\gamma \in \mathbb{R}_{\geq 0}$

Approach: compute a sequence of probability assignments $p_0, p_1, p_2, \ldots$ such that $L(c, p_i) \leq L(c, p_{i+1})$.

for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
  construct the RTG $\mathrm{Prod}(\xi, \mathcal{N}, \zeta)$

for all $i = 1, 2, 3, \ldots$ **do**
  $\mathrm{count}(\rho) := 0$ for every $\rho \in R$;
  $p := p_{i-1}$

  for all $(\xi, \zeta) \in \mathrm{supp}(h)$ **do**
    let $\mathcal{G} = (\mathrm{Prod}(\xi, \mathcal{N}, \zeta), p)$ and $\mathrm{Prod}(\xi, \mathcal{N}, \zeta) = (N, R, S, R')$;
    compute $out_\mathcal{G}$ and $in_\mathcal{G}$;
    for all $(A \to \tau) \in R'$ **do**
      $\gamma := out_\mathcal{G}(A) \cdot p(\tau(\varepsilon)) \cdot in_\mathcal{G}(\tau)$;
      $\mathrm{count}(\tau(\varepsilon)) := \mathrm{count}(\tau(\varepsilon)) + c(\xi, \zeta) \cdot \frac{\gamma}{in_\mathcal{G}(S)}$

  for all $\rho = (q(l) \to r) \in R$ **do**
    $p_i(\rho) := \mathrm{count}(\rho) \cdot \Big(\sum_{\rho' \in R_q} \mathrm{count}(\rho') \Big)^{-1}$

  $\mathrm{output}(p_i)$

References:

- [Arnold, Dauchet 82] Morphismes et bimorphismes d'arbes, *Theoretical Computer Science*, 20(1):33-93.
- [Dempster, Laird, Rubin 77] Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39:1–38.
- [Fülöp, Maletti, Vogler 11] Weighted extended tree transducers, *Fundamenta Informaticae* 111(2): 163–202.
- [Galley, Hopkins, Knight, Marcu 04] What's in a translation rule? Proc. HLT-NAACL, Association for Computational Linguistics, 273–280.
- [Graehl, Knight 04] Training tree transducers, HLT-NAACL, Association for Computational Linguistics, 105-112.
- [Lopez 08] Statistical Machine Translation, *ACM Computing Surveys*, 40(3): 8:1–8:9.
- [Liang, Bouchard-Côté, Klein, Taskar 06] An End-to-End Discriminative Approach to Machine Translation, Proc. 21st Int. Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. Comput. Ling., 761–768.
- [Prescher 05] A Tutorial on the Expectation-Maximization Algorithm ..., arXiv:cs/0412015v2
- [Stüber 12] personal communication

References:

- [Arnold, Dauchet 82] Morphismes et bimorphismes d'arbes, *Theoretical Computer Science*, 20(1):33-93.
- [Dempster, Laird, Rubin 77] Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39:1–38.
- [Fülöp, Maletti, Vogler 11] Weighted extended tree transducers, *Fundamenta Informaticae* 111(2): 163–202.
- [Galley, Hopkins, Knight, Marcu 04] What's in a translation rule? Proc. HLT-NAACL, Association for Computational Linguistics, 273–280.
- [Graehl, Knight 04] Training tree transducers, HLT-NAACL, Association for Computational Linguistics, 105-112.
- [Lopez 08] Statistical Machine Translation, *ACM Computing Surveys*, 40(3): 8:1–8:9.
- [Liang, Bouchard-Côté, Klein, Taskar 06] An End-to-End Discriminative Approach to Machine Translation, Proc. 21st Int. Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. Comput. Ling., 761–768.
- [Prescher 05] A Tutorial on the Expectation-Maximization Algorithm ..., arXiv:cs/0412015v2
- [Stüber 12] personal communication

thanks

set of all distributions over Y: $\mathcal{B}(Y)$
probability model over Y: $\mathcal{B} \subseteq \mathcal{B}(Y)$
corpus: $c : Y \to \mathbb{R}_{\geq 0}$
size of c: $|c| = \sum_y c(y)$
relative frequency estimation: $\mathrm{rfe}(c)(y) = \frac{c(y)}{|c|}$

likelihood of Y-corpus c under $p \in \mathcal{B}(Y)$:

$$L(c, p) = \prod_y p(y)^{c(y)}$$

maximum-likelihood estimation of Y-corpus c in $\mathcal{B} \subseteq \mathcal{B}(Y)$:

$$\mathrm{mle}(c, \mathcal{B}) = \mathrm{argmax}_{p \in \mathcal{B}}\, L(c, p)$$

likelihood of X-corpus c under $p \in \mathcal{B}(Y)$:

$$L(c, p) = \prod_x \Big(\sum_{y : \pi(y) = x} p(y)\Big)^{c(x)}$$

maximum-likelihood estimation of X-corpus c in $\mathcal{B} \subseteq \mathcal{B}(Y)$:

$$\mathrm{mle}(c, \mathcal{B}) = \mathrm{argmax}_{p \in \mathcal{B}}\, L(c, p)$$