重み付き有限オートマトンに対する曖昧性解消演算の一般化とその応用

A generalized disambiguation algorithm for weighted finite automata and its application to NLP tasks

林 克彦 [1*]　　永田 昌明 [1]
Katsuhiko Hayashi[1]　　Masaaki Nagata[1]

[1] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
[1] NTT Communication Science Laboratories, NTT Corp.

Abstract: We present a disambiguation algorithm for weighted finite tree automata (FTA). This algorithm converts ambiguous FTA into equivalent non-ambiguous one where no two accepting paths labeled with the same tree exists. The notion of non-ambiguity is similar to that of determinism in the automata theory, but we show that disambiguation is applicable to the wider class of weighted automata than determinization. We conduct experiments on Natural Language Processing (NLP) tasks, and also show that disambiguated automata become much smaller than determinized automata in practice.

1　はじめに

この約 50 年の間，**有限木オートマトン** (Finite tree automata: FTA) [16] は盛んに研究が行われてきた [4]. FTA は有限オートマトン (Finite state automata: FSA) の自然な一般化であり，記号列ではなく，木を受理できるように拡張されたオートマトンである．重み付き FTA [6] もまた盛んに研究が行われ，自然言語処理 [8, 3] や画像分析 [5] などの分野で応用が進んでいる．

そのような応用で問題となるのは，重み付き FTA の**曖昧性解消**である．曖昧性解消は FTA を等価で**無曖昧な** FTA へと変換する操作である．無曖昧な FTA では全ての受理可能な木に対して，高々1 つの受理経路 (実行) だけが存在する [4]．曖昧性のある重み付き FTA では，ある受理可能な木に対する真の重みは複数の受理実行に分割されてしまう．よって，曖昧性解消の目的は，

- ある木に対する複数の受理実行を 1 つの実行にまとめること，
- それら複数の受理実行に対する重みの合計をその 1 つの実行に割り当てること

である．このような無曖昧な重み付き FTA 上では，最短経路問題は重みが最良の木を求める問題に一致する．よって，ダイクストラ法 (クヌース 1977 法) [10] や K 最短路ビタビ法 [9] によって，最良の木や重みが上位 K 個の木を効率的に求めることができる．

従来，重み付き FTA に対する曖昧性解消は，**決定化演算**によって行われてきた [11]．決定性の FTA では，ある状態の集合 (ベクトル) から同じ記号による遷移は 1 つ以上存在しない．決定性の FTA は無曖昧であり [4], May と Knight は自然言語の構文解析や機械翻訳タスクにおいて，決定化が最良解の質向上や重複の無い重み上位 K 個の解を求めるのに有用であることを示した [11]. しかし一方で，サイズ N のオートマトンに対して，最悪の場合，その等価な決定性のオートマトンのサイズは $\Omega(2^n)$ となることが知られており [15, 4]，決定性のオートマトンは記憶容量の点で大きな問題がある．

より小さな無曖昧のオートマトンを直接的に構築するため，Mohri は FSA に対する曖昧性解消演算 [13] を提案し，さらにそれを重み付き FSA へと拡張した [14]. 本稿では曖昧性解消演算を重み付き FTA へ適用可能な形に拡張する．通常の FSA と FTA の最も重要な違いは，FTA では k 個の状態ベクトルから 1 つの状態への遷移を許す点である．よって，提案法はそのような遷移の扱いが可能なように拡張した点で，従来法とは大きく異なる．また，提案法が閉路を含む重み付き FTA へ適用可能な十分条件を示し (**weak twins property**: **弱双子性質**)，May と Knight の決定化演算よりも広い範囲の FTA が扱えることを示す．ちなみに，提案法は閉路を含まないどんな重み付き FTA にも適用可能である．

本稿では自然言語処理における文圧縮 [3] と機械翻訳 [8] タスクで実験を行った．これらは重み付き木トランスデューサ [12] によってモデル化でき，その出力空間は重み付き FTA で表現される．実験ではこの出力空間に対し，May と Knight の決定化，及び，提案法を適用したときに構築されるオートマトンのサイズを比較した．結果からは決定化に比べ，提案法は約 90 倍小さな無曖昧のオートマトンを構築できることがわかった．

*連絡先: NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: hayashi.katsuhiko@lab.ntt.co.jp

2 諸々の定義

モノイドは組 $(\mathbb{S},\otimes,\bar{1})$ で定義され，$\otimes$ は $\mathbb{S}$ 上において閉じており，結合性を満たす二項演算である．$\bar{1}$ は $\otimes$ に対する単位元である．モノイドが**可換性**を満たすとは，$\otimes$ が可換性を満たすときをいう．**半環**は 5 つ組 $(\mathbb{S},\oplus,\otimes,\bar{0},\bar{1})$ で定義され，$(\mathbb{S},\oplus,\bar{0})$ は可換性を満たすモノイド，$(\mathbb{S},\otimes,\bar{1})$ はモノイドである．$\otimes$ は $\oplus$ に対して**分配性**が成り立ち，$\bar{0}$ は $\otimes$ に対する零化域である．半環が可換性を満たすとは，$\otimes$ が可換性を満たすときをいう．もし，$\mathbb{S}$ 中の要素 x，x'，z (ただし，$z\neq\bar{0}$) に対して，$x\otimes z = x'\otimes z$ ならば $x = x'$ となるとき，$\otimes$ は**消約性**を満たすという．半環が消約性を満たすとは，$\otimes$ が消約性を満たすときをいう．もし，$\mathbb{S}\backslash\{\bar{0}\}$ 中の要素 x が $x'\otimes x = \bar{1}$ となる**左逆元** x' ($x'\in\mathbb{S}$) を持つとき，半環を**左可除性**を満たすという．もし，$x\otimes x'\neq\bar{0}$ となる要素 $x,x'\in\mathbb{S}$ に対して，$x = (x\otimes x')\otimes z$ ($z\in\mathbb{S}$) となる z が少なくとも 1 つ存在するとき，半環は**弱左可除性**を満たすという．$\otimes$ が消約性を満たすとき，z は単一であり，$z = (x\times x')^{-1}\times x$ と書ける．

ランク付きアルファベットは組 (Σ,rk) とし，記号の有限集合 Σ と写像 $rk:\Sigma\to\mathbb{N}$ から成る．写像 rk は Σ の全ての要素にランクを付与する．全ての $k\in\mathbb{N}$ に対し，$\Sigma^{(k)}\subseteq\Sigma$ を $rk(\sigma)=k$ となる記号 $\sigma\in\Sigma$ の集合とする．$\sigma\in\Sigma^{(k)}$ を $\sigma^{(k)}$ と書くが，自明な場合は (k) を省略する．T_Σ は次の条件を満たす最小の集合とする．

- 全ての $\sigma\in\Sigma^{(0)}$ に対して，σ から成る単一の頂点 $\sigma()$ は T_Σ に含まれる木である．
- 全ての $\sigma\in\Sigma^{(k)}$ $(k\geq 1)$ と $t_1,\ldots,t_k\in T_\Sigma$ に対して，σ でラベル付けされ，木 $t_1,\ldots,t_k$ を子に持つ $\sigma(t_1,\ldots,t_k)$ は T_Σ に含まれる木である．

木 $t=\sigma(t_1,\ldots,t_k)$ の位置集合 $Pos(t)$ を

$$Pos(t)=\{root\}\cup\{i.pos|1\leq i\leq k, pos\in Pos(t_i)\}$$

として定義する．$root$ は t のルート記号の位置とする．葉の位置集合 $leaves(t)$ は

$$leaves(t)=\{pos|pos\in Pos(t),\forall i\in\mathbb{N}, pos.i\notin Pos(t)\}$$

として定義する．木 t のサイズは $|Pos(t)|$ とし，位置 pos にある記号は $t(pos)$ として書く．また，木 t の位置 pos の部分木を $t|_{pos}$，位置 pos の部分木の木 s との置き換えを $t[s]_{pos}$ とする．木 t の高さは $height(t)=1+\max(height(t|_i)|1\leq i\leq rk(t(root)))$ として定義する．例えば，木 $t=\sigma(\gamma(\alpha,\beta(z)),\alpha,z(z))$ を考えると，$Pos(t)=\{root,1,1.1,1.2,1.2.1,2,3,3.1\}$，$leaves(t)=\{1.1,1.2.1,2,3.1\}$ $size(t)=8$，$t(root)=\sigma$，$t(1.1)=\alpha$，$t|_{1.2}=\beta(z)$，$t[\beta]_1=\sigma(\beta,\alpha,z(z))$，$height(t)=4$ となる．

3 重み付き有限木オートマトン

重み付き木オートマトン (FTA) A は半環 $(\mathbb{S},\oplus,\otimes,\bar{0},\bar{1})$ 上において (Σ,Q,i,F,E,ρ) として定義される．

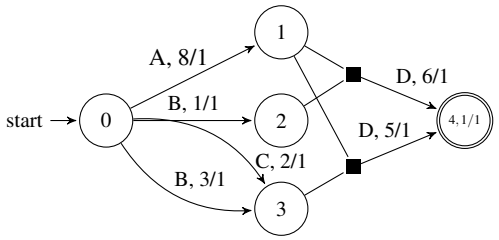

図 1: 半環 $(R_+,+,\times,0,1)$ 上での重み付き木オートマトンの例．全ての重みは分母も表示している．

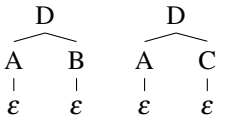

図 2: 図 1 のオートマトンで受理可能な木.

- Σ はランク付きアルファベット,
- Q は状態の有限集合,
- $i\in Q$ は初期状態,
- $F\subseteq Q$ は終了状態の集合,
- $E\subseteq\overbrace{Q\times\cdots\times Q}^{k}\times\Sigma^{(k)}\times Q\times\mathbb{S}$ は辺の集合である．辺は $\sigma(q_1,\ldots,q_k)\xrightarrow{w}q$ と書き，記号 $\sigma\in\Sigma^{(k)}$ を読み込んで，状態ベクトル $\{q_1,\ldots,q_k\}$ から状態 q へ重み $w\in\mathbb{S}$ で遷移する．i への特殊な遷移として，辺 $\varepsilon()\xrightarrow{\bar{1}}i$ を定義し，ε は空記号とする ($\varepsilon\notin\Sigma$).
- ρ は F から $\mathbb{S}$ への終了重み関数である．

オートマトンのサイズ $|A|$ は $|Q|+|E|$ と定義する．

木 $t(t\in T_\Sigma)$ に対する A の**実行**を写像 $r:Pos(t)\to Q$ として定義し，全ての位置 $pos\in Pos(t)$ に対して，$t(pos)$ が $\sigma^{(k)}$ となるとき，辺 $\sigma^{(k)}(r(pos.1),\ldots,r(pos.k))\xrightarrow{w}r(pos)\in E$ を満たす必要がある．t に対する実行の集合を $Run_A(t)$ と書く．もし，$r(root)\in F\wedge(\forall pos\in leaves(t), r(pos)=i)$ となる r が $Run_A(t)$ に存在するならば，t は A によって**受理可能**と呼ぶ．また，そのような実行 r を**受理可能な実行**と呼ぶ．A の言語 $L(A)=\{t\mid\forall t\in T_\Sigma, t$ は A によって受理可能 $\}$ として定義する．$L(A)=L(A')$ のとき，A は FTA A' と**等価**と呼ぶ．A が**無曖昧**であるとき，全ての $t\in L(A)$ に対して，$Run_A(t)$ には高々 1 つの受理可能な実行しか存在しない．

実行 $r\in Run_A(t)$ の重みは

$$w(r)=\Big(\bigotimes_{pos\in Pos(t)}w\Big)\tag{1}$$

として定義され，ここでは $\sigma^{(k)}(r(pos.1),\ldots,r(pos.k))\xrightarrow{w}r(pos)$ または $\varepsilon()\xrightarrow{w}r(pos)$ が E 中に存在する．また，受理可能な実行 $r\in Run_A(t)(t\in L(A))$ の重みは

$$w_A(r)=w(r)\otimes\rho(r(root)).\tag{2}$$

として定義する．複数の状態から構成されるベクトルの集合 $V\subseteq Q^*$，状態の集合 $U\subseteq Q$，木 $t\in T_\Sigma$ に対して，

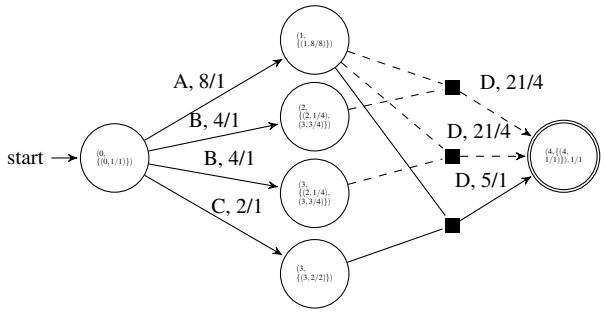

図 3: 図 1 のオートマトンに対する曖昧性解消の結果.

V 中のベクトルから U 中の状態へ木 t で遷移する実行の集合を $Run(V,t,U)$ として書き，その重みの合計を

$$W(V,t,U) = \bigoplus_{r \in Run(V,t,U)} w(r). \tag{3}$$

として書く．V が単集合 $\{\{p_1,\ldots,p_k\}\}$ のとき，単純に $W(\{p_1,\ldots,p_k\},t,U)$ として書き，U に対しても同様に書く．さらに，V 中のベクトルから始まり，木 $t \in T_\Sigma$ で到達できる状態の集合を $\delta(V,t)$ として書く．2 つの状態 p_j と q_j に対して $Run(\{p_1,\ldots,p_j,\ldots,p_k\},t,F) \cap Run(\{q_1,\ldots,q_j,\ldots,q_k\},t,F) \neq \emptyset\ (1 \le j \le k)$ となるような木 $t \in T_\Sigma$ が存在するとき，p_j と q_j は**将来を共有する**と呼ぶ．また，ある木 $t \in T_\Sigma$ に対して，状態 $p \in Q$ における**閉路** c は $root, pos \in Pos(t)$ が $c(root) = p$，$c(pos) = p$ で $j \in \mathbb{N}$, $pos.j \notin Pos(t)$ である実行のことを言う．状態 p における閉路の集合を $Cyc(p,t,p)$，その中に含まれる全ての閉路の重みの合計を $W(p,t,p)$ として書く．

図 1 に半環 $(R_+,+,\times,0,1)$ 上での重み付き FTA の例を示し，それは次のような要素から成る．

- $\Sigma = \{\mathrm{A}^{(1)},\mathrm{B}^{(1)},\mathrm{C}^{(1)},\mathrm{D}^{(2)}\}$,
- $Q = \{0,1,2,3,4\}$ and $i = 0$ and $F = \{4\}$,
- $E = \{\varepsilon() \xrightarrow{1/1} 0, \mathrm{A}(0) \xrightarrow{8/1} 1, \mathrm{B}(0) \xrightarrow{1/1} 2, \mathrm{C}(0) \xrightarrow{2/1} 3, \mathrm{B}(0) \xrightarrow{3/1} 3, \mathrm{D}(1,2) \xrightarrow{6/1} 4, \mathrm{D}(1,3) \xrightarrow{5/1} 4\}$,
- $\rho(4) = 1/1$.

全ての重みには分母を付けて書く．図 2 に示した木はこのオートマトンで受理可能であるが，左側の木は次の 2 つの実行によって受理できる．

$$r_1 = \{0,1,2,4\} \text{ and } r_2 = \{0,1,3,4\}.$$

これは図 1 が曖昧であることを意味する．実行の重み $w(r_1)$ は 48，$w(r_2)$ は 120 となる．

4 一般化曖昧性解消演算

木 $t \in T_\Sigma$ と状態 $p \in \delta(V,t)$ に対し，$\delta(V,t)$ 中で p と将来を共有する状態の集合を

$$\delta_p(V,t) = \delta(V,t) \cap \{q : (p,q) \in B\}$$

として定義する．B は $trim(A \cap A)$ で，A と A の積集合 [4] を計算し，そこから終了状態に到達できない状態と辺を除去したオートマトン ($trim$) である．(p,q) が B 中に存在するかをチェックすることで，p と q が将来を共有するか効率的に決定できる [17]．木 $t \in T_\Sigma$ と状態 $p \in \delta(\{i\},t)$ に対し，重み付き部分集合

$$s(p,t) = \Big\{(q_1,w_1),\ldots,(q_\ell,w_\ell) : \{q_1,\ldots,q_\ell\} = \delta_p(\{i\},t),$$
$$\forall j \in [1,\ell], w_j = W(\{i\},t,\{q_1,\ldots,q_\ell\})^{-1} \otimes W(\{i\},t,q_j)\Big\}.$$

として定義する．場合によって p と t は省略し，重み付き部分集合 $s(p,t)$ を単に s と書くときがある．重み付き部分集合 $s(p,t)$ に対して，$set(s) = \{q_1,\ldots,q_\ell\}$ を定義する．重み付き FTA A に対し，$A' = \{\Sigma, Q', i', F', E', \rho'\}$ を次のように定義する．

$$Q' = \Big\{(p, s(p,t)) : t \in T_\Sigma, p \in \delta(\{i\},t)\Big\},\ \ i' = (i, \{(i,\overline{1})\}),$$
$$F' = \Big\{(p, s(p,t)) : t \in L(A), p \in \delta(\{i\},t) \cap F\Big\},$$
$$E' = \Big\{\sigma^{(k)}((p_1,s_1),\ldots,(p_k,s_k)) \xrightarrow{w} (p,s) : (p_1,s_1),\ldots,$$
$$(p_k,s_k),(p,s) \in Q', \sigma^{(k)} \in \Sigma^{(k)}, \exists t_1,\ldots,t_k \in T_\Sigma \mid$$
$$s_1 = s(p_1,t_1) = \{(p_{1,1},w_{1,1}),\ldots,(p_{1,\ell_1},w_{1,\ell_1})\},\ldots,$$
$$s_k = s(p_k,t_k) = \{(p_{k,1},w_{k,1}),\ldots,(p_{k,\ell_k},w_{k,\ell_k})\},$$
$$s = s(p,\sigma^{(k)}(t_1,\ldots,t_k)) = \{(p'_1,w'_1),\ldots,(p'_\ell,w'_\ell)\},$$
$$p \in \delta(\{p_1,\ldots,p_k\},\sigma^{(k)}), w = \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} \{w_{1,j_1} \otimes \cdots \otimes$$
$$w_{k,j_k} \otimes W(\{p_{1,j_1},\ldots,p_{k,j_k}\},\sigma^{(k)},set(s))\},$$
$$\forall j \in [1,\ell], w'_j = w^{-1} \otimes \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} \{w_{1,j_1} \otimes \cdots \otimes w_{k,j_k} \otimes$$
$$W(\{p_{1,j_1},\ldots,p_{k,j_k}\},\sigma^{(k)},p'_j)\}\Big\},$$
$$\forall (p,s) \in F', s = \{(p_1,w_1),\ldots,(p_\ell,w_\ell)\}, \rho'((p,s)) =$$
$$\bigoplus_{j=1,p_j \in F}^{\ell} (w_j \otimes \rho(p_j)).$$

この定義から A' の受理経路の重みは，それが対応する木の A 上での全受理経路の重みの合計となる (ただし，半環の定義による．)．この詳細な分析は次節で与える．図 3 には図 1 の FTA に対する演算結果を示した．曖昧性を排除するため，点線の辺のうち一方は排除する必要があるが，これにはまず初期状態から同じ木で到達できる状態同士を同値関係として定義する．そして，

- ある状態へ同じ記号で遷移する 2 つの状態ベクトルが各次元の状態間で同値関係を持つとき，一方の遷移を排除する，
- 同値関係を持つ終了状態の一方を排除する，

ことで曖昧性が排除できる．この具体的なアルゴリズムの一例は文献 [17] で述べられている．

5 分析

ここでは，トロピカル半環 $(R_+^{+\infty}, \min, +, +\infty, 0)$ 上で定義された閉路を含む重み付き FTA に対して，一般化曖昧性解消演算が適用できる十分条件を示す．この条件を**弱双子性質** (weak twins property) と呼ぶ．この性質の定義や証明を与える前に，一般化曖昧性解消演算の以下の別の重要な性質について証明を与える．

命題 5.1 *FTA* $A = (\Sigma, Q, i, F, E, \rho)$ に対して，一般化曖昧性解消演算を適用することで $A' = (\Sigma, Q', i', F', E', \rho')$ が構築されたとする．このとき，どんな実行 $r \in Run(\{i'\}, t, (q,s))$，木 $t \in T_\Sigma$，状態 $(q,s) \in Q'$ $(s = \{(p_1, w_1), \ldots, (p_\ell, w_\ell)\})$ に対しても次の等式が成り立つ*:*

$$w(r) = W(\{i\}, t, set(s)) \text{ and } \forall n \in [1,\ell], w(r) \otimes w_n = W(\{i\}, t, p_n).$$

証明 木 t の高さに対する帰納法によって，この命題を証明する．$height(t) = 1$ の場合，$t = \varepsilon()$ であり，$Run(\{\}, t, (i, \{(i, \overline{1})\}))$ には高々 *1* つの実行 r が存在する．r は特殊な遷移 $\varepsilon() \xrightarrow{\overline{1}} (i, (i, \overline{1}))$ から構築されたもので，$w(r) = \overline{1}$，$W(\{\}, t, (i, \{(i, \overline{1})\})) = \overline{1}$ となる．また，$w(r) \otimes \overline{1} = \overline{1}$ が成り立つので，命題の等式は成り立つ．

$height(t) \leq n$ $(n \in \mathbb{N})$ のときに命題の等式が成り立つと仮定する．ある $\sigma^{(k)} \in \Sigma^{(k)}$ と高さ n 以下で，かつ，少なくとも高さ n の木を一本は含む $t_1, \ldots, t_k \in T_\Sigma$ に対して，$t = \sigma^k(t_1, \ldots, t_k)$ とする $(height(t) = n+1)$．また，実行 r が $Run(\{i'\}, t, (q,s))$ $(q \in Q)$ に含まれ，$\forall j \in [1,k]$ に対して，r_j が $Run(\{i'\}, t_j, (q_j, s_j))$ $(s_j = \{(q_{j,1}, w_{j,1}), \ldots, (q_{j,\ell_j}, w_{j,\ell_j})\})$ に含まれるとすると，E' には辺 $\sigma^{(k)}((q_1,s_1), \ldots, (q_k,s_k)) \xrightarrow{w} (q,s)$ が存在する．このとき，次の式*:*

$$\begin{aligned} w(r) &= \bigotimes_{j=1}^{k} w(r_j) \otimes w = \bigotimes_{j=1}^{k} w(r_j) \otimes \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} w_{1,j_1} \otimes \\ &\quad \cdots \otimes w_{k,j_k} \otimes W(\{q_{1,j_1}, \ldots, q_{k,j_k}\}, \sigma^{(k)}, set(s)) \\ &= \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} W(\{i\}, t_1, q_{1,j_1}) \otimes \cdots \otimes W(\{i\}, t_k, q_{k,j_k}) \otimes \\ &\quad W(\{q_{1,j_1}, \ldots, q_{k,j_k}\}, \sigma^{(k)}, set(s)) \end{aligned} \tag{4}$$

が書ける．ここで仮定から式 $\forall m \in [1,k], w(r_m) \otimes w_{m,j_m} = W(\{i\}, t_m, q_{m,j_m})$ を使った．

次に，A 上において木 $t = \sigma^{(k)}(t_1, \ldots, t_k)$ に対する $Run(\{i\}, t, set(s))$ 中のどんな実行 v も $set(s_1), \ldots, set(s_k)$ を経由することを証明する．まず，実行 v は k 個の実行に分割できる．$\forall j \in [1,k]$ に対して，その実行は初期状態 i から木 t_j で $set(s_j)$ 中の *1* つの状態へ到達する．さらに，それら k 個の状態で構成されたベクトルから記号 $\sigma^{(k)}$ で $set(s)$ のある状態へと遷移する辺が存在する．このとき，式 *(4)* は次のことを意味する*:*

$$w(r) = W(\{i\}, t, set(s)). \tag{5}$$

r を k 個の実行 $v_1, \ldots, v_k$ に分割し，$\forall j \in [1,k]$ に対して，v_j は初期状態 i から木 t_j で状態 q'_j $(\in Q)$ に到達する実行とする．また，それらの状態から構成された k 次元ベクトル $\{q'_1, \ldots, q'_k\}$ は記号 $\sigma^{(k)}$ による遷移を表す辺 e' によって，$set(s)$ 中の状態 q' へ到達できるとする．q' は $set(s)$ 中に存在するので，q' と q は将来を共有する．さらに，q' は $\delta(\{q'_1, \ldots, q'_k\}, \sigma^{(k)})$ に含まれ，q は $\delta(\{q_1, \ldots, q_k\}, \sigma^{(k)})$ に含まれるので，$\forall j \in [1,k]$ に対して，q'_j と q_j もまた将来を共有する．また，$\forall j \in [1,k]$ に対して，$q'_j \in \delta(\{i\}, t_j)$ となるので，これは q'_j が $set(s_j)$ に含まれることを意味し，実行 v_j が $set(s_j)$ で終了することを意味する．

これより $w(r) = \bigotimes_{j=1}^{k} w(r_j) \otimes w$ という観点に立つと，どんな $n \in [1,\ell]$ に対しても，次のように書ける*:*

$$\begin{aligned} w(r) \otimes w_n &= \bigotimes_{j=1}^{k} w(r_j) \otimes w \otimes w^{-1} \otimes \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} w_{1,j_1} \otimes \\ &\quad \cdots \otimes w_{k,j_k} \otimes W(\{q_{1,j_1}, \ldots, q_{k,j_k}\}, \sigma^{(k)}, p_n) \\ &= \bigoplus_{j_1=1}^{\ell_1} \cdots \bigoplus_{j_k=1}^{\ell_k} W(\{i\}, t_1, q_{1,j_1}) \otimes W(\{i\}, t_k, q_{k,j_k}) \otimes \\ &\quad W(\{q_{1,j_1}, \ldots, q_{k,j_k}\}, \sigma^{(k)}, p_n). \end{aligned} \tag{6}$$

ここで仮定から式 $\forall m \in [1,k], w(r_m) \otimes w_{m,j_m} = W(\{i\}, t_m, q_{m,j_m})$ を使った．

初期状態 i から始まり，木 $t = \sigma^{(k)}(t_1, \ldots, t_k)$ で p_n へ到達して終了する実行 v は，$\forall j \in [1,k]$ に対して，木 t_j で $set(s_j)$ のある状態へ到達するとき，そのような k 個の状態から構成されるベクトルを必ず経由する．この観点に立つと，式 *(6)* は次のことを意味する*:*

$$w(r) \otimes w_n = W(\{i\}, t, p_n). \tag{7}$$

これより命題が成り立つことは明らかである． □

FTA に対する弱双子性質を定義し，トロピカル半環上で定義された閉路を含む重み付きの FTA がこの性質を満たすとき，一般化曖昧性解消演算が適用可能であることを示す．

定義 5.1 *2* つの状態 $p, q \in Q$ が**兄弟** *(sibling)* であるとは，$Run(\{i\}, t_1, p) \neq \emptyset$ かつ $Run(\{i\}, t_1, q) \neq \emptyset$，及び，$Cyc(p, t_2, p) \neq \emptyset$ かつ $Cyc(q, t_2, q)$ となるような木 $t_1, t_2 \in T_\Sigma$ が存在するときをいう．兄弟 p, q が**双子** *(twins)* であるとは，そのような t_2 に対して，$W(p, t_2, p) = W(q, t_2, q)$ となるときをいう．*FTA A* が**双子性質** *(twins property)* を満たすとは，どんな兄弟も双子であるときをいう *[1]*．また，A が**弱双子性質**を満たすとは，将来を共有するどんな兄弟も双子であるときをいう．

定理 5.1 トロピカル半環上で定義された *FTA A* が弱双子性質を満たすとする．このとき，一般化曖昧性解消演算は A に適用可能である．

証明 A が弱双子性質を満たし，一般化曖昧性解消演算が無限に多くの異なる状態 (q,s) を作り出すと仮定する．この仮定は一般化曖昧性解消演算が，$n \in \mathbb{N}$ に対して，無限に多くの (q,s_n) を作り出すことを意味する．ここで $\ell < +\infty$ に対し，$set(s_n) = \{p_1,\ldots,p_\ell\}$ は全て同一で，$s_n = \{(p_1,w_n(p_1)),\ldots,(p_\ell,w_n(p_\ell))\}$ と書ける．

どんな $n \in \mathbb{N}$ にも，$\forall p \in \{p_1,\ldots,p_\ell\}$ に対して，木 $t_n \in T_\Sigma$ が存在し，トロピカル半環上での命題 *5.1* から，

$$w_n(p) = W(\{i\},t_n,p) - W(\{i\},t_n,\{p_1,\ldots,p_\ell\}). \quad (8)$$

と書ける．ここで，ℓ は有限なので，無限に多くのインデックス集合 $J \subseteq \mathbb{N}$ に対し，$W(\{i\},t_n,\{p_1,\ldots,p_\ell\}) = W(\{i\},t_n,p)$ となるような状態 $p \in \{p_1,\ldots,p_\ell\}$ が少なくとも *1* つは存在する．そして，式 *(8)* から，$\forall n \in J$ に対し，$w_n(p) = 0$ が成り立つ．$\forall i \in [1,\ell]$ に対して，集合 $\{w_n(q) - w_n(p_i) : n \in J\}$ は有限になることはできない．さもなければ，$\{w_n(q) - w_n(p) : n \in J\} = \{w_n(q) : n \in J\}$ は有限となり，これは，全ての i に対して $\{w_n(q) - w_n(p_i) : n \in J\}$ が有限となることから，全ての i に対して $\{w_n(p_i) : n \in J\}$ が有限であることを意味する．これは $\{s_n : n \in J\}$ が無限であることに違反する．よって，$\{w_n(q) - w_n(u) : n \in J\}$ を無限にする状態 $u \in \{p_1,\ldots,p_\ell\}$ が少なくとも *1* つは存在しなければならない．

そこで，集合 $\{w_n(q) - w_n(u) : n \in J\}$ が有限集合*:*

$$C = \big\{w(r_1) - w(r_2) : r_1 \in Run(\{i\},t,q), r_2 \in Run(\{i\},t,u), \\ height(t) \le |Q|^2\big\}.$$

に含まれることを証明し，これが仮定に違反することを示す．ある $n \in \mathbb{N}$ に対して，$t = t_n$ を考え，r_1 を $Run(\{i\},t,q)$ 中の重みが最短の実行 *(*最短実行*)*，r_2 を $Run(\{i\},t,u)$ 中の最短実行とする．式 *(8)* より，$w_n = w_n(q) - w_n(u)$ は次のように書ける．

$$\begin{aligned} w_n &= (w(r_1) - W(\{i\},t,p)) - (w(r_2) - W(\{i\},t,p)) \\ &= w(r_1) - w(r_2). \end{aligned} \quad (9)$$

q と u は共に初期状態 i から t で到達可能なので，$A \cap A$ には i' から (q,u) への実行が存在する．ここで $height(t) > |Q|^2$ と仮定すると，この実行はある状態 (q_1,u_1) において少なくとも *1* つの閉路を含むことになり，$pos_1 < pos_2$ かつ $r_1(pos_1) = r_1(pos_2) = q_1$ かつ $r_2(pos_1) = r_2(pos_2) = u_1$ となるような位置 $pos_1, pos_2 \in Pos(t)$ を見つけることができる．r_1 と r_2 は最短実行なので，閉路 $c_1 \in Cyc(q_1, t|_{pos_1}, q_1)$ と $c_2 \in Cyc(u_1, t|_{pos_1}, u_1)$ もまた最短実行となる．さらに，実行 $r'_1 \in Run(\{i\}, t|_{pos_2}, q_1)$ と $r'_2 \in Run(\{i\}, t|_{pos_2}, u_1)$ もまた最短実行となる．一般化曖昧性解消演算によって作られる状態の定義から，$\{p_1,\ldots,p_\ell\}$ 中の全ての状態，特に u は q と将来を共有する．これは明らかに q_1 と u_1 もまた将来を共有していることを意味する．それゆえ，弱双子性質より，$w(c_1) = w(c_2)$ とならなければならない．よって，$r''_1 \in Run(\{i\},t',q)$ と $r''_2 \in Run(\{i\},t',u)$ に対して，$w_n = w(r''_1) - w(r''_2)$ となるような木 $t' = t[t|_{pos_2}]_{pos_1}$ を作ることができる．$height(t') < height(t)$ となるので，$height(t)$ に対する帰納法より，$r'''_1 \in Run(\{i\},t'',q)$ と $r'''_2 \in Run(\{i\},t'',u)$ に対して，$height(t'') \le |Q|^2$ で $w_n = w(r'''_1) - w(r'''_2)$ となる木 t'' を作ることができる．全ての t_n に対して，この過程を適用することで，$\{w_n : n \in J\}$ が有限集合 C に含まれることが示せる．これは一般化曖昧性解消演算が無限に多くの状態を作り出すことに違反するので，仮定に反する．□

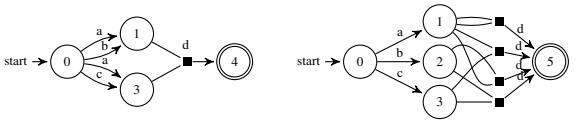

図 4: 無曖昧な木オートマトンとその決定化の結果.

6 実験

文圧縮と機械翻訳タスクで実験を行う．これらは (拡張的，線形，削除無し) の重み付き木トランスデューサ [12] でモデル化し，日英機械翻訳に対するその規則の例は次のようなものである．

q.S(N(kare wa) VP(x1:N x2:V)) $\xrightarrow{0.8}$ S(N(He) VP(q.x2 q.x1))
q.V(suki desu) $\xrightarrow{0.3}$ V(likes), q.N(ryouri ga) $\xrightarrow{0.5}$ N(cooking).

ここで q は状態名，x#は変数を表す．規則の左辺を木 S(N(kare wa) VP(N(ryouri ga) V(suki desu))) へ適用すると，上のトランスデューサは木オートマトン: $\Sigma = \{\mathrm{S}^{(2)}, \mathrm{N}^{(1)}, \mathrm{VP}^{(2)}, \mathrm{V}^{(1)}, \mathrm{He}^{(1)}, \mathrm{likes}^{(1)}, \mathrm{cooking}^{(1)}\}$, $Q = \{\mathrm{q0}, \mathrm{q1}, \mathrm{q2}, \mathrm{q3}, \mathrm{q4}, \mathrm{q5}, \mathrm{q6}, \mathrm{q7}, \mathrm{q8}\}$, $i = \mathrm{q4}$, $F = \{\mathrm{q0}\}$, $E = \{\varepsilon() \xrightarrow{1.0} \mathrm{q4}, \mathrm{S(q1\ q2)} \xrightarrow{0.8} \mathrm{q0}, \mathrm{N(q3)} \xrightarrow{1.0} \mathrm{q1}, \mathrm{He(q4)} \xrightarrow{1.0} \mathrm{q3}, \mathrm{VP(q6\ q5)} \xrightarrow{1.0} \mathrm{q2}, \mathrm{N(q7)} \xrightarrow{0.5} \mathrm{q5}, \mathrm{cooking(q4)} \xrightarrow{1.0} \mathrm{q7}, \mathrm{V(q8)} \xrightarrow{0.3} \mathrm{q6}, \mathrm{likes(q4)} \xrightarrow{1.0} \mathrm{q8}\}$, $\rho(\mathrm{q0}) = 1.0$ を生成する．これらの規則で左辺の高さが 1 より高い場合，標準化 [4] を行った．例えば，規則 $\mathrm{V(likes)(q4)} \xrightarrow{0.3} \mathrm{q6}$ は 2 つの規則 $\mathrm{V(q8)} \xrightarrow{0.3} \mathrm{q6}$ と $\mathrm{likes(q4)} \xrightarrow{1.0} \mathrm{q8}$ に分割することで平坦化される．実験ではさらに ε 遷移を取り除いた (初期状態への特殊な遷移は除く)．構文解析済みの対訳文，及び，その単語アライメントのデータから harvest2[1] を使って，木トランスデューサの規則を取り出した．単語アライメントには GIZA++[2]を使い，英語と日本語文の構文解析には内製システムを使った．さらに，Tiburon[3] を木トランスデューサの規則の重み学習，及び，オートマトンの決定化に使用した．

6.1 文圧縮

文圧縮コーパス (480 文対) [2] を使い，木トランスデューサの規則抽出と学習を行った．このデータは規模が小さくデータスパースネスの問題が大きいため，再

[1] http://staffwww.dcs.shef.ac.uk/people/T.Cohn/t3/
[2] http://code.google.com/p/giza-pp/
[3] http://www.isi.edu/licensed-sw/tiburon/

表 1: 文圧縮タスクにおける各オートマトンの平均状態数，遷移数，実行数．

	baseline	determinize	disambiguate
# of states	135.49	68.39	93.19
# of transitions	167.73	190.99	153.62
# of runs	7.95×10^{15}	5.15×10^{11}	5.15×10^{11}

表 2: 機械翻訳タスクにおける各オートマトンの平均状態数，遷移数，実行数．

	baseline	determinize	disambiguate
# of states	3591.55	2469.33	3552.00
# of transitions	5046.88	1031625.88	15225.55
# of runs	4.01×10^{35}	2.47×10^{35}	2.47×10^{35}

度訓練データ 480 文を使って，テストを行い，そのうち 466 文だけ木オートマトンに変換できた．表 1 には 466 個の元のオートマトンとそれらへ決定化，及び，提案法を適用したときの結果の平均状態数，遷移数，実行数を示した．このデータでは，提案法の方が決定化よりも若干小さなオートマトンを構築できていることがわかる．元のオートマトンでは平均 7.95×10^{15} の実行が存在したが，決定化，曖昧性解消共に平均 5.15×10^{11} の実行数となり，約 99%の冗長な解を排除できている．

6.2 機械翻訳

特許翻訳タスク NTCIR-10 の日英データを使って実験を行う [7]. その訓練データから 5,000 文対をランダムに抽出し，木トランスデューサの規則抽出と学習を行った．テストにはテストセットのうち 20 文長以下の日本語 278 文を使った．そのうち，76 文だけオートマトンへと変換でき，さらにそのうち 9 文だけが決定化と提案法が現実的な時間で停止した．表 2 からは提案法が決定化よりも約 90 倍小さなオートマトンを構築できていることがわかる. 図 4 の左図に簡単な木オートマトンを示したが，これは無曖昧であり，提案法は入力と同じオートマトンを出力する．図 4 の右図に示すように，決定化では遷移数が増幅されてしまう．文圧縮データとは異なり，機械翻訳データではこのようなタイプのオートマトンが多く含まれる．なぜなら，入力単語に対して多様な訳語が存在し，初期状態からの遷移で同一単語による遷移が多く出来上がるためである．

7 むすび

本稿では曖昧性解消演算 [14] を重み付き FTA に適用可能なように一般化した．また，重み付き FTA が閉路を含む場合に適用可能な十分条件を示すことで，決定化 [1] よりも広いクラスの FTA が扱えることを示した．実験では，自然言語処理タスク，特に分野で最重要タスクとなっている機械翻訳においてその有効性を示した．今後は，曖昧性解消演算を木トランスデューサの合成演算や前向き適用演算 [12] と併せて動的に処理する手法の開発が実用では重要な課題となる．

参考文献

[1] Matthias Büchse, Jonathan May, and Heiko Vogler. Determinization of weighted tree automata using factorizations. *Journal of Automata, Languages and Combinatorics*, 15(3/4):229–254, 2009.

[2] Trevor Cohn and Mirella Lapata. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144. Association for Computational Linguistics, 2008.

[3] Trevor Anthony Cohn and Mirella Lapata. Sentence compression as tree transduction. *Journal of Articial Intelligence Research*, 34:637–674, 2009.

[4] H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. Tree automata techniques and applications. Available on: `http://www.grappa.univ-lille3.fr/tata`, 2007. release October, 12th 2007.

[5] Frank Drewes. *Grammatical Picture Generation: A Tree-Based Approach (Texts in Theoretical Computer Science. An EATCS Series)*. Springer-Verlag New York, Inc., 2006.

[6] Clarence A Ellis. Probabilistic tree automata. *Information and control*, 19(5):401–416, 1971.

[7] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR conference*, pages 260–286, 2013.

[8] Jonathan Graehl, Kevin Knight, and Jonathan May. Training tree transducers. *Computational Linguistics*, 34(3):391–427, September 2008.

[9] Liang Huang and David Chiang. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 53–64. Association for Computational Linguistics, 2005.

[10] Donald E Knuth. A generalization of dijkstra's algorithm. *Information Processing Letters*, 6(1):1–5, 1977.

[11] Jonathan May and Kevin Knight. A better n-best list: Practical determinization of weighted finite tree automata. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 351–358. Association for Computational Linguistics, 2006.

[12] Jonathan David Louis May. *Weighted Tree Automata and Transducers for Syntactic Natural Language Processing*. PhD thesis, University of Southern California, 2010.

[13] Mehryar Mohri. A disambiguation algorithm for finite automata and functional transducers. In *Implementation and Application of Automata*, pages 265–277. Springer, 2012.

[14] Mehryar Mohri and Michael D. Riley. On the disambiguation of weighted automata. *CoRR*, abs/1405.0500, 2014.

[15] Erik Meineche Schmidt. Succinctness of descriptions of context-free, regular, and finite languages. *DAIMI Report Series*, 7(84), 1978.

[16] James W. Thatcher and Jesse B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Mathematical systems theory*, 2(1):57–81, 1968.

[17] 林克彦 永田昌明. 有限オートマトンに対する一般化曖昧性解消演算の正当性. In **第 *95* 回人工知能学会基本問題研究会**, 2014.