Construction of a Bottom-Up Deterministic n-Gram Weighted Tree Automaton

Matthias Büchse, Tobias Denkinger, and Heiko Vogler

Department of Computer Science
Technische Universität Dresden

Abstract. We propose a novel construction to lift an n-gram model to trees. The resulting weighted tree automaton is bottom-up deterministic in contrast to the weighted tree automaton constructed using the Bar-Hillel, Perles, Shamir algorithm.

1 Introduction

Recent approaches to machine translation are mostly statistical [4]. Researchers define a class of translation functions, and they use training algorithms to select a function that fits a given set of existing translations. Translation functions that are considered in research are often syntax-directed, i.e., the grammatical structure of a sentence, represented by a tree, is of special interest.

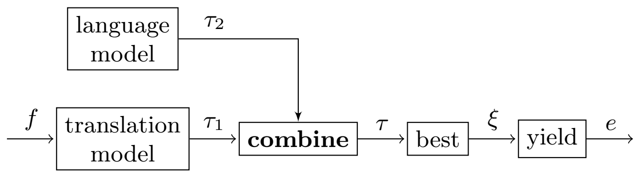

Fig. 1. Translation function with translation and language model.

A typical translation function is shown in Fig. 1. The translation model consumes the input sentence f and emits a weighted tree language (WTL) τ_1 over $(\mathbb{R}_{\geq 0}, +, \cdot, 0, 1)$, in which each translation of f is assigned a real number (as weight). The language model provides a weight for every sentence of the target language by means of a WTL τ_2. Both WTLs are then combined into one WTL τ. Best followed by yield outputs the string e of the best tree ξ in τ.

Extended top-down tree transducers [3], synchronous tree-adjoining grammars [6], and synchronous context-free grammars [2] are some of the most prominent examples of translation models. Examples of language models are n-gram models, hidden Markov models, weighted string automata (WSA), and probabilistic context-free grammars.

All language models mentioned above generate weighted string languages (WSL). But in order to make the combination of τ_1 and τ_2 possible, τ_2 must be lifted to a WTL. In this paper we show a construction that lifts the n-gram model to a WTL by constructing a weighted tree automaton (WTA), called n-gram WTA.

The classical approach to construct the n-gram WTA is the following: for the n-gram model N, we can construct a WSA $\mathcal{A}$ that recognizes N. Then we construct the product of $\mathcal{A}$ and the WTA that recognizes every tree with weight 1. For this, we employ the extension [5, Section 4] of the Bar-Hillel, Perles, Shamir algorithm [1, Lemma 4.1]. The constructed product is the n-gram WTA.

We propose a direct construction for the n-gram WTA. We show that the resulting WTA is bottom-up deterministic, which is in contrast to the n-gram WTA produced by the classical approach. Our construction is inspired by [2] where it appears interleaved with the other steps shown in Fig. 1.

An efficient implementation of the translation function in Fig. 1 computes the two functions best and combine interleaved, where best is usually computed via dynamic programming, i.e., bottom-up. Thus such an algorithm can profit when τ_2 is specified in a bottom-up deterministic manner.

2 Preliminaries

We let Γ^* denote the set of all words over an alphabet Γ. For $w \in \Gamma^*$ and $k \in \mathbb{N}$, $\mathrm{fst}_k(w)$ and $\mathrm{lst}_k(w)$ denote the sequences of the first and the last k symbols of w, respectively. A *ranked alphabet* is a tuple (Σ, rk) where Σ is an alphabet and $\mathrm{rk}\colon \Sigma \to \mathbb{N}$ is a *rank mapping*. In the following, we assume that Γ is an alphabet and (Σ, rk), or just Σ, is a ranked alphabet with $\Gamma \subseteq \mathrm{rk}^{-1}(0)$.

Let Q be an alphabet, the set of *unranked trees over* Q is denoted by $\mathcal{U}_Q$. The set of *positions of* ξ is denoted by $\mathrm{pos}(\xi)$. For $p \in \mathrm{pos}(\xi)$, the *symbol of* ξ *at* p is denoted by $\xi(p)$. The set of *(ranked) trees over* Σ is denoted by T_Σ. The Γ*-yield of* ξ is the mapping $\mathrm{yield}_\Gamma : T_\Sigma \to \Gamma^*$ where $\mathrm{yield}_\Gamma(\xi)$ is the sequence of all symbols σ in ξ with $\sigma \in \Gamma$ read from left to right.

A *weighted tree automaton (WTA)* is a tuple $\mathcal{A} = (Q, \Sigma, \delta, \nu)$ where Q is an alphabet, δ is a Σ-family of functions $\delta_\sigma : Q^{\mathrm{rk}(\sigma)} \times Q \to \mathbb{R}_{\geq 0}$, and $\nu\colon Q \to \mathbb{R}_{\geq 0}$. The set of all *runs of* $\mathcal{A}$ *on* ξ is the set $R_\mathcal{A}(\xi) = \{\kappa \in \mathcal{U}_Q \mid \mathrm{pos}(\kappa) = \mathrm{pos}(\xi)\}$. For $\kappa \in R_\mathcal{A}(\xi)$, the *weight of* κ is $\mathrm{wt}(\kappa) = \prod_{p \in \mathrm{pos}(\xi)} \delta_{\xi(p)}(\kappa(p1), \ldots, \kappa(p\,\mathrm{rk}(\xi(p))), \kappa(p))$. The *semantics of* $\mathcal{A}$ is the mapping $[\![\mathcal{A}]\!]\colon T_\Sigma \to \mathbb{R}_{\geq 0}$ where for every $\xi = \sigma(\xi_1, \ldots, \xi_{\mathrm{rk}(\sigma)}) \in T_\Sigma$ we define $[\![\mathcal{A}]\!](\xi) = \sum_{\kappa \in R_\mathcal{A}(\xi)} \mathrm{wt}(\kappa) \cdot \nu(\kappa(\varepsilon))$. We call $\mathcal{A}$ *bottom-up deterministic* if for every $\sigma \in \Sigma$ and $q_1, \ldots, q_{\mathrm{rk}(\sigma)} \in Q$ there exists at most one $q \in Q$ such that $\delta_\sigma(q_1, \ldots, q_{\mathrm{rk}(\sigma)}, q) > 0$.

In the following we assume that $n \geq 1$. An *n-gram model over* Γ is a tuple $N = (\Gamma, \mu)$ where $\mu\colon \Gamma^n \to \mathbb{R}_{\geq 0}$ is a mapping (n-gram weights). The *semantics of an n-gram model* N is the mapping $[\![N]\!]\colon \Gamma^* \to \mathbb{R}_{\geq 0}$ where for every $l \geq 0$ and $w_1, \ldots, w_l \in \Gamma$ we define $[\![N]\!](w_1 \cdots w_l) = \prod_{i=0}^{l-n} \mu(w_{i+1} \cdots w_{i+n})$ if $l \geq n$, and $[\![N]\!](w_1 \cdots w_l) = 0$ otherwise. In the following, N denotes an n-gram model.

Proposition 1. *Let $u, v \in \Gamma^*$, $|u| \geq n$, and $|v| \geq n$. We have*

$$[\![N]\!](uv) = [\![N]\!](u) \cdot [\![N]\!](\mathrm{lst}_{n-1}(u)\,\mathrm{fst}_{n-1}(v)) \cdot [\![N]\!](v) \ .$$

3 Direct Construction

In order to define a WTA $\mathcal{A}_{N,\Sigma}$ with $[\![\mathcal{A}_{N,\Sigma}]\!] = [\![N]\!] \circ \text{yield}_\Gamma$ we have to compute $[\![N]\!] \circ \text{yield}_\Gamma(\xi)$ while traversing a given tree ξ bottom-up. At each node in ξ, we only see the current symbol and the states of the computations in the subtrees. A closer look at Proposition 1 suggests to (1) compute the semantics of the currently visible substrings under N and (2) propagate the left and right $n-1$ symbols of the substring in the state. In the following construction, parts (1) and (2) are handled by the functions g and f, respectively.

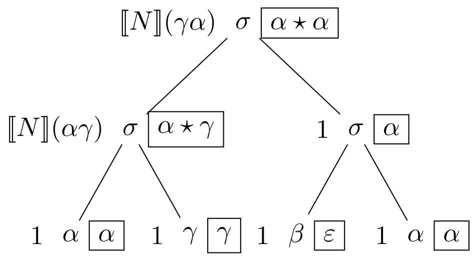

Fig. 2. Tree over $\Sigma = \{\sigma^{(2)}, \beta^{(0)}\} \cup \Gamma$ and $\Gamma = \{\alpha^{(0)}, \gamma^{(0)}\}$, with a run (states appear in boxes) and transition weights due to a 2-gram model N over Γ.

Let $\star$ be a new symbol, i.e., $\star \notin \Sigma$. We define $f\colon (\Gamma \cup \{\star\})^* \to (\Gamma \cup \{\star\})^*$ and $g\colon (\Gamma \cup \{\star\})^* \to \mathbb{R}_{\geq 0}$ as follows. Let $w \in (\Gamma \cup \{\star\})^*$. Then $f(w) = \text{fst}_{n-1}(w) \star \text{lst}_{n-1}(w)$ if $|w| \geq n$, and $f(w) = w$ otherwise. Note that there are $u_0, \ldots, u_k \in \Gamma^*$ such that $w = u_0 \star u_1 \cdots \star u_k$. We define $g(w) = \prod_{i=0}^{k} N'(u_i)$ where $N'(u_i) = [\![N]\!](u_i)$ if $|u_i| \geq n$, and $N'(u_i) = 1$ otherwise.

The *n-gram WTA over Σ* is the WTA $A_{N,\Sigma} = (Q, \Sigma, \delta, \nu)$ where $Q = Q_1 \cup Q_2$ with $Q_1 = \bigcup_{i=0}^{n-1} \Gamma^i$ and $Q_2 = \Gamma^{n-1} \times \{\star\} \times \Gamma^{n-1}$, $\nu(q) = 1$ if $q \in Q_2$, otherwise $\nu(q) = 0$, and for every $k \in \mathbb{N}$, $\sigma \in \Sigma^{(k)}$, and $q_1, \ldots, q_k, q \in Q$ (cf. Fig. 2 for an example):

$$\delta_\sigma(q_1, \ldots, q_k, q) = \begin{cases} g(q) & \text{if } k = 0 \text{ and } q = \text{yield}_\Gamma(\sigma) \\ g(q_1 \cdots q_k) & \text{if } k \geq 1 \text{ and } q = f(q_1 \cdots q_k) \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 1. *Let N be an n-gram model over Γ and Σ be a ranked alphabet. Then $[\![\mathcal{A}_{N,\Sigma}]\!] = [\![N]\!] \circ \text{yield}_\Gamma$ and the WTA $\mathcal{A}_{N,\Sigma}$ is bottom-up deterministic.*

References

1. Y. Bar–Hillel, M. Perles, and E. Shamir. On formal properties of simple phrase structure grammars. *Z. Phonetik. Sprach. Komm.*, 14:143–172, 1961.
2. D. Chiang. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228, June 2007.
3. K. Knight and J. Graehl. An overview of probabilistic tree transducers for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *LNCS*, pages 1–24. Springer, 2005.
4. A. Lopez. Statistical machine translation. *ACM Comp. Surv.*, pages 8:1–8:49, 2008.
5. A. Maletti and G. Satta. Parsing algorithms based on tree automata. In *Proc. of IWPT '09*, pages 1–12. ACL, 2009.
6. S. M. Shieber and Y. Schabes. Synchronous tree-adjoining grammars. In *Proceedings of the 13th Int. Conf. on Comp. Ling.*, volume 3, pages 253–258, 1990.