Lexicographic Semirings for Exact Automata Encoding of Sequence Models

Brian Roark, Richard Sproat, and Izhak Shafran
{roark,rws,zak}@cslu.ogi.edu

Abstract

In this paper we introduce a novel use of the lexicographic semiring and motivate its use for speech and language processing tasks. We prove that the semiring allows for exact encoding of backoff models with epsilon transitions. This allows for off-line optimization of exact models represented as large weighted finite-state transducers in contrast to implicit (on-line) failure transition representations. We present preliminary empirical results demonstrating that, even in simple intersection scenarios amenable to the use of failure transitions, the use of the more powerful lexicographic semiring is competitive in terms of time of intersection.

1 Introduction and Motivation

Representing smoothed n-gram language models as weighted finite-state transducers (WFST) is most naturally done with a failure transition, which reflects the semantics of the "otherwise" formulation of smoothing (Allauzen et al., 2003). For example, the typical backoff formulation of the probability of a word w given a history h is as follows

$$\mathrm{P}(w \mid h) \quad = \quad \begin{cases} \overline{\mathrm{P}}(w \mid h) & \text{if } c(hw) > 0 \\ \alpha_h \mathrm{P}(w \mid h') & \text{otherwise} \end{cases} \qquad (1)$$

where $\overline{\mathrm{P}}$ is an empirical estimate of the probability that reserves small finite probability for unseen n-grams; α_h is a backoff weight that ensures normalization; and h' is a backoff history typically achieved by excising the earliest word in the history h. The principle benefit of encoding the WFST in this way is that it only requires explicitly storing n-gram transitions for observed n-grams, i.e., count greater than zero, as opposed to all possible n-grams of the given order which would be infeasible in for example large vocabulary speech recognition. This is a massive space savings, and such an approach is also used for non-probabilistic stochastic language models, such as those trained with the perceptron algorithm (Roark et al., 2007), as the means to access all and exactly those features that should fire for a particular sequence in a deterministic automaton. Similar issues hold for other finite-state sequence processing problems, e.g., tagging, bracketing or segmenting.

Failure transitions, however, are an implicit method for representing a much larger explicit automaton – in the case of n-gram models, all possible n-grams for that order. During composition with the model, the failure transition must be interpreted on the fly, keeping track of those symbols that have already been found leaving the original state, and only allowing failure transition traversal for symbols that have not been found (the semantics of "otherwise"). This compact implicit representation cannot generally be preserved when composing with other models, e.g., when combining a language model with a pronunciation lexicon as in widely-used FST approaches to speech recognition (Mohri et al., 2002). Moving from implicit to explicit representation when performing such a composition leads to an explosion in the size of the resulting transducer, frequently making the approach intractable. In practice, an off-line approximation to the model is made, typically by treating the failure transitions as epsilon transitions (Mohri et al., 2002; Allauzen et al., 2003), allowing large transducers to be composed and optimized off-line. These complex approximate transducers are then used during first-pass decoding, and the resulting pruned search graphs (e.g., word lattices) can be rescored with exact language models encoded with failure transitions.

Similar problems arise when building, say, POS-taggers as WFST: not every pos-tag sequence will have been observed during training, hence failure transitions will achieve great savings in the size of models. Yet discriminative models may include complex features that combine both input stream (word) and output stream (tag) sequences in a single feature, yielding complicated transducer topologies for which effective use of failure transitions may not

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 1–5,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

be possible. An exact encoding using other mechanisms is required in such cases to allow for off-line representation and optimization.

In this paper, we introduce a novel use of a semiring – the lexicographic semiring (Golan, 1999) – which permits an exact encoding of these sorts of models with the same compact topology as with failure transitions, but using epsilon transitions. Unlike the standard epsilon approximation, this semiring allows for an exact representation, while also allowing (unlike failure transition approaches) for off-line composition with other transducers, with all the optimizations that such representations provide.

In the next section, we introduce the semiring, followed by a proof that its use yields exact representations. We then conclude with a brief evaluation of the cost of intersection relative to failure transitions in comparable situations.

2 The Lexicographic Semiring

Weighted automata are automata in which the transitions carry weight elements of a *semiring* (Kuich and Salomaa, 1986). A semiring is a ring that may lack negation, with two associative operations $\oplus$ and $\otimes$ and their respective identity elements $\overline{0}$ and $\overline{1}$. A common semiring in speech and language processing, and one that we will be using in this paper, is the *tropical semiring* $(\mathbb{R} \cup \{\infty\}, \min, +, \infty, 0)$, i.e., $\min$ is the $\oplus$ of the semiring (with identity ∞) and $+$ is the $\otimes$ of the semiring (with identity 0). This is appropriate for performing Viterbi search using negative log probabilities – we add negative logs along a path and take the min between paths.

A $\langle W_1, W_2 \ldots W_n \rangle$-lexicographic weight is a tuple of weights where each of the weight classes $W_1, W_2 \ldots W_n$, must observe the *path property* (Mohri, 2002). The path property of a semiring K is defined in terms of the *natural order* on K such that: $a <_K b$ iff $a \oplus b = a$. The tropical semiring mentioned above is a common example of a semiring that observes the path property, since:

$$\begin{aligned} w_1 \oplus w_2 &= \min\{w_1, w_2\} \\ w_1 \otimes w_2 &= w_1 + w_2 \end{aligned}$$

The discussion in this paper will be restricted to lexicographic weights consisting of a pair of tropical weights — henceforth the $\langle T, T \rangle$-lexicographic semiring. For this semiring the operations $\oplus$ and $\otimes$ are defined as follows (Golan, 1999, pp. 223–224):

$$\begin{aligned} \langle w_1, w_2 \rangle \oplus \langle w_3, w_4 \rangle &= \begin{cases} \langle w_1, w_2 \rangle & \begin{array}{r} \text{if } w_1 < w_3 \text{ or} \\ (w_1 = w_3 \;\& \\ w_2 < w_4) \end{array} \\ \langle w_3, w_4 \rangle & \text{otherwise} \end{cases} \\ \langle w_1, w_2 \rangle \otimes \langle w_3, w_4 \rangle &= \langle w_1 + w_3, w_2 + w_4 \rangle \end{aligned}$$

The term "lexicographic" is an apt term for this semiring since the comparison for $\oplus$ is like the lexicographic comparison of strings, comparing the first elements, then the second, and so forth.

3 Language model encoding

3.1 Standard encoding

For language model encoding, we will differentiate between two classes of transitions: backoff arcs (labeled with a ϕ for failure, or with ϵ using our new semiring); and n-gram arcs (everything else, labeled with the word whose probability is assigned). Each state in the automaton represents an n-gram history string h and each n-gram arc is weighted with the (negative log) conditional probability of the word w labeling the arc given the history h. For a given history h and n-gram arc labeled with a word w, the destination of the arc is the state associated with the longest suffix of the string hw that is a history in the model. This will depend on the Markov order of the n-gram model. For example, consider the trigram model schematic shown in Figure 1, in which only history sequences of length 2 are kept in the model. Thus, from history $h_i = w_{i-2}w_{i-1}$, the word w_i transitions to $h_{i+1} = w_{i-1}w_i$, which is the longest suffix of $h_i w_i$ in the model.

As detailed in the "otherwise" semantics of equation 1, backoff arcs transition from state h to a state h', typically the suffix of h of length $|h| - 1$, with weight $(-\log \alpha_h)$. We call the destination state a backoff state. This recursive backoff topology terminates at the unigram state, i.e., $h = \epsilon$, no history.

Backoff states of order k may be traversed either via ϕ-arcs from the higher order n-gram of order $k+1$ or via an n-gram arc from a lower order n-gram of order $k-1$. This means that no n-gram arc can enter the zeroeth order state (final backoff), and full-order states — history strings of length $n-1$ for a model of order n — may have n-gram arcs entering from other full-order states as well as from backoff states of history size $n-2$.

3.2 Encoding with lexicographic semiring

For an LM machine M on the tropical semiring with failure transitions, which is deterministic and has the

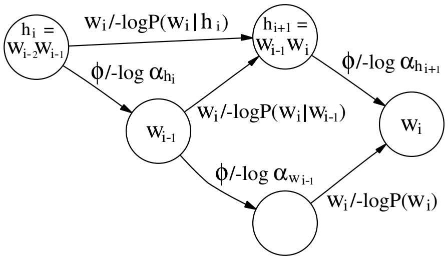
$h_i = w_{i-2}w_{i-1}$
w_i/-logP(w_i | h_i)
$h_{i+1} = w_{i-1}w_i$
ϕ/-log α_{h_i}
ϕ/-log $\alpha_{h_{i+1}}$
w_{i-1}
w_i/-logP(w_i|w_{i-1})
w_i
ϕ/-log $\alpha_{w_{i-1}}$
w_i/-logP(w_i)

Figure 1: Deterministic finite-state representation of n-gram models with negative log probabilities (tropical semiring). The symbol ϕ labels backoff transitions. Modified from Roark and Sproat (2007), Figure 6.1.

path property, we can simulate ϕ-arcs in a standard LM topology by a topologically equivalent machine M' on the lexicographic $\langle T, T \rangle$ semiring, where ϕ has been replaced with epsilon, as follows. For every n-gram arc with label w and weight c, source state s_i and destination state s_j, construct an n-gram arc with label w, weight $\langle 0, c \rangle$, source state s'_i, and destination state s'_j. The exit cost of each state is constructed as follows. If the state is non-final, $\langle \infty, \infty \rangle$. Otherwise if it final with exit cost c it will be $\langle 0, c \rangle$.

Let n be the length of the longest history string in the model. For every ϕ-arc with (backoff) weight c, source state s_i, and destination state s_j representing a history of length k, construct an ϵ-arc with source state s'_i, destination state s'_j, and weight $\langle \Phi^{\otimes(n-k)}, c \rangle$, where $\Phi > 0$ and $\Phi^{\otimes(n-k)}$ takes Φ to the $(n-k)^{\text{th}}$ power with the $\otimes$ operation. In the tropical semiring, $\otimes$ is $+$, so $\Phi^{\otimes(n-k)} = (n-k)\Phi$. For example, in a trigram model, if we are backing off from a bigram state h (history length = 1) to a unigram state, $n - k = 2 - 0 = 2$, so we set the backoff weight to $\langle 2\Phi, -\log \alpha_h \rangle$ for some $\Phi > 0$.

In order to combine the model with another automaton or transducer, we would need to also convert those models to the $\langle T, T \rangle$ semiring. For these automata, we simply use a default transformation such that every transition with weight c is assigned weight $\langle 0, c \rangle$. For example, given a word lattice L, we convert the lattice to L' in the lexicographic semiring using this default transformation, and then perform the intersection $L' \cap M'$. By removing epsilon transitions and determinizing the result, the low cost path for any given string will be retained in the result, which will correspond to the path achieved with ϕ-arcs. Finally we project the second dimension of the $\langle T, T \rangle$ weights to produce a lattice in the tropical semiring, which is equivalent to the result of $L \cap M$, i.e.,

$$\mathcal{C}_2(\textbf{det}(\textbf{eps-rem}(L' \cap M'))) = L \cap M$$

where $\mathcal{C}_2$ denotes projecting the second-dimension of the $\langle T, T \rangle$ weights, $\textbf{det}(\cdot)$ denotes determinization, and $\textbf{eps-rem}(\cdot)$ denotes ϵ-removal.

4 Proof

We wish to prove that for any machine N, $\text{ShortestPath}(M' \cap N')$ passes through the equivalent states in M' to those passed through in M for $\text{ShortestPath}(M \cap N)$. Therefore determinization of the resulting intersection after ϵ-removal yields the same topology as intersection with the equivalent ϕ machine. Intuitively, since the first dimension of the $\langle T, T \rangle$ weights is 0 for n-gram arcs and > 0 for backoff arcs, the shortest path will traverse the fewest possible backoff arcs; further, since higher-order backoff arcs cost less in the first dimension of the $\langle T, T \rangle$ weights in M', the shortest path will include n-gram arcs at their earliest possible point.

We prove this by induction on the state-sequence of the path p/p' up to a given state s_i/s'_i in the respective machines M/M'.

Base case: If p/p' is of length 0, and therefore the states s_i/s'_i are the initial states of the respective machines, the proposition clearly holds.

Inductive step: Now suppose that p/p' visits $s_0...s_i/s'_0...s'_i$ and we have therefore reached s_i/s'_i in the respective machines. Suppose the cumulated weights of p/p' are W and $\langle \Psi, W \rangle$, respectively. We wish to show that whichever s_j is next visited on p (i.e., the path becomes $s_0...s_i s_j$) the equivalent state s' is visited on p' (i.e., the path becomes $s'_0...s'_i s'_j$).

Let w be the next symbol to be matched leaving states s_i and s'_i. There are four cases to consider: (1) there is an n-gram arc leaving states s_i and s'_i labeled with w, but no backoff arc leaving the state; (2) there is no n-gram arc labeled with w leaving the states, but there is a backoff arc; (3) there is no n-gram arc labeled with w and no backoff arc leaving the states; and (4) there is both an n-gram arc labeled with w and a backoff arc leaving the states. In cases (1) and (2), there is only one possible transition to take in either M or M', and based on the algorithm for construction of M' given in Section 3.2, these transitions will point to s_j and s'_j respectively. Case (3) leads to failure of intersection with either machine. This leaves case (4) to consider. In M, since there is a transition leaving state s_i labeled with w, the backoff arc, which is a failure transition, cannot be traversed, hence the destination of the n-gram arc s_j will be the next state in p. However, in M', both the n-gram transition labeled with w and the backoff transition, now labeled with ϵ, can be traversed. What we will now prove is that the shortest path through M' cannot include taking the backoff arc in this case.

In order to emit w by taking the backoff arc out of state s'_i, one or more backoff (ϵ) transitions must be taken, followed by an n-gram arc labeled with w. Let k be the order of the history represented by state s'_i, hence the cost of the first backoff arc is $\langle (n-k)\Phi, -\log(\alpha_{s'_i}) \rangle$ in our semiring. If we traverse m backoff arcs prior to emitting the w, the first dimension of our accumulated cost will be $m(n-k+\frac{m-1}{2})\Phi$, based on our algorithm for construction of M' given in Section 3.2. Let s'_l be the destination state after traversing m backoff arcs followed by an n-gram arc labeled with w. Note that, by definition, $m \leq k$, and $k-m+1$ is the order of state s'_l. Based on the construction algorithm, the state s'_l is also reachable by first emitting w from state s'_i to reach state s'_j followed by some number of backoff transitions. The order of state s'_j is either k (if k is the highest order in the model) or $k+1$ (by extending the history of state s'_i by one word). If it is of order k, then it will require $m-1$ backoff arcs to reach state s'_l, one fewer than the path to state s'_l that begins with a backoff arc, for a total cost of $(m-1)(n-k+\frac{m-1}{2})\Phi$ which is less than $m(n-k+\frac{m-1}{2})\Phi$. If state s'_j is of order $k+1$, there will be m backoff arcs to reach state s'_l, but with a total cost of $m(n-(k+1)+\frac{m-1}{2})\Phi = m(n-k+\frac{m-3}{2})\Phi$ which is also less than $m(n-k+\frac{m-1}{2})\Phi$. Hence the state s'_l can always be reached from s'_i with a lower cost through state s'_j than by first taking the backoff arc from s'_i. Therefore the shortest path on M' must follow $s'_0 ... s'_i s'_j$. □

This completes the proof.

5 Experimental Comparison of ϵ, ϕ and $\langle T, T \rangle$ encoded language models

For our experiments we used lattices derived from a very large vocabulary continuous speech recognition system, which was built for the 2007 GALE Arabic speech recognition task, and used in the work reported in Lehr and Shafran (2011). The lexicographic semiring was evaluated on the development set (2.6 hours of broadcast news and conversations; 18K words). The 888 word lattices for the development set were generated using a competitive baseline system with acoustic models trained on about 1000 hrs of Arabic broadcast data and a 4-gram language model. The language model consisting of 122M n-grams was estimated by interpolation of 14 components. The vocabulary is relatively large at 737K and the associated dictionary has only single pronunciations.

The language model was converted to the automaton topology described earlier, and represented in three ways: first as an approximation of a failure machine using epsilons instead of failure arcs; second as a correct failure machine; and third using the lexicographic construction derived in this paper.

The three versions of the LM were evaluated by intersecting them with the 888 lattices of the development set. The overall error rate for the systems was 24.8%—comparable to the state-of-the-art on this task[1]. For the shortest paths, the failure and lexicographic machines always produced identical lattices (as determined by FST equivalence); in contrast, 81% of the shortest paths from the epsilon approximation are different, at least in terms of weights, from the shortest paths using the failure LM. For full lattices, 42 (4.7%) of the lexicographic outputs differ from the failure LM outputs, due to small floating point rounding issues; 863 (97%) of the epsilon approximation outputs differ.

In terms of size, the failure LM, with 5.7 million arcs requires 97 Mb. The equivalent $\langle T, T \rangle$-lexicographic LM requires 120 Mb, due to the doubling of the size of the weights.[2] To measure speed, we performed the intersections 1000 times for each of our 888 lattices on a 2993 MHz Intel® Xeon® CPU, and took the mean times for each of our methods. The 888 lattices were processed with a mean of 1.62 seconds in total (1.8 msec per lattice) using the failure LM; using the $\langle T, T \rangle$-lexicographic LM required 1.8 seconds (2.0 msec per lattice), and is thus about 11% slower. Epsilon approximation, where the failure arcs are approximated with epsilon arcs took 1.17 seconds (1.3 msec per lattice). The

[1] The error rate is a couple of points higher than in Lehr and Shafran (2011) since we discarded non-lexical words, which are absent in maximum likelihood estimated language model and are typically augmented to the unigram backoff state with an arbitrary cost, fine-tuned to optimize performance for a given task.

[2] If size became an issue, the first dimension of the $\langle T, T \rangle$-weight can be represented by a single byte.

slightly slower speeds for the exact method using the failure LM, and $\langle T, T \rangle$ can be related to the overhead of computing the failure function at runtime, and determinization, respectively.

6 Conclusion

In this paper we have introduced a novel application of the lexicographic semiring, proved that it can be used to provide an exact encoding of language model topologies with failure arcs, and provided experimental results that demonstrate its efficiency. Since the $\langle T, T \rangle$-lexicographic semiring is both left- and right-distributive, other optimizations such as minimization are possible. The particular $\langle T, T \rangle$-lexicographic semiring we have used here is but one of many possible lexicographic encodings. We are currently exploring the use of a lexicographic semiring that involves different semirings in the various dimensions, for the integration of part-of-speech taggers into language models.

An implementation of the lexicographic semiring by the second author is already available as part of the OpenFst package (Allauzen et al., 2007). The methods described here are part of the NGram language-model-training toolkit, soon to be released at `opengrm.org`.

Acknowledgments

This research was supported in part by NSF Grant #IIS-0811745 and DARPA grant #HR0011-09-1-0041. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or DARPA. We thank Maider Lehr for help in preparing the test data. We also thank the ACL reviewers for valuable comments.

References

Cyril Allauzen, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Twelfth International Conference on Implementation and Application of Automata (CIAA 2007), Lecture Notes in Computer Science*, volume 4793, pages 11–23, Prague, Czech Republic. Springer.

Jonathan Golan. 1999. *Semirings and their Applications*. Kluwer Academic Publishers, Dordrecht.

Werner Kuich and Arto Salomaa. 1986. *Semirings, Automata, Languages*. Number 5 in EATCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin, Germany.

Maider Lehr and Izhak Shafran. 2011. Learning a discriminative weighted finite-state transducer for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, July.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88.

Mehryar Mohri. 2002. Semiring framework and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press, Oxford.

Brian Roark, Murat Saraclar, and Michael Collins. 2007. Discriminative n-gram language modeling. *Computer Speech and Language*, 21(2):373–392.