Statistical machine translation using streaming string transducers

Thomas Ruprecht
Technische Universität Dresden
thomas.ruprecht@tu-dresden.de

Abstract

Nondeterministic streaming string transducers (*nsst*) [1] are string-to-string transducers and feature a finite set of states, as usual. Each transition reads exactly one input symbol and composes several output strings that are stored in registers; this may introduce new symbols and arbitrarily compose the strings previously stored in the registers, but not duplicate them. Since there are no further restrictions, the input language of each nsst is regular. The properties of the string-to-string transductions imposed by nsst were examined intensively by Alur and Deshmukh [1], yet – compared to well-established transducer formalisms, such as finite state transducers – this formalism is rather young. Apart from the transduction computed by nsst, Bojańczyk [2] also introduced origin semantics for this transducer model. They consider the relation between input and output positions with respect to which output positions were introduced when the input position was read. Bojańczyk, Daviaud, Guillon, and Penelle [3] characterized the class of these origin graphs for nsst and compared it to the classes of origin graphs defined transductions by two-way automata and monadic second-order logic.

Statistical machine translation (*smt*) deals with the automatic translation from sentences of one natural language into sentences of another natural language [6]. Typically, large bilingual corpora are used to automatically induce a finite set of probabilistic translation rules of some underlying formalism. This set of rules serves for the translation of previously unseen sentences. Recently Nederhof and Vogler [8] investigated the string-to-string transduction imposed by nsst whose input is restricted to a multiple context-free language, resulting in transductions of synchronous multiple context-free grammars. As synchronous multiple context-free grammars were successfully applied in smt [5], the question arises how nsst perform in this setting.

Here we consider *probabilistic nondeterministic streaming string transducers* (*pnsst*) as an underlying formalism for smt. We present an induction for pnsst from bilingual corpora with word alignments (which are treated as origin graphs). Using techniques known for the induction of multiple context-free grammars [7], we define nsst that recognize a given set of origin graphs. Weights for each transition are induced according to the maximum likelihood estimate [4] with respect to the set of origin graphs. Some properties of the induced pnsst are discussed in the context of inference, i.e. the prediction of the most probable origin graph for a fixed input.

References

[1] R. Alur and J. V. Deshmukh. “Nondeterministic streaming string transducers”. In: *ICALP*. Springer. 2011.

[2] M. Bojańczyk. “Transducers with origin information”. In: *ICALP*. Springer. 2014.

[3] M. Bojańczyk, L. Daviaud, B. Guillon, and V. Penelle. “Which classes of origin graphs are generated by transducers?” In: *ICALP*. 2017.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977).

[5] M. Kaeshammer. “Synchronous Linear Context-Free Rewriting Systems for Machine Translation”. In: *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*. ACL, 2013.

[6] A. Lopez. “Statistical Machine Translation”. In: *ACM Comput. Surv.* 40.3 (2008).

[7] W. Maier and A. Søgaard. “Treebanks and mild context-sensitivity”. In: *Proceedings of Formal Grammar*. 2008.

[8] M.-J. Nederhof and H. Vogler. “Regular transductions with MCFG input syntax”. In: *Proceedings of FSMNLP*. ACL, 2019.