Abstract

The persistent efforts to make valuable annotated corpora in more diverse, morphologically rich languages has driven research in NLP into considering more explicit techniques to incorporate morphological information into the pipeline. Recent efforts have proposed combined strategies to bring together the transducer paradigm and neural architectures, although ingesting one character at a time in a context-agnostic setup. In this thesis, we introduce a technique inspired by the *byte-pair-encoding* (BPE) compression algorithm in order to obtain transducing actions that resemble word formations more faithfully. Then, we propose a neural transducer architecture that operates over these transducing actions, ingesting one word token at a time and effectively incorporating sentence-level context by encoding per-token action representations in a hierarchical fashion. We investigate the benefit of this word formation representations for the tasks of lemmatization and context-aware morphological tagging for a typologically diverse set of languages.

For lemmatization, we use investigate an optimization technique that explores possible action sequences and scores them based on task-specific metrics instead of standard log-likelihood. We find that our approach benefits greatly languages that use less commonly studied morphological processes such as templatic processes, with up to 55.73% error reduction in lemmatization for Arabic. Furthermore, we find that projecting these word formation representations into a common multilingual space enables our models to group together action labels signaling the same phenomena in several languages, e.g. Plurality, irrespective of the language-specific morphological process that may be involved.

For morphological tagging, we investigate the effect of different tagging strategies, e.g. bundle vs individual tag prediction, as well as the effect of multilingual action representations. We find that our taggers are able to obtain up to 20% error reduction by leveraging multilingual actions with respect to the monolingual scenario.