pipilika.com
Department of CSE,
SUST

Report on Bangla Stemmer

4th December 2017

Suffix stripping

- suffixes in tri with dictionary(root words) search
- Dataset contains 2,50,000 root words, 72 suffix (but now we have 526 suffix)

Correct examples:

Original	output
শিক্ষাঙ্গনের	শিক্ষাঙ্গন
সুস্থতা	সুস্থ
নির্ভর	নির্ভর
প্রতিযোগিতামূলক	প্রতিযোগিতা
পরিকল্পিত	পরিকল্প
পাশ্চাত্যের	পাশ্চাত্য

Incorrect examples:

রাজনৈতিক	রাজনৈত
বেসরকারি	বেসরক
আসাদুজ্জামান	আসাদুজ্জা

Things can be done :

- Enriching dictionary
- Separating verb, noun, pronoun suffixes (stripping suffixes based on pos)

Seq2Seq modeling (under process)

- Seq2seq modeling with LSTM
- Training data

    সংযোগগুলোও সংযোগ

    সংযোগগুলোর সংযোগ

    হিটউইসের হিটউইস

    সিবিএসের সিবিএস

    নৌদুর্ঘটনায় নৌদুর্ঘটনা

Correct examples:

Input sequence	output sequence
তৈরির	তৈরি
দেশগুলোর	দেশ
ভাবমূর্তি	ভাবমূর্তি
প্রতিযোগিতায়	প্রতিযোগিতা
অনির্দিষ্টকালের	অনির্দিষ্টকাল

Incorrect examples:

Input sequence	output sequence
বেড়ে	বেড
স্তিমিত	স্তিমি
পড়লে	পডল
হোসেন	হোসন
জুবায়ের	জুবায়ু

Things can be done :

- Making a good training dataset which covers almost every possible word inflection
- Trying different RNN structures
- A training dataset can be made by permutation of root words and prefix list

Comparision :

- Almost **55%** of output from both approach are same

Where RNN beats suffix stripping:

Original	seq2seq	suffix stripping
একাডেমিক	একাডেমি	একাডেম
মেধাবী	মেধাবী	মেধাব
শিক্ষা	শিক্ষা	শিক্ষ
শিক্ষার্থী	শিক্ষার্থী	শিক্ষার্থ
কর্মজীবনেও	কর্মজীবন	কর্মজীবনে
চর্চায়	চর্চা	চর্চায়
ভর্তি	ভর্তি	ভর্
যায়	যা	যায়
এক্ষেত্রেও	এক্ষেত্র	এক্ষেত্রে

Where suffix stripping beats RNN:

Original	seq2seq	suffix stripping
অন্যসব	অন্যসব	অন্য
নিয়মানুবর্তিতা	নিযমানুবর্তিতা	নিয়মানুবর্তি
পর্যায়ের	পর্যায	পর্যায়
সরকারি	সরকা	সরকার

Original	seq2seq	suffix stripping
বিক্রয়	বিক্য	বিক্রয়
দীর্ঘমেয়াদি	দীর্ঘমেযুদা	দীর্ঘমেয়াদ

Where RNN greatly fall:

Original	seq2seq
বা	বাং
কি	কিং
পর	অরপ
প্রক্রিয়ার	প্রক্যিরা

More tests on newspaper data:

1.original strings(sports) :

সমুদ্র সৈকত নামের একটি বাস ঢাকা থেকে কুয়াকাটার দিকে যাচ্ছিল পথে চালক নিয়ন্ত্রণ হারালে বাসটি রাস্তার পাশে একটি গাছের সাথে ধাক্কা খেয়ে খাদে পড়ে যায় এতে ঘটনাস্থলেই একজনের মৃত্যু হয় ভারতের ইনিংসের ৫০তম ওভারটি করেছিলেন শানাকা টিভি রিপ্লেতে দেখা গিয়েছে সেই ওভারে বলের সিম খুঁটছিলেন তিনি এতে আইসিসির আচরণবিধির অনুচ্ছেদ ভঙ্গের অপরাধ করেন এই পেসার জরিমানার পাশাপাশি তাঁর নামে তিনটি ডিমেরিট পয়েন্ট যোগ করা হয়েছে

suffix stripping output (dictionary size 2,52,000):

সমুদ্র সৈকত নামে এক বাস ঢাক থেক কুয়াকাটা দিক যাচ্ছিল পথ চালক নিয়ন্ত্রণ হারাল বাস রাস্তা পাশ এক গাছ সাথে ধাক্কা খেয়ে খাদ পড়ে যা এত ঘটনাস্থল এক মৃত্যু হয় ভারতের ইনিংস ৫০তম ওভার করেছিলে শানাকা টিভি রিপ্লে দেখ গিয়েছ সেই ওভার বল সিম খুঁটছিলেন তিন এত আইসিসি আচরণবিধি অনুচ্ছেদ ভঙ্গ অপরাধ করেন এই পেস জরিমানা পাশাপাশি তাঁর নাম তিন ডিমেরিট পয়েন্ট যোগ করা হয়েছ

seq2seq model output (trained with 60500 example,80% training,20% val):

সমুদ্র সৈকত নামে এক বাস ঢাকা থেকে কুয়াকাটা দিক যাচ্ছি পথ চালক নিয়ন্ত্রণ হারাল বাস রাস্তা পাশ এক গাছ সাথ ধাক্কা খেয় খাদ পড় যায় এত ঘটনাস্থল এক মৃত্যু হয় ভারত ইনিংস পতম ওভার করেছিলেন শানাকা টিভি রিপ্লে দেখা গিয়েছ সেই ওভার বল সিম খুঁটছিল তিনি এত আইসিসি আচরণবিধি অনুচ্ছেদ ভঙ্গ অপরাধ করেন এই পেসা জরিমানা পাশাপাশি তাঁর নাম তিন ডিমের পয়েন্ট যোগ করাণ্ হয়েছ

2.original string (politics):

ঢাকা বিশ্ববিদ্যালয় কেন্দ্রীয় ছাত্র সংসদ ডাকসু নির্বাচনের দাবিতে ওয়ালিদ আশরাফের অনশন চলছে বিশ্ববিদ্যালয়ের পক্ষ থেকে গতকাল রোববার পর্যন্ত কোনো ঘোষণা আসেনি উল্টো প্রক্টর বলছেন আন্দোলনকারীরা নৈরাজ্য চালিয়ে বিশ্ববিদ্যালয়কে অস্থিতিশীল করার চেষ্টা করছেনগত ২৫ নভেম্বর বিকেল পাঁচটার দিকে বিশ্ববিদ্যালয়ের শিক্ষার্থী ওয়ালিদ এককভাবে অনশনে বসেন গতকাল অষ্টম দিন পর্যন্ত বামপন্থী বিভিন্ন ছাত্র সংগঠন পৃথক ও জোটগতভাবে তাঁর সঙ্গে সংহতি জানিয়েছে ঢাকা জগন্নাথসহ কয়েকটি বিশ্ববিদ্যালয়ের শিক্ষকেরাও একাত্মতা পোষণ করেছেনএর মধ্যে গত শনিবার সন্ধ্যায় ওয়ালিদ অসুস্থ হয়ে পড়লে আন্দোলনকারী লোকজন তাঁকে ঢাকা মেডিকেল কলেজ হাসপাতালের জরুরি বিভাগে নিয়ে যান সেখানে চিকিৎসক তাঁকে স্যালাইন নেওয়ার পরামর্শ দেন কিন্তু শিক্ষার্থীরা বলছেন গতকাল পর্যন্ত তিনি স্যালাইন না নিয়ে পানি খেয়েই অনশন চালিয়ে যাচ্ছেন এখনো তিনি স্মৃতি চিরন্তনে উপাচার্যের বাসভবনের সামনে অবস্থান করছেন

suffix stripping output (dictionary size 2,52,000):

ঢাক বিশ্ববিদ্যালয় কেন্দ্রীয় ছাত্র সংসদ ডাকসু নির্বাচন দাবি ওয়ালিদ আশরাফ অনশন চলছ বিশ্ববিদ্যালয় পক্ষ থেক গতকাল রোবব পর্য কোন ঘোষণ আসেন উল্ট প্রক্টর বলছ আন্দোলনকারী নৈরাজ্য চাল বিশ্ববিদ্যালয় অস্থিতিশীল করার চেষ্ট করছেনগত ২৫ নভেম্বর বিকেল পাঁচটা দিক বিশ্ববিদ্যালয় শিক্ষার্থ ওয়ালিদ একক অনশন বসেন গতকাল অষ্টম দিন পর্য বামপন্থ বিভিন্ন ছাত্র সংগঠন পৃথক ও জোটগত তাঁর সঙ্গ সংহত জানিয়েছ ঢাক জগন্নাথ কয়েক বিশ্ববিদ্যালয় শিক্ষক একাত্ম পোষণ করেছেন মধ্য গত শনিব সন্ধ্যা ওয়ালিদ অসুস্থ হয় পড়ল আন্দোলনকার লোক তাঁ ঢাক মেডিকেল কলেজ হাসপাতাল জরুর বিভাগ নিয় যান সেখান চিকিৎসক তাঁ স্যালাইন নেওয়া পরামর্শ দেন কিন্তু শিক্ষার্থী বলছ গতকাল পর্য তিন স্যালাইন না নিয় পান খেয়ে অনশন চাল যাচ্ছ এখন তিন স্মৃ চিরন্তন উপাচার্য বাসভবন সামন অবস্থ করছ

seq2seq model output (trained with 60500 example,80% training,20% val):

ঢাকা বিশ্ববিদ্যালয়ে কেন্দ্যী ছাত্ সংসদ ডাকসু নির্বাচন দাবি ওয়ালিদ আশরাফ অনশন চলছে বিশ্ববিদ্যালয়ে পক্ষ থেক গতকাল রোববা পর্যন্ত কোনো ঘোষণা আসেনি উল্টো প্রক্ট বলছেন আন্দোলনকারী নৈরাজ্য চালি বিশ্ববিদ্যালয়ে অস্থিতিশী করা চেষ্টা করছেনগু অি নভেম্ব বিক পাঁচটা দিক বিশ্ববিদ্যালয়ে শিক্ষার্থী ওয়ালিদ এককভাবে অনশন বসেন গতকাল অষ্টম দিন পর্যন্ত বামপন্থী বিভিন্ন ছাত্ সংগঠন পৃথক ওং জোটগতভাবে তাঁ সঙ্গ সংহতি জানিয়েছ ঢাকা জগন্নাথ কয়েক বিশ্ববিদ্যালয়ে শিক্ষক একাত্মতা পোষণ করেছেন মধ্য গতি শনিবা সন্ধ্যা ওয়ালিদ অসুস্থ হয় পড়ল আন্দোলনকারী লোক তাঁক ঢাকা মেডিক কলেজ হাসপাতাল জরু বিভাগ নিয়ে যান সেখান চিকিৎসক তাঁক স্যালাইন নেওয়া পরামর্শ দেন কিন্তু

শিক্ষার্থী বলছেন গতকাল পর্যন্ত তিনি স্যালাইন না নিয়ে পানি খেয়ে অনশন চালি যাচ্ছেন এখনো তিনি স্মৃতি চিরন্তন উপাচার্য বাসভবন সামন অবস্থান করছেন