Project: spaCy

2015-01-29T12:35:09Z aikinci 36comments

Additional Language Support

Web url: https://github.com/explosion/spaCy/issues/15

API url: https://api.github.com/repos/explosion/spaCy/issues/15

How can I add another language?

-------------------------------------------------------------------------

2015-02-01T22:26:59Z chasseurmic

Same question here: where one should begin in order to add support for other languages? Would be glad to lend a hand...

-------------------------------------------------------------------------

2015-02-14T17:29:05Z scari

+1. I would gladly lend a hand.

-------------------------------------------------------------------------

2015-02-14T17:51:57Z honnibal

Thanks everyone. I'll say a bit more about what's blocking this.

The lexemes.bin data file has been constructed in a way that depends on various intermediate data files --- for instance, I processed an unannotated corpus into a list of word counts, and then smoothed the counts with another script, and then consumed the smoothed probabilities with the current quick-and-dirty make\_lexicon.py script, which isn't even in this repository yet. I also need to set up a program to generate Brown clusters, and configure word2vec to generate word vectors.

Finally, I need to document the process, and document the tokenizer file formats, so that I can describe what you'll actually need to do to add new languages.

Instead of doing these things, I've mostly been doing bug-fixes, improving the API docs, and trying to improve my deployment process, which at the moment feels very error-prone.

I'll say a little bit about what will be required to add new languages.

1. Select an unannotated corpus. This will probably be Wikipedia --- it's a nice way to streamline things across languages. Another nice solution would be to run a language identification program over Common Crawl dumps, so that we can get text from wider genres.

2. Select an annotated corpus. This will define the tokenization standards that we have to target. I'll be licensing the data, so you probably won't have direct access to it --- I'll have to do the actual training. It will probably be nice to give you a web API, so you can run things. If the API you call is emailing me "hey, train this model", well...That sucks.

3. Define tokenization rules. This is mostly a list of prefix tokenization, a list of suffix tokenization, and a list of special-cases, which are exact-matched. The "How It Works" page says a little bit more about this, but not enough.

4. Write a lemmatizer and morphological analyser. If there's a WordNet for the language you're targeting, and it's any good, I would prefer to lemmatize to WordNet sense-keys. The BabelNet project is probably the useful way to go about this.

Having written all this, I'm thinking it might be nice to inter-operate closely with Gensim on this. Gensim will give us the word2vec implementation, and would be a good way to handle the boot-strapping problem: if getting spaCy to work on a new language initially depends on processing a bunch of unannotated data with spaCy, then things are awkward.

I'll think more about this, and probably reach out to Radim about it.

-------------------------------------------------------------------------

2015-05-07T23:20:55Z geovedi

When I was still using redshift, I used word2vec and gensim to generate word cluster with `window\_size=2`. It was fast, but wasn't quite happy with the result until I found [Percy Liang's](https://github.com/percyliang/brown-cluster) word clustering tool. Took days to generate since I have large corpus.

-------------------------------------------------------------------------

2015-08-20T14:38:25Z brunoalano

If you need, I've a lot of resources for Portuguese (word2vec, Stanford NE Extractor, ConLL Floresta Sint√°tica, POS Tagger (trigram, bigram, n-gram)...) that I've trained for my project.

-------------------------------------------------------------------------

2015-08-20T15:05:39Z aikinci

@brunoalano thanks but actually I am interested in Turkish for the beginning.

-------------------------------------------------------------------------

2015-09-15T06:53:20Z honnibal

Working on docs for this here: http://spacy.io/tutorials/add-a-language/

-------------------------------------------------------------------------

2015-09-19T10:58:34Z bittlingmayer

+1

Ideally we could develop the concept of explicit \_partial support\_ of a new language. For example, tokenisation really works already for most languages.

-------------------------------------------------------------------------

2015-09-20T18:09:45Z dpk

I would be very interested in spaCy support for German, especially official support. Thanks for the documentation on adding languages!

-------------------------------------------------------------------------

2015-10-03T03:20:02Z honnibal

See Issue #124

-------------------------------------------------------------------------

2016-01-11T13:27:24Z evanmiltenburg

> Select an unannotated corpus. This will probably be Wikipedia --- it's a nice way to streamline things across languages. Another nice solution would be to run a language identification program over Common Crawl dumps, so that we can get text from wider genres.

Perhaps the corpora from http://corporafromtheweb.org/ are useful here. They're large, come in several different languages (Dutch, English, French, German, Spanish, Swedish), and are tokenized pretty well. They're also POS-tagged and lemmatized, but I don't know whether that's needed here.

-------------------------------------------------------------------------

2016-05-05T09:16:58Z honnibal

I'm pleased to say that there's now excellent support for German, thanks to the great work from our first NLP employee Wolfgang Seeker. We're still finishing up the blog post etc, but the model is uploaded and can be used from spaCy 0.100.7.

```

import spacy

de\_nlp = spacy.load('de')

```

We're still refactoring and working on better processes for adding more languages. I'm going to close this issue because it's old and most of the information here is now out of date.

Obviously, there's still a lot to do to support more languages. And I think the idea of partial language support is important and overdue. But --- progress :)

-------------------------------------------------------------------------

2016-05-13T15:25:33Z davidsbatista

I'm also glad to help in adding Portuguese tools to spaCy.

By the way, anyone knows if there is corpora to train dependency parsers or just PoS-taggers to Dutch?

-------------------------------------------------------------------------

2016-05-13T15:36:36Z evanmiltenburg

Yes there are corpora for that: Lassy-Klein, Lassy-Groot, and SoNaR. Probably the best thing is to contact the [TST-centrale](http://tst-centrale.org/nl/). They are undergoing some changes in management, so they might be slow to respond.

For questions about Lassy, you can email [Gertjan van Noord](http://www.let.rug.nl/~vannoord/).

For questions about SoNaR you can email [Nelleke Oostdijk](http://www.ru.nl/taalwetenschap/wie\_wat\_waar-oud/medewerkers/nelleke-oostdijk/).

I trained a POS-tagger on the NLCOW14 corpus, which was automatically tagged. But if you need something fast, then [here](https://github.com/evanmiltenburg/Dutch-tagger) is the repository.

If you just need a parser, try [Alpino](http://www.let.rug.nl/vannoord/alp/Alpino/).

-------------------------------------------------------------------------

2016-05-13T15:43:12Z davidsbatista

Thanks for your reply! One more questions, does the license associated to those datasets allows to be incorporated into spaCy or use it to process dutch text not within a research context?

-------------------------------------------------------------------------

2016-05-13T15:58:15Z evanmiltenburg

I think if you're a researcher yourself, and you don't contribute to SpaCy for money I think it should be OK.

If you're a SpaCy employee it might technically be commercial use and the terms and conditions from the TST-centrale are different: I just checked and the license for Lassy-klein for commercial use costs 2000 euros. Same for SoNar-klein. The larger corpora cost more. Then again, SpaCy is open-source. So you can always try. If it doesn't work, contact Gertjan to see whether he can help you.

Else there's also [this free smaller treebank](https://www.let.rug.nl/vannoord/trees/) on his website. That should at least get you started. And the people at the university of Groningen would probably be happy to see another parser for Dutch so they can compare Alpino to it :)

Also look around on [Maarten van Gompel](http://proycon.anaproy.nl/)'s GitHub [here](https://github.com/proycon). He's worked on memory-based POS tagging for Dutch, among other things.

-------------------------------------------------------------------------

2016-05-13T16:03:02Z davidsbatista

Dank u wel for all the clarifications!

-------------------------------------------------------------------------

2016-05-16T14:07:22Z evanmiltenburg

You're welcome, but I just realized I forgot to mention http://universaldependencies.org/ which also covers Dutch, and seems to overlap with Lassy. Sorry!

-------------------------------------------------------------------------

2016-07-18T10:43:14Z kaizenx

I new here but have been trolling for a while, would be glad to add Bahasa support

-------------------------------------------------------------------------

2016-07-18T12:23:02Z syllog1sm

@geovedi , were you working on Bahasa?

-------------------------------------------------------------------------

2016-07-18T13:07:44Z geovedi

@syllog1sm yes. it's been awhile tho. will catch up with the latest commit and regenerate the model.

-------------------------------------------------------------------------

2016-07-19T07:28:55Z kaizenx

@geovedi do you have a fork or branch i can check out?

-------------------------------------------------------------------------

2016-07-29T09:19:09Z korobool

Does anybody plan to add a support for Russian?

I wanted to understand how much effort and what road-map should I plan to add support for Russian, but unfortunately page http://spacy.io/tutorials/add-a-language/ is not available.

Could you please point me to corresponding doc/guide?

I'm also reading pull requests about adding of a German support, do you think the work for adding of Russian support is going to be kind of similar?

-------------------------------------------------------------------------

2016-08-23T14:11:33Z savkov

Same question as @korobool -- I would be interested in adding support for Bulgarian.

-------------------------------------------------------------------------

2016-08-23T14:32:40Z tpetmanson

Just to push towards adding for additional language, I would be willing to work on Estonian support as I have experience developing https://github.com/estnltk/estnltk, which is a NLP library specific to the language.

-------------------------------------------------------------------------

2016-09-06T05:06:03Z jasonmhead

http://spacy.io/tutorials/add-a-language/

is 404'ing

-------------------------------------------------------------------------

2016-10-21T06:50:47Z Biswajit2902

http://spacy.io/tutorials/add-a-language/

is not working. It seems page is not available (404 Error).

-------------------------------------------------------------------------

2016-10-30T14:22:50Z kaustubhn

Hi, I am interested in adding a language I have a corpus for 4 Indian languages, would be willing to help add a spacy model.

P.S: the URL is not working.

Thanks

-------------------------------------------------------------------------

2017-01-10T04:06:23Z ausiddiqui

https://spacy.io/docs/usage/adding-languages

-------------------------------------------------------------------------

2017-01-10T10:03:30Z ines

@kaustubhn Sounds great ‚Äì for more info on how to add languages, see the new link posted above.

There's also an open discussion on getting started with the Indic language tokenizers in #641, so this would probably be the best place to talk more about this.

-------------------------------------------------------------------------

2017-09-12T14:52:58Z martinbel

Hi @davidsbatista. Do you have some code to train a model in portuguese?

-------------------------------------------------------------------------

2017-09-12T15:31:38Z davidsbatista

Train what exactly? For PoS and syntactic parsing there is public available data, for named-entity recognition, is not so easy, there is no public available dataset

-------------------------------------------------------------------------

2017-09-12T19:27:10Z martinbel

The training would be what's mentioned in the docs above. At this point PoS would be enough for me. Were you able to implement this in spaCy?

-------------------------------------------------------------------------

2017-09-13T08:31:54Z davidsbatista

I could started it, although at the moment I have a bit of too much in my plate, but if you give me some guidelines on how to start I can try to plug-in the open PoS corpora in spaCy.

-------------------------------------------------------------------------

2017-09-13T08:49:37Z davidsbatista

got it: https://spacy.io/docs/usage/adding-languages

:)

-------------------------------------------------------------------------

2018-05-08T17:27:29Z lock[bot]

This thread has been automatically locked since there has not been any recent activity after it was closed. Please open a new issue for related bugs.

-------------------------------------------------------------------------