Differential approach to webcorpus construction

TATIANA SHAVRINA
NRU HSE

02.06.2018
RSUH, MOSCOW

Taiga project

Home | Corpus Annotation | Segments | Downloads | View on GitHub

- open-source corpus for machine learning tasks
- 14 resources – news, social media, fiction, journals
- 6 billion tokens
- morpho- and syntactic annotation
- conllu, universal dependencies
- different types of text annotation

Corpus that can be a useful instrument not only for linguists, but for data scientists and engineers
tatianashavrina.github.io/taiga_site/

Corpus annotation

Morphology, lemmatization and syntax – UDPipe trained on SynTagRus

Additional annotation:

- genres and text types
- difficulty (readability)
- dialogs (with parallel data in English, German, Italian)
- authorship
- rubrics and themes
- tags
- dates, etc

Usage examples

- Embeddings on stihi.ru
- Closest words to ‘роза’ (rose)

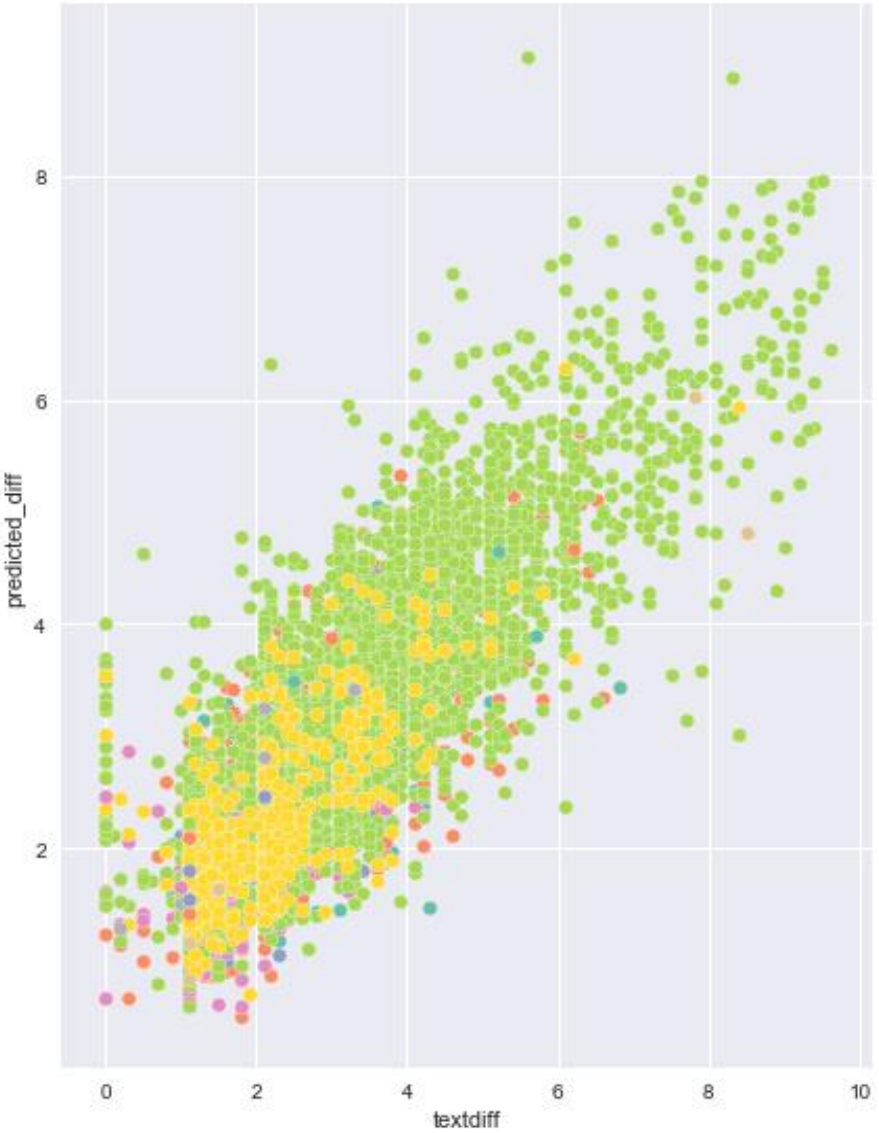
News		**Araneum fastText**		**RNC**		**Taiga poems**	
фрезия	freesia	орхидея	orchid	цветы	flowers	оброза	icons
гвоздика	carnation	гортензия	hydrangea	гиацинт	hyacinth	береза	birch
сябитова	syabit	фрезия	freesia	хризантема	chrysanthemum	проза	prose
хризантема	chrysanthemum	хризантема	chrysanthemum	лилия	lily	фроза	frost
ирис	iris	гербера	gerbera	маргаритка	daisy	эмброза	embrose
рымбаева	rymbayeva	маргаритка	daisy	нарцисс	narcissus	розалия	rosalia
гербера	gerbera	фиалка	violet	гелиотроп	heliotrope	гроза	storm
яснення	clarification	ирис	iris	букет	bouquet	лукроза	lucrose
сирень	lilac	лилия	lily	астр	astern	бяроза	birch
яснювальнуть	**error**	цветок	flower	цветок	flower	розали	Rosalie

Usage examples

- Regression model trained on NPlus1 : text difficulty from 0 to 10

- also: chat-bot training, classification
- and clustering,
- genre-specific generators, etc

- not suitable for POS and syntactic
- annotation – now we have manually
- verified UD subcorpus of Taiga
- **universaldependencies.org/**
- **Treebanks/ru_taiga/index.html**



Webcorpus construction approaches

- **classic crawling** – crawling every resource, address to search engines and walk through the pages; and after the material is cleaned from spam and is deduplicated. (example – Common Crawl)

- **fitted** – all the materials from the listed thousands of URLs are crawled. Sometimes a fit-function is used, which decodes whereas URL is suitable or not, while crawler addresses to the search engines, like in the first approach (example – Aranea corpora)

- **differential** – a small number of large resources are crawled, but they are downloaded as completely as possible, entirely, if possible. This download allows us to state that linguistics variation of the resource is covered entirety (example – General Internet-Corpus of Russian, Taiga).

Perspectives of differential corpora

- The idea: there are homogeneous sources on the web, that are not difficult to collect completely – news sites, journals, etc.

(yet some resources cannot be collected on the whole (social networks, scientific papers, etc) – but we already have GICR, arxiv.org datasets)

- Taiga unites that kind of web resources, that can be collected completely, if they are useful for ML tasks or have interesting annotation – this way we are pretending to have representative subcorpora.

  The resulting corpus segments can be used for further experiments with corpus segmentation and obtaining representativity definition

  → Which small data is good and which is not for the task?

  This question better be answered if you have the information about variation in the whole big data corpus.

Thank you!