Neural Machine Translation: English-Russian
Weekly report #1

Temur Kholmatov	Abdurasul Rakhimov	Jafar Badour
Computer Science, DS-01	Computer Science, DS-02	Computer Science, DS-01
Innopolis University	Innopolis University	Innopolis University
t.holmatov@innopolis.ru	a.rahimov@innopolis.ru	j.badour@innopolis.ru

Introduction

In this project, two datasets for Neural Machine Translation will be explored and compared.

- **News Commentary dataset** :

  The News Commentary Dataset is retrieved from the News Commentary website. The website contains datasets for parallel translated sentences of news in 13 different languages. The dataset can be downloaded here.

- **Yandex NLP dataset** :

  Yandex offers a dataset for parallel sentences in Russian and English. This dataset outsizes the News Commentary dataset in sentence count.

In the first week, the tasks completed are:

a. Search for proper datasets with maximal relation to our client base.
b. Make collective decisions for dataset choices.
c. Exploring two different datasets and comparing their statistical characteristics.

News Commentary dataset

The dataset contains 280984 pairs of sentences; each pair contains two sentences: The English and the Russian versions of the sentence.

Some samples from the dataset:

- Wouldn’t you know it?	- И что бы вы думали?
- Since their articles appeared, the price of gold has moved up still further.	- С тех пор как вышли их статьи, стоимость золота повысилась еще больше.

Yandex dataset

Yandex dataset contains one million pairwise English-Russian sentences. The Yandex dataset contains more data points as well as it covers more topics in its sentences.

Some samples from the dataset:

- Now you have Black Sabbath and Kiss tribute albums.	- А сейчас куча триьютов тем же самым BLACK SABBATH и KISS.
- I was the one who sat down and copied them.	- Я был единственным, кто занялся копированием демо на кассете.

Note that the first example in the English section pairs with the first example in the Russian section.

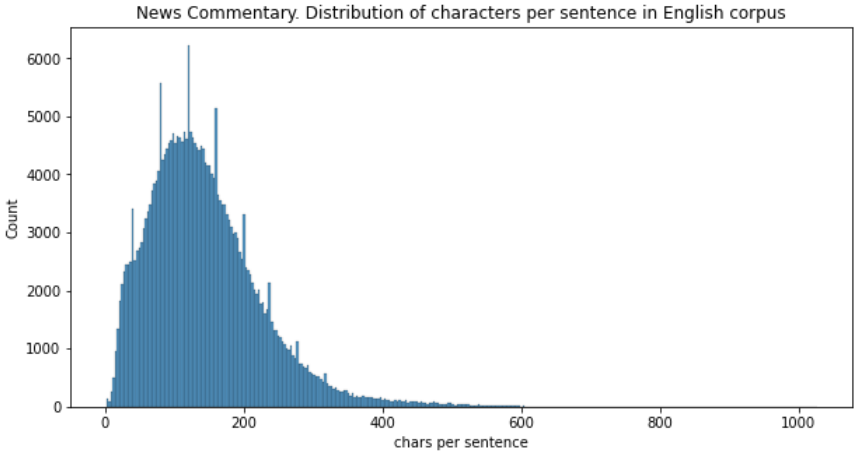
Analyzing text statistics

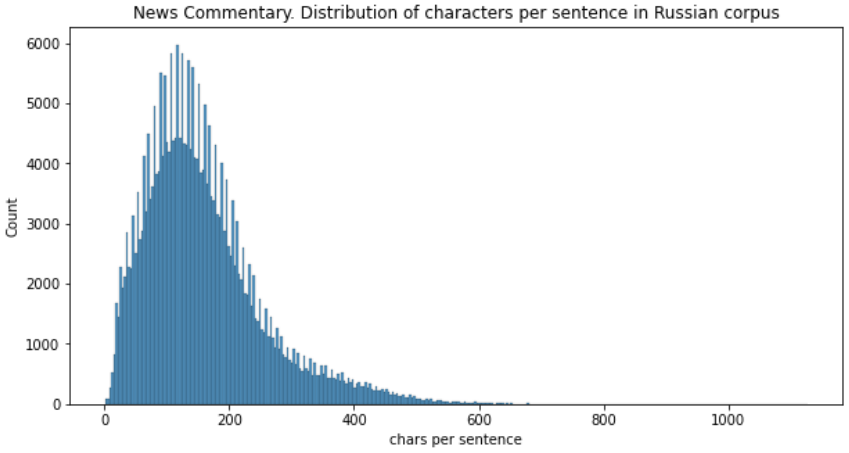
Basic description of the datasets

	Language	# of tokens	# of unique words	Average sentence word count	Frequent tokens
News Commentary	English	6470649	182981	23.02	'the', 'of', 'to', 'and', 'in', 'a', 'is', 'that', 'for', 'be'
	Russian	5995739	356638	21.33	'в', 'и', 'на', 'не', 'что', 'с', 'к', 'для', 'как', '_'

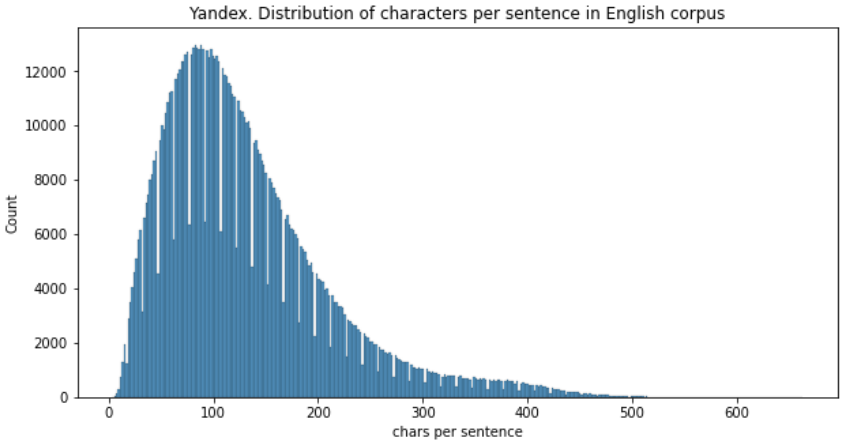
Yandex	English	21252975	796290	21.25	'the', 'of', 'and', 'to', 'in', 'a', 'is', 'for', 'that', 'with'
	Russian	18680351	1323932	18.68	'и', 'в', 'на', 'с', 'не', 'что', '-', 'для', 'по', 'к'

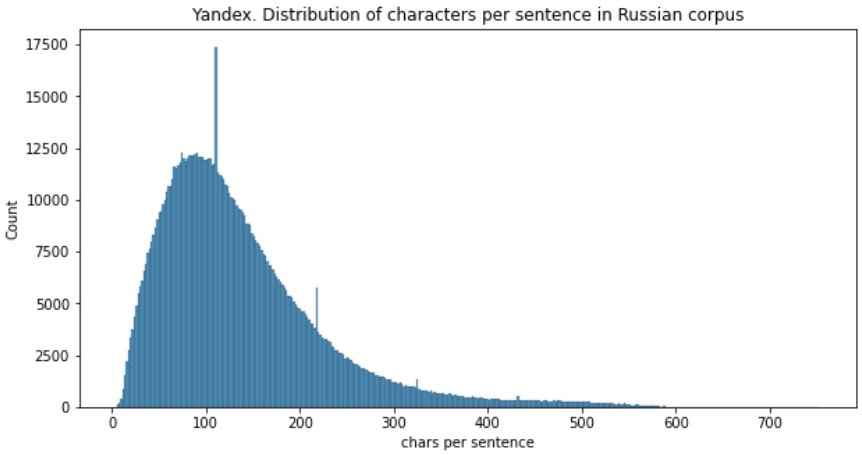
Characters per sentence





We can see that the histogram for the Russian corpus is less blunt on the sides than the one for the English corpus for the News Commentary dataset.

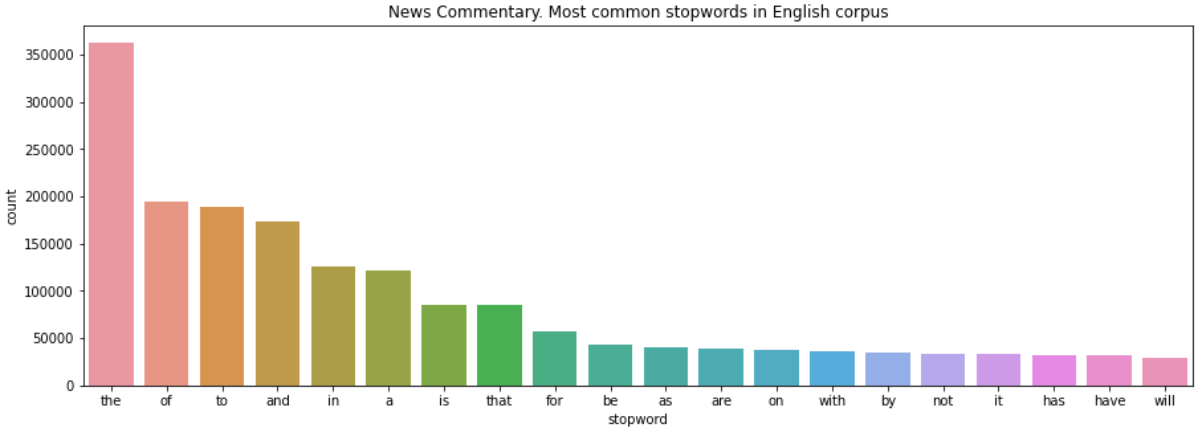


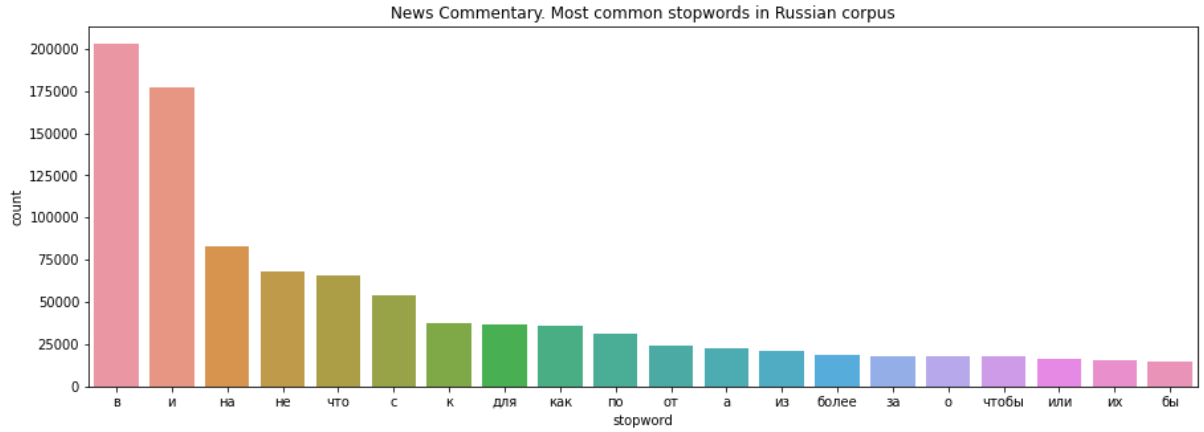


The English corpus has more variance in terms of the number of characters per sentence than the Russian corpus for the Yandex dataset.

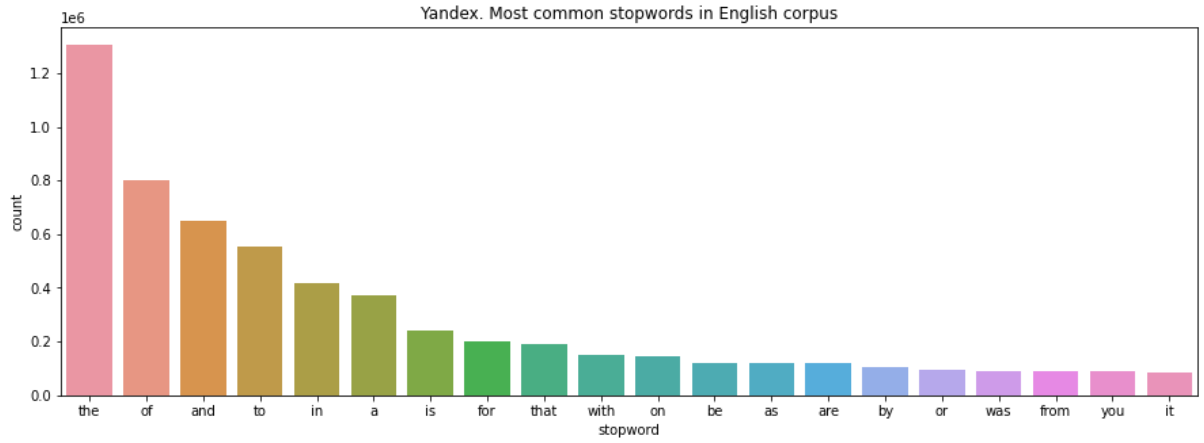
Most common stop words

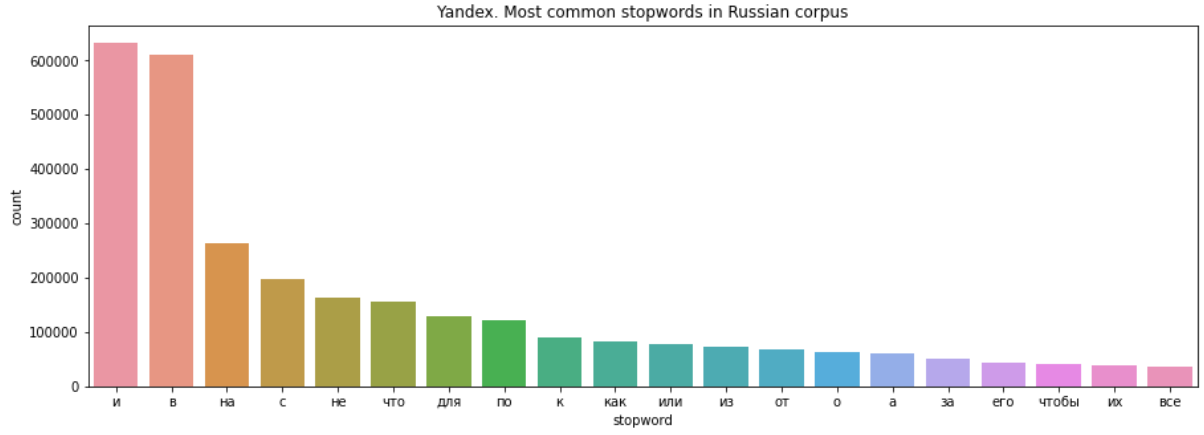
News Commentary dataset





Yandex dataset

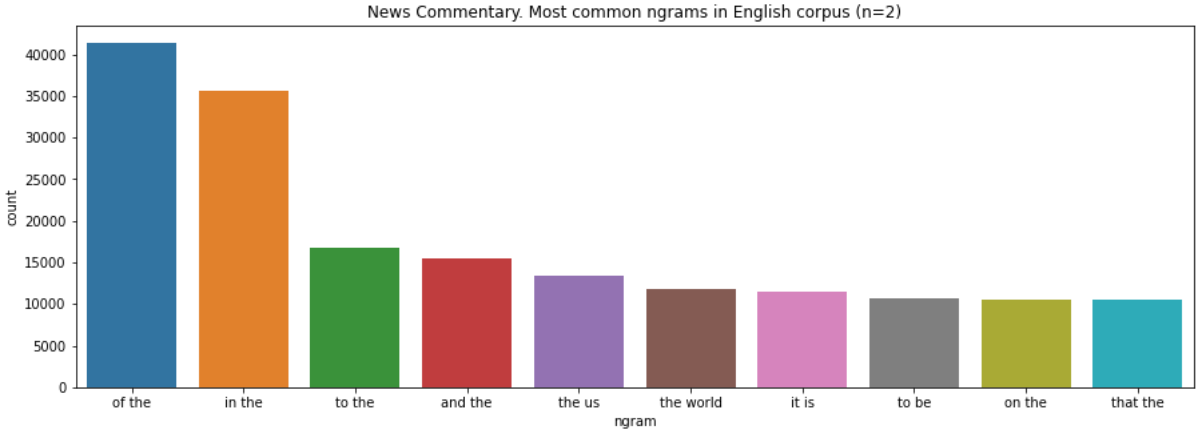


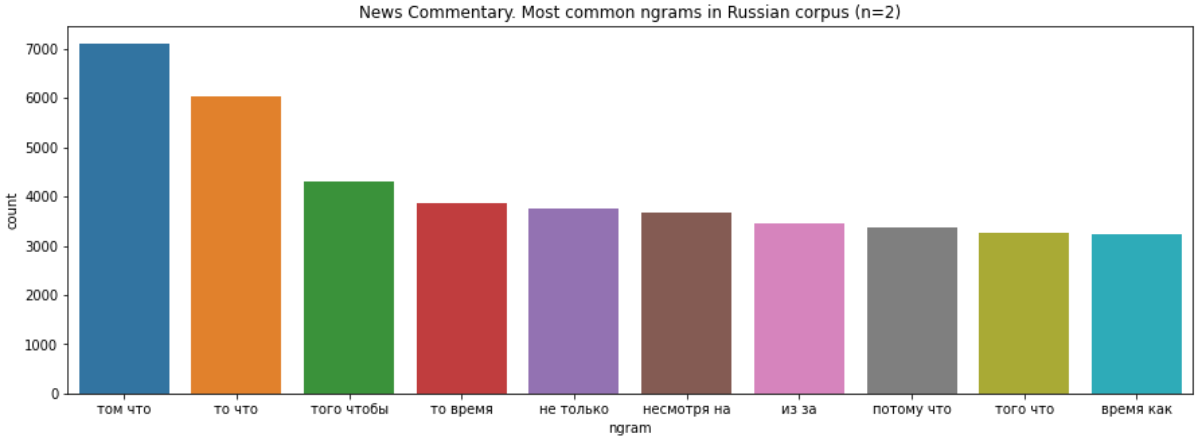


The most common stop words in both datasets are highly overlapping.

The most common n-grams

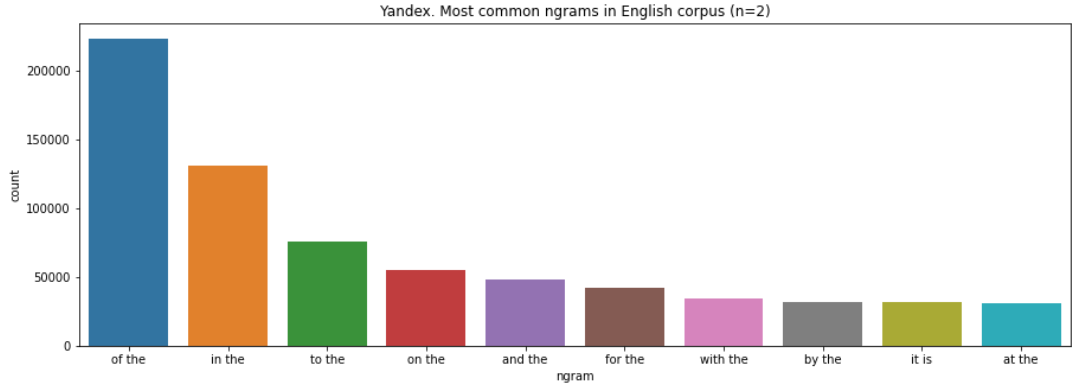
News Commentary

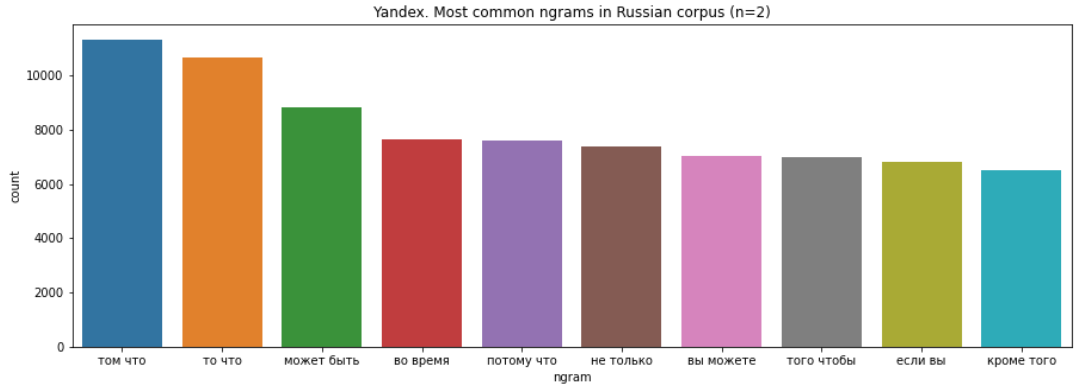




We can see that bigrams of stopwords are highly common in both corpora. 2-grams in the English version mostly contains a combination of article and preposition. The list of most common 3-grams in the English version contains common expressions, US (in different forms and combinations). Most common n-grams in the Russian version mostly contain common expressions.

Yandex





The bigrams sets in both datasets consist of similar components. N-grams in the English version mostly contain a combination of article and preposition for 2-grams and common expressions for 3-grams. Most common n-grams in the Russian version mostly contain common expressions.

Workflow

Scrum meeting	Tasks	Assignees		
		#1	#2	#3
Meeting #1	Search online for the best datasets	**AR**	**TK**	**JB**
	Licensing issues investigation	**AR**		
Meeting #2	Basic Utils functions implementation.	**AR**	**TK**	
	Testing and debugging	**JB**		
Meeting #3	Reporting structure and outlining	**AR**	**TK**	**JB**
	Report writing	**JB**		

Abdurasul Rahimov: **AR**, Temur Kholmatov: **TK**, Jafar Badour: **JB**

Final observations

News Commentary

The total number of sentences in the dataset is 280984. Most of the sentences contain around 20 words and 100 characters. Some sentences contain only special characters. The average length of the words is around 5-10. The most common words in the corpus are stopping words. Most frequent n-grams contain common expressions and combinations of articles and prepositions (for English).

Yandex

The total number of sentences in the dataset is 1 million. Most of the sentences contain around 20 words and 100 characters. Some sentences contain links to websites. The average length of the words is around 5-10. The most common words in the corpus are stopping words. Most frequent n-grams contain common expressions and combinations of articles and prepositions (for English).