Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

Яцишин Егор Сергеевич

Имплементация BERT для разрешения кореференции на материале русскоязычных медицинских форумов

Выпускная квалификационная работа студента 4 курса бакалавриата группы 171

Научный руководитель
Кандидат филологических наук, доцент

Д. А. Рыжова

____

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

____

«   » __________ 2021 г.

Научный консультант
М. А. Пономарева

____

Москва, 2021

BERT Implementation for Coreference Resolution in Russian Medical Forums

Yatsishin Egor

Fundamental and Computational Linguistics

School of Linguistics

National Research University Higher School of Economics

Bachelor Thesis

Research Advisor: Daria Ryzhova

Research Consultant: Maria Ponomareva

June 2021

Table of contents

1 Abstract 3

2 Introduction 3

3 Coreference types 5

4 Constraints for reference resolution 5

5 Analysis of coreference strategies in real-world data 7

6 The production pipeline overview 10

7 Data and format 11

8 Models 14

9 Methods 17

10 Future work 19

11 Conclusions 19

12 List of references 21

13 Appendix 23

1. Abstract

This work describes the first coreference resolution system for Russian medical forum data based on transformer architecture. To get the understanding of the data we did some preparatory work: around 700 of messages from different forums and social networks were read and typical coreference scenarios were observed, described and classified by hand. The proposed solution for the stated task of creating a coreference resolution system is an ensemble architecture, witch combines fine-tuned BERT-based model, coarse-to-fine coreference resolution model and a preprocessing pipeline. Following research and production were done on a free basis by request from Semantic Hub and with support from the mentioned company.

2. Introduction

R&D in the medical domain is a growing trend in pharmaceutical industry. Companies take risks producing new drugs, since common lack of information about real patient experience same as a great number of misdiagnosed or even undiagnosed patients can easily cause a loss of significant investments. Patients at the same time are suffering without a proper medication or a way to find the one that suits their case. Semantic Hub and other companies specialise on implementing modern natural language processing technologies to reduce the described risks significantly. We foresee a way to make a possible improvement for the system Semantic Hub executes by producing a BERT-based model for single-message-level coreference resolution on the data gathered from Russian medical forums.

The task of coreference resolution can be described as determination of mentions in the text which refer to the same entity in the real world. A group of several mentions which are coreferent is usually called a coreference chain.

(1) Andrew is the best student in class, therefore he was selected as the headman.

In sentence *(1)* mentions «Andrew» and «he» are considered coreferent as they refer to the same person.

Coreference is a complex phenomenon relevant to discourse understanding on a lower level. It has a rich history of research in natural language processing (NLP), but it is one of the fields which has seen a really slow progress in the recent years. Still, the task is vital for a large number of higher-level fields such as machine translation (Preuss, 1992), text summarisation (Steinberger et al., 2007), sentiment analysis (Cambria, 2016), question answering (Castagnola, 2002) and others. The development in the field of coreference resolution started with systems based on rule-approach (Lappin and Leass, 1994) and recently came to modern deep-learning methods (Wiseman et al, 2016; Clark and Manning, 2016a,b; Young et al, 2018b) which demonstrated a great step in performance for most NLP spheres.

Hereafter are introduced some examples to show the importance of correctly resolved coreference for some NLP tasks.

(2) *I wasn’t able to put the cupboard$_1$ in the kitchen. It$_1$ was too big.*

(3) *I wasn’t able to put the cupboard in the kitchen$_2$. It$_2$ was too small.*

Considering the machine translation task, it is impossible to correctly translate to Russian the second sentence in *(2)* and *(3).* That is due to differentiation of the grammatical gender of inanimate nouns. The translations for *(2)* and *(3)* differ:

(4) *Он$_1$ был слишком большим. (On$_1$ byl slishkom bol’shim)*

(5) *Она$_2$ была слишком маленькая. (Ona$_2$ byla slishkom mal’en’kaya)*

Another case showing the importance of coreference resolution task is from the field of question answering:

(6) *Maria$_1$ told Sasha to close the door because she$_1$ was cold.*

(7) *Maria told Sasha$_2$ to close the door because she$_2$ was screaming too loud.*

Without resolving the coreference in the sentences it is not possible to answer for sure to questions such as «Who of the girls was cold?» or «Who of the girls was screaming too loud?»

The latest significant solutions for coreference resolution in Russian were introduced at Dialog-19 conference (Budnikov and Toldova et al., 2019). Presented models showed great results on the subset of OpenCorpora entries, which are definitely different from forum threads by their genre, consisting of news for the most part. This can be observed in the tables below.

chaskor.ru (articles)	8 %
chaskor.ru (news)	28 %
Wikipedia	10 %
Wikinews	15 %
Blogs	20 %
Fiction	3 %
Non-fiction	4 %
Legal texts	11 %
Other	1 %

Table 1. Genre distribution of Dialog-19 dataset

Mentions	21486
Mentions in coreference chains	18282
Coreference chains	4110

Table 2. Train set statistics of Dialog-19 dataset

Mentions	7475
Mentions in coreference chains	6877
Coreference chains	1568

Table 3. Test set statistics of Dialog-19 dataset

At the moment the algorithms implemented at the Dialog-19 are either outdated or unavailable for commercial usage.

3. Coreference types

The task of resolving coreference becomes especially complex due to the amount of possible types of referential links that could be established between words. Sadly, this causes most resolving algorithms to face the issue of bad coverage meaning that most systems are designed for specific types of references. The pipeline described in this work is focused on resolving pronominal anaphora.

Pronominal anaphora is one of the most common type of anaphora. It usually occurs in everyday speech and prevails on other types of references. This kind of anaphora was firstly introduced by (Heim, 1982) and has been described in many works since. There are two important sub-types for pronominal anaphora:

- Indefinite pronominal

(8) Many$_1$ of the students$_1$ nowadays are politically active.

This type of reference is about pronoun referring to an entity that is not being explicitly defined and the exact relations are ambiguous.

- Definite pronominal

(9) He saw this book$_1$ in the drawer two days ago. It$_1$ had a classy hard cover and creamy-coloured pages.

The type is definite because the reference is aimed to a single object that is clearly distinguished and defined.

The system proposed in this work should be capable of solving both, but will most possibly be better at resolving the definite type since it provides more explicit syntactic and semantic features for the model to learn.

4. Constraints for reference resolution

Semantics and syntax play a major role in the approaches proposed for the coreference resolution because in the natural language there are some obvious constraints for referent linking. The phase of mention-filtering is commonly relied on these constraints as well as most of syntactic approaches. These constraints are presented below. They are crucial, but usually they are not enough to filter everything that does not fit.

- Number agreement

(10) Maria and Andrew$_1$ carried over the snacks$_2$ for the party$_3$. Unfortunately, they$_4$ were spoiled.

The coreference can not be established between entities that are of different number. The referential link is possible only in case of agreement on plural or singular form between the words. In the sentence *(10)* the pronominal reference *4* can be referent to *1* and *2* since they are all plural, while *3* is not considered as a valid option for establishing coreference.

- Person agreement

(11) Andrew and Sasha$_1$ are getting married. No one ever expected that they$_2$ could be together.

There are only three persons in linguistics and coreferent objects must agree on the person characteristic. In the sentence above the *2* refers to *1* as they agree on their person, but if the pronoun *2* is changed to «we» it becomes impossible to establish such a link.

- Gender agreement

(12) Maria$_1$ found a kitten$_2$ wandering near her flat. She took it$_2$ home.

References that are connected have to agree on their gender. This constrain is really important (Mitkov, 2014) and those entities which do not agree on this feature can be successfully eliminated.

- Grammatical role constraint

(13) Sasha$_1$ has never eaten a single shrimp. She asks her mom$_2$ to eat them for her. She$_3$ thinks they smell disgusting.

The well-known subject, object, verb system is represented in this constraint. There is evidence (Kennedy and Boguraev, 1996) that these roles can play a crucial role in detecting valid referents. The idea behind it is that subject-entities are more salient than object-entities, which means that in sentences *(13)* entity *3* refers to *1*, since *2* has lower priority.

- Mention repetition

(14) Andrew$_1$ is a real fan of Chet Baker. Tomorrow he is going to see a jazzman$_2$ playing Bakers' masterpieces at a restaurant. He$_3$ is amazed about this event.

Those entities or objects that have been introduced several times in the text or proved to be the topic of the story are considered more viable to be coreferent than others. In *(14)* it is *1* that is coreferent to *3* and not *2*, because the story is focused on Andrew in the previous sentence, making *1* more salient.

- Mention recency

(15) Maria now has two cats. Alexander$_1$ is small and calm. Arthur$_2$ who is a Persian is arrogant and snooty. Maria is obliged to take him$_3$ to the vet every week, since he has a weak spine.

Additionally to the previous point, recency is a valid and important factor for consideration when resolving coreference (Carbonell and Brown, 1988). The recently introduced entities are given more importance than those mentioned previously in the text. In the example *(15)* according to syntactic relations *3* may refer either to *1* or to *2* and the rule of recency allows to give *2* a higher priority over *1* because it was mentioned closer to *3*.

- Verbal constraint

(16) Andrew tried to grab the cat$_1$ wearing my glasses expecting it to be a hat$_2$, but it$_3$ suddenly wiggled.

In the natural language some verbs can cooccur with a certain type of entities. Here in the example the verb «wiggled» can refer only to an animate object, so *2* should be filtered out of the coreference chain because of the constraint on verbal animacy. The conclusions presented above are strongly based on world knowledge, which is better be incorporated in the system.

- World knowledge

It has been a long time since we dreamed of incorporating real world knowledge in the computer systems, but still the wide scope of the information and its' generality hold back the progress — at this moment world knowledge can be incorporated in the resolution system only to some extent. Fortunately, having a large enough corpora can sometime compensate this drawback when using transformers. This makes resolution of some obvious concepts like «rabbit-jumps» and «snake-crawls» possible at some extent.

5. Analysis of coreference strategies in real-world data

As already mentioned, the first step of the task was to determine which types of coreferential relationships are most common and important. Having multiple sources of data available: Q&A forums, social networks, opinion aggregators, web-sites of medical centres and more, we thought that starting with looking mostly at the patient forums is the best option. That is mostly because of some extra-linguistic factors important for the way people communicate and structure the information:

- Looking at the structure of different sources it is clear that most of them go with the same pattern. Every source consists of clusters — threads, which in turn consists of a header — the first main message which opens up the topic of the thread, and other messages — the answers to the stated topic. Almost every topic on patient forums is a great representation of the described structure, while other sources tend to be less strict to the unification of the data. This leads to less amount of unrelated chatting between users and less spamming overall. Considering the above, we suppose that the structure of the data-source affects its' saturation with relevant linguistic phenomenons.

- Contrary to social networks, patient forums are meant only for helping people deal with health problems. Such a restrictive position leads to people posting being more attentive to their language and the way the information is presented. Regarding the task it is easier to read and, in future, automatically process texts which have less lexical, syntactic and spelling errors.

- Threads on patient forums are covering a broader spectre of topics related to health issues: from pills prescribed to finding the best doctor for a very special case. This suggests a wider range of language forms used as well as better representation of the health-related texts. The texts are also focused on how people feel while they live through their illnesses and what they experience as patients, while the rest of the sources favour to be focused on other information. For example, review-based web-sites usually feature a lot of price comparisons and discussions about the economic benefits of some product, which is not actually interesting for the task.

Patient forums are better structured, more unified and linguistically clear while being thematically broader than other resources. For these reasons some of the threads from the patient forums, which provide their data to Semantic Hub, were manually read and classified by the type of coreference encountered. More than 700 messages of different sizes were examined and classified by the form of coreferential relation across 6 major threads. Such an overview of real specific cases is an important step in the whole task of resolving coreference in the medical domain forums since it introduces the data to the researcher and thus gives a better understanding of the possible natural language concepts related to coreference.

It was decided by Semantic Hub that in the scope of this work the main focus, apart from obvious pronominal coreference, should be on those types of specific coreference classes:

- Message mentions a drug or an illness by their names and later they are referred to by a demonstrative pronoun like «тот», «этот», «такой» («tot» - that, «etot» - this, «takoi» - this kind of) or other forms of them. The fact that this type of relation is close to adjectival anaphoric type may serve as a great support when adapting the system to a more broad spectre of reference types.

(17) Я долго принимала Топамакс$_1$ и он$_1$ оказывал должное действие, как и заявлено. Такой$_1$ нам подходит!

- I took Topamax$_1$ for a long time and it$_1$ had the proper effect, as stated. That$_1$ (referencing the Topamax as an object) suits us!

- It can be noticed that when describing a drug and the dosage people tend to skip the drug name and refer to it by just leaving the amount of the drug used or prescribed. Sometimes the dosage is changed making it even harder to resolve such referential relations.

(18) Принимаю Оземпик 200мг$_1$ утром. На 250$_1$ побочки были ужасные. Снижали до 120 мг$_1$, эффективность слишком сильно снижалась.

- I take Ozempic 200mg$_1$ in the morning. At 250$_1$, the side effects were terrible. Did reduce to 120 mg$_1$, the effectiveness decreased too much.

6. The production pipeline overview

Now, with the understanding of linguistic part of the coreferential relations, the system for the resolution of these relations has to be constructed. In other words there are three points which require to make a decision:

1. Learning algorithms
   Which machine learning algorithm should be used? Are there any additional systems that can contribute to the results?

2. Dataset
   Which parameters should the dataset contain to represent the previously mentioned constraints? What are possible ways to create the dataset and adapt it to the described needs? Are there already any free-distributed corpora for coreference resolution task in Russian language? What type of data can be obtained from Semantic Hub?

3. Genre adaptation
   Should the model be tuned to conversational, medical or any other field of genre?

Regarding the choice of the model architecture for coreference resolution task there are plenty of great state-of-the-art solutions available at the moment. Having read a large amount of publications about coreference resolution with deep learning, we selected some of the most recent and high-scoring systems created by reputable authors accordingly (Clark and Manning, 2016; Joshi et al., 2019a; Lee et al. 2017; Joshi et al., 2019b) in the table below.

		MUC			B3			CEAFφ4			Avr F1
Model	year	prec	rec	f1	prec	rec	f1	prec	rec	f1	
Reinforcement learning. Average.	2016	79.63	70.1	74.56	69.95	57.61	63.17	63.56	54.65	58.77	65.5
SpanBERT	**2020**	**85.8**	**84.8**	**85.3**	**78.3**	**77.9**	**78.1**	**76.4**	**74.2**	**75.3**	**79.6**
End-To-End	2017	81.2	73.6	77.2	72.3	61.7	66.6	65.2	60.2	62.6	68.8
BERT-large + c2f-coref (independent)	2019	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9

Table 4. Overview of SOTA models.

As can be observed, the SpanBERT (Joshi et al. 2019) presented by Facebook currently resolves coreference better than the other models on CoNLL 2012 English Test dataset (Pradhan et al., 2012). Other models, including those, which are not present in the table, will be described further in the work.

Considering the data, already existing sets should be logically mentioned firstly. There are two different freely available datasets for coreference resolution in Russian we found: RuCor (Toldova et al., 2014) dataset, and Dialog-2019 (Budnikov and Toldova et al., 2019) datasets. Both are made by random sampling from OpenCorpora and consist of 181 documents for RuCor, and 395 train with 127 test texts of various genre for Dialog-2019. While Dialog-2019 set consists of four different markup files for every text, RuCor markup is presented in CoNLL-like format. The latter option seems to be more desirable, since (Joshi et al. 2019) implicitly provides a script (Joshi and Kummerfeld, 2019) for converting ConLL to the SpanBERT accepted **jsonlines** format. Still, the RuCor is not fully CoNLL, so it needs to be preprocessed.

For the needs of the present research Semantic Hub produced a dataset for coreference resolution. It is based on the the domain-driven features and specifics of previously described anaphora types, resolution of which was selected as the main task of this paper. Currently it has been marked up only for coreference resolution on a message-level. The data is presented in json format and provides:

- a span for the mention
- the mention string for defined span from the text
- an id for every coreference chain
- antecedent or anaphor parameter

This should be the first system based on transformer architecture and created for such a genre and type of Russian texts, so a variety of adaptation problems are expected. Therefore we expect to fine-tune the model on some big chunks of Russian texts and dialog information similar to what DeepPavlov have done for their ConversationalRuBert (Neural Networks and Deep Learning lab, MIPT, 2018) before fine-tuning on RuCor or other tagged corpora.

7. Data and format

Today a lot of research on coreference resolution topic has been done and some of the results seem promising, but most of the proposed top-performing systems are still mentioned for use with English language. Moreover, the amount of datasets available is limited to such extent that at the present moment the researchers have only 9200 coreferential chains consisting of 41500 tokens for Russian language if summed up. The OntoNotes (Weischedel and Ralph et al., 2013) corpus, which serves as the source for most of the works on automated coreference resolution for English consists of 2.9 million words with 1.5 million of English entries. Compared to this, Russian data is 36 times smaller.

Semantic Hub provided a dataset of 386 254 documents. Each document represents a json with two main parts: «annotations» and «text». «Annotations» is a list of entities tagged as participating in some coreference relations, the features of each entity was described in the previous chapter. The «text» part represents the raw text of a message, usually with some xml leftovers. Every single message can contain more than one sentence. Some messages do not have any coreferential relations in them, but are still present. Most of the tagged referential entities are represented by one word, usually a noun or a pronoun. The dataset consists of 1 693 773 coreferential chains and there are as much as 2 968 033 tokens present in these chains. Empirical knowledge shows that usually a corpora of such a size is sufficient for achieving adequate results on the evaluation. And continued training on the available data is consistently effective as proved in (Xia and Van Durme, 2021). Which means that a model tuned on OntoNotes can be successfully adapted to another domain having a large enough dataset for the adaptation The mark-up of the coreference chains was done by the linguistic processor of Semantic Hub. Some infographics and facts about the produced corpus can be found in Appendix (1-5).

Another concern regarding the data is the format. The way the data is stored affects its usability and greatly affects the effectiveness of the system. In the case of the present paper it is easy to say that CoNLL-12 is close to be the prefect format for the task. To better understand the benefits of using CoNLL-12 for coreference resolutions it is best to see what types of information in can contain.

The format is presented as a set of columns. Each column represents either a linear sentence annotation, part-of-speech for example, or there are multiple columns which are taken together and synced with another one to represent the roles of other words regarding the selected one. Possible contents the CoNLL-12 columns can be seen in the table below.

Column	Type	Description
1	Document ID	This is a variation on the document filename
2	Part number	Some files are divided into multiple parts.
3	Word number	
4	Word itself	This is the token as met in the text
5	Part-of-Speech	
6	Parse bit	This is the bracketed structure broken before the first open parenthesis in the parse, and the word/part-of-speech leaf replaced with a *. Used when word itself is represented by a TreeBank entity. Mostly OntoNotes feature.
7	Predicate lemma	The predicate lemma is mentioned for the rows for which we have semantic role information. All other rows are marked with a "-"
8	Predicate Frameset ID	This is the PropBank frameset ID of the predicate in Column 7.
9	Word sense	This is the word sense of the word in Column 3.
10	Author	This is the author name, can be useful for future adaptations.
11	Named Entities	These columns identifies the spans representing various named entities.
12	Predicate Arguments	There is one column each of predicate argument structure information for the predicate mentioned in Column 7.
13	Coreference	Coreference chain information encoded in a parenthesis structure.

Table 5, CoNLL-12 format description

The basics of transformers and their behaviour when taught on large quantities of data combined with such a versatile dataset is sure to produce some great results. This is due to the fact that CoNLL implicitly allows to take into account most types of anaphoric constraints, while BERT-based models are known to catch and distinguish the syntactic and semantic habits of different entities by itself. Moreover, CoNLL-12 is accepted by most of the coreference resolution systems, since the largest tagged dataset for coreference - OntoNotes, is presented in this format. Everything mentioned means that creating a script which transfers Semantic Hub json-type data to CoNLL-12 format can serve as a universal foundation for every coreference resolution project of the company in the near future.

Worth mentioning is that RuCor is known (Sysoev et al., 2017) to have mistakes which can be fixed only manually and are not to be addressed by the corpus creators in the near future. Two major problems of the corpus are duplicated mentions and cyclic chains which are caused by incautious merging of the tagged versions from different annotators. So, wishing to utilise the corpus, we should obviously read through it manually. This is considered really time consuming, but using such a set of data without preparation can lead to inexplicable mistakes and that is unwanted. Finding some way to automatically filter the best texts in the corpus could be a good solution for the problem.

Also there is a problem of formatting. The data is stored in different formats depending on the corpora so there is a need to create a parser and a converter for the data to be acceptable for the BERT system. Fortunately, NeuralCoref (HuggingFace, 2017) implicitly makes use of some **rucor2conll.py** script to convert RuCor to CoNLL formatting. Nevertheless, the script itself comes out to be outdated and contains a lot of coding errors. A similar situation is observed with SpanBERT script for CoNLL adaptation. Launching the stock SpanBERT script for transferring RuCor in CoNLL to jsonlines causes it to terminate prematurely due to the program error. Coping with such a problem is a difficult task, which is, unfortunately, not solvable by editing the existing script because of the hardcoded dependencies. Concluding everything said before, the necessity of a script for data transformation and preprocessing becomes obvious. Unfortunately, the available data differs drastically. This leads to the fact that every single data source of Russian coreference-tagged texts needs an individual program written for the data modification.

8. Models

«In NLP all they talk about is BERT», – while being a joke that is close to be the actual truth: most of the current SOTA models are implementing the transformer architecture and, as we see it for today, half of the transformer-based models are BERTs. The dominant position of transformer-based models, especially BERT-based, can be observed on the coreference related tasks. The models reviewed in the table above are produced by reputable groups of authors as the result of colossal work with vast datasets, powerful GPUs. Regarding this proposal, we think, it is logical to give a detailed overview for the models which are planned to be implemented directly. Other systems for coreference resolution will be briefly presented in the main work for comparison purposes.

The model presented in (Joshi et al. 2019) – the SpanBERT performs 79.6% F1 on CoNLL-2012 shared task (OntoNotes) and exceeds the previous top model by 6.6% absolute.The performance of the model is still outstanding, though it was released in November of 2019.

Hereafter we present a description of SpanBERT main features:

Data

- Pretraining of BERT$_{large}$ was done on 800M words from BooksCorpus and 2,500M words from English Wikipedia using cased Wordpiece tokens (playing → play, ##ing)
- Fine-tuning was performed on multiple datasets using standard BERT cased encoder by HuggingFace.

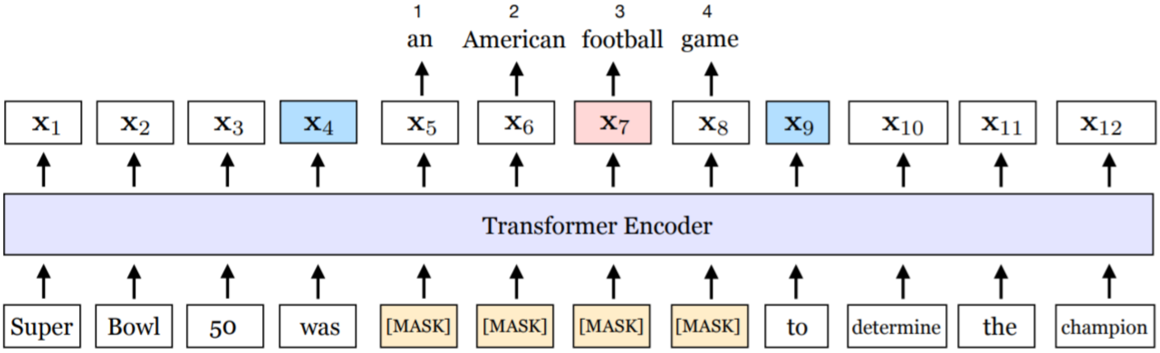
Method

- Random spans of tokens are masked, instead of full random tokens.
- New auxiliary objective - Span boundary objective (SBO), which tries to predict the entire masked span using only the representations of the tokens at the span's boundary.
- Pre-training is performed on single segments for each example instead of two half-length segments as BERT does for the next sentence prediction (NSP) objective. Doing so considerably improves performance on most downstream tasks, as the result of the abandonment of the NSP subtask in SpanBERT.

Span Masking

The model is sampling spans of text until the masking budget – 15%, has been spent:

1. At each iteration, sample a span length, which is number of words, from a geometric distribution *L ~ Geo(p)*, which is skewed in such a way that shorter spans are more probable.
2. Uniformly randomly select the starting point for the span to be masked, which must be at the beginning of a word.
3. Sample a sequence of complete words instead of usual subword tokens.



Img. 1, Predicting x_5-x_8 span by x_4 and x_9

Span Boundary Objective

Span boundary objective (SBO) – is a task of making a prediction for each token of a masked span using only the representations of the observed tokens at the boundaries.

Overview of this objective:

1. The model output is a function of x_{s-1} – token to the left of the span; x_{e+1} – token to the right of the span; p_{i-s+1} – embedded position of the target with respect to the start of the span.
2. SpanBERT representation function is a 2-layer feed-forward network with GeLU activations and layer normalisation.
3. Vector representation y_i is utilised to predict the token x_i and compute the cross-entropy loss exactly like the in the masked language modelling (MLM) objective.
4. SpanBERT sums the loss from both the *span boundary* and the *regular masked language model* objectives for each token x_i in the masked span (x_s , ..., x_e), while reusing the input embedding for the target tokens in both MLM and SBO:

$$\begin{aligned} \mathcal{L}(x_i) &= \mathcal{L}_{\text{MLM}}(x_i) + \mathcal{L}_{\text{SBO}}(x_i) \\ &= -\log P\left(x_i \mid \mathbf{x}_i\right) - \log P\left(x_i \mid \mathbf{y}_i\right) \end{aligned}$$

Since it is impossible for SpanBERT to make the actual understandable predictions by itself a need of an additional module arises. For the purposes of this work it is convenient to apply *coarse to fine coreference model (c2f-coref)* presented in (Lee et al. 2018). Moreover, c2f-coref was used in the (Joshi et al. 2019) to make predictions, but replacing LSTM-based encoder inside with the BERT transformer.

Overview of c2f-coref model:

For each mention span x, the model learns a distribution *P(·)* over possible antecedent spans *Y*.

$$P(y) = \frac{e^{s(x,y)}}{\sum_{y' \in Y} e^{s(x,y')}}$$

Where *s(x,y)* is the scoring function:

$$\begin{aligned} s(x, y) &= s_m(x) + s_m(y) + s_c(x, y) \\ s_m(x) &= \text{FFNN}_m(\mathbf{g_x}) \\ s_c(x, y) &= \text{FFNN}_c(\mathbf{g_x}, \mathbf{g_y}, \phi(x, y)) \end{aligned}$$

- $s_m(x)$ represents how likely is it for the span *x* to be a mention;
- $s_m(y)$ is the same, but for *y*;
- $s_c(x,y)$ represents how likely is that *x* and *y* refer to the same entity assuming they are both mentions;
- $\boldsymbol{g_x}$ and $\boldsymbol{g_y}$ are the span representations of first and last tokens concatenated with the attention vector calculated over the output representations of the token in the span;
- *FFNN(·)* stands for feedforward neural network;
- *φ(x, y)* represents speaker and genre metadata features.

9. Methods

In the present work the implemented pipeline of coding was like this:

1. Prepare the data from RuCor and Semantic Hub to make it ready for SpanBERT.
    1. Preprocessing of the texts.
    2. Transformation from Semantic Hub json to CoNLL-12.
2. Fine-tune SpanBERT + c2f-coref models on the preprocessed texts.
3. Obtain the scores for the test data

A significant amount of coding was related to the first stage. Having the negative experience of applying BERT-models to heavily-processed texts (Yatsishin et al., 2020), texts in this work were preprocessed very lightly, without any stemming or lemmatisation. Any leftover xml from the Semantic Hub processor was eliminated from the message text. Multiple whitespaces of every kind were removed same as special «smileys» represented by text between colons (:smile:). The messages without a finished sentence had a dot symbol appended to the last word.

The script for conversion of Semantic Hub json to CoNLL-12 format is based on the **pandas dataframe** architecture thus allowing easy modification and addition of the data. Also it allows to add custom columns to the desired CoNLL-type document which can be really convenient when adapting other SOTA models for a spectre of different tasks in the future. The part of speech is being determined with nltk module (Bird et al., 2009). While processing the data it came out that for the script to process such vast amounts – 400k messages, more processing power than available is obligatory. At the moment the program has been working for over 150 hours filling the columns 1, 2, 3, 4, 5 and 13, as the most important ones for coreference resolution task. The process is launched on «1.6 GHz 2-core Intel Core i5 with 8 Gb, 2133 MHz, LPDDR3 RAM» and «Intel UHD Graphics 617 1536 Mb» which are available. Unfortunately, it is impossible to obtain more GPU power, hence the only thing possible is launching a parallel Google Colab notebook and await for the process to finish. At the moment of the writing the preprocessed was stored in a dataframe of 1,3+ Gb consisting of 11 932 678 tokens out of 64 851 795, meaning that there are 52 919 117 more tokens to work on. The data was split for train, test and development parts in 95 : 2.5 : 2.5 ratio referring to the amount of the documents. Such a ratio has empirically proved to be optimal for such amounts of information.

For the second step from the above pipeline Tensorflow realisations of the described modules were implemented. While utilising the scripts for the SpanBERT implementation some problems were faced related to the correctness of the existing snippets. Additionally, large amount of requirements and dependancies for the system are heavily outdated as for the day, so some of them were deleted or safely updated, if possible. After that the data was processed by a BERT tokeniser and converted to the desirable jsonlines format. Next the model was tuned on the training documents in the Google Colab environment. Having a very limited amount of processing power the fine-tuning was done with a reduced configuration compared to the original SpanBERT models. The data was accessible for the model in batches of 20 000 documents. The biggest challenges met at this step of the pipeline were changing the snippets based on the needs of the present paper and preparing the launch of the model with the optimal configurations to fit in the GPU available and present decent results at the same time.

The limitations set by the amount of processing power affected the results greatly, since it was impossible to implement several practices known to enhance the results. Firstly and mainly, the model trained was the **spanbert-base** inside of a **bert-cased-base** so to say, since, as mentioned, SpanBERT architecture is a modification for **bert-cased-base** tasks and masking. The vocabulary of the **bert-cased-base** is actually very limited considering Russian language, since it was trained on English Wikipedia. There are still all the letters present as a single token for every letter in Russian except for the capital «Ë». The presence of Cyrillic symbols can be explained by a limited number of them encountered in English Wikipedia topics. Without having enough power to set the SpanBERT weights to zero and retrain them side to side with **rubert-base-cased** (Kuratov and Arkhipov, 2019) the model is faced with many challenges in the tuning process and some of them are impossible to overcome for such a system. The metric for the scoring was the CEAFm metric (Luo, 2005) which is used for identification of how similar are the guessed entities with the original ones. The final average high-score combination on the document level achieved after training on the whole train data was:

Average F_1: 30.16% on 1843 docs
Average precision: 65.14%
Average recall: 20.58%

Let's take a look at some of the interesting clusters predicted by the model for a single message. The text of the examples cannot be shown being under the NDA with Semantic Hub.

(A) ['нас', 'нам', 'нам', 'мы'],
['меня', 'меня']

(B) ['нас', 'нас', 'нам'],
['дочь', 'дочь', 'дочь'],
['мне', 'меня', 'меня']

(C) ['мне', 'меня', 'мне', 'мне', 'меня'],
['сын', 'нас', 'нас', 'сына', 'сына', 'нам'],
['дочь', 'нее']

(D) ['Мне', 'мне', 'меня', 'меня', 'Нне'],
['Мажу', 'мужа']

One could expect all the predicted words to rarely be longer than four letters since the model was trained on the English data. It is unusual that a word consists of more than four tokens, which can be easily observed in the **bert-cased-base** vocabulary. As for the Russian the only tokens available are the letters, so the words predicted within the same algorithm are really short. The same premises are behind the 'typos' that can be clearly noticed in *(D).* Modern English is known to have mostly analytic marking pattern with fairly little inflection in contrast to Russian being a synthetic language with nouns possible to inflect for six cases. All that leads to the point that having the words created from letters where each letter is considered a meaningful part of the word is not the best way to represent the Russian language.

Still, as can be seen, the model obtained the knowledge needed and is working within the confines of the anaphoric constraints described previously. The model can detect different numbers and cases *(A).* It can as well agree the referents on the person parameter, which can be noticed in the *(C),* where the model captures the usual parental «we», which is specific for the medical domain.

Moreover, the model easily diversifies the cases were «we» is used in a different context as in *(B).* Also is can be observed in *(C)* that the gender for the nouns and pronouns is captured well: the words for «son» and «daughter» are in different clusters, but together with the right pronoun.

The achieved results are actually pretty good, having in mind the amount of processing power available. As explained above, despite working on English-based principles and vocabulary the system is capable of capturing the information needed for reference resolution and it resolves greatly on short words, especially on pronouns. The whole pipeline is actually working and is suitable for resolving coreference in Russian. It is obvious that much better results could be achieved if the tokeniser of the system was replaced with a cased multilingual or Russian one. Filling all the columns of the CoNLL format can also provide a great boost for the model performance. More training epochs on the whole dataset will surely provide greater results. Much more about the produced files, the code, problems with it and its algorithms can be found on GitHub[1].

10. Future work

The most important system upgrade possible is definitely the adaptation of the presented pipeline to work with a **rubert** model. Apart from standard methods of score enhancement in machine learning there are plenty of things that can be modified to make the obtained model more suitable for the task. First of all, it is important to fill as many columns in the CoNLL-12 format of the data as possible. It is also possible to try launching the model on the dataset where only entities related to medical facts are tagged by the Semantic Hub processor, as this can turn out to be a big change for the type of entities recognised in the future. For a multi language BERT realisation mapping the part of speech tags to the universal tagging scheme possibly also can make the system give better predictions, as all the pos-tags between languages are made in a unified manner. Obviously, some rule-based preprocessing should be added to correctly process the abbreviations or special medical information like drug dosage, the results of medical analysis or drug names. In addition, the whole dataset should be filtered to get out the entries without coreference and to deal with the exceptions inside the data. Exceptions can be observed as sparks on the heatmap in Appendix (5).

11. Conclusions

This work is aimed at adapting a system created for different domain and language to the reality of medical forums in Russian. The code of the system was totally adapted and at the moment is suitable for easy use in the format of two python notebooks. There are a lot of ways to make it work better, and the current experiments show that many of the major parameters for coreference resolution are well-captured. Having only the processing power as the main obstacle the system is definitely capable of producing a result twice as good as it is at the moment.

As essential components of the system construction process there were a lot of intermediate steps: analysis of specific types of anaphoric relations on forum data, comparing different SOTA models

[1] https://github.com/toskn/thesis

for coreference resolution, critical analysis of existing corpora with the data gathered inside and many more. Besides the analysis and explanations the paper presents an overview of anaphoric constraints and relations types meaningful for the task of coreference resolution, a universal script for data conversion from Semantic Hub json to every CoNLL-like format and jsonlines format after that. The SpanBERT and coarse to fine coreference modules are prepared and are finally ready to be used on the Russian data.

12. List of references

1. Bird et al., 2009 — Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.
2. Budnikov and Toldova et al., 2019 — Budnikov, E. A., Toldova, S. Y., Zvereva, D. S., Maximova, D. M., & Ionov, M. I. (2019). *RU-EVAL-2019: Evaluating Anaphora and Coreference Resolution for Russian*. http://www.dialog-21.ru/media/4689/budnikovzverevamaximova2019evaluatinganaphoracoreferenceresolution.pdf
3. Cambria, 2016 — Cambria E. (2016) Affective computing and sentiment analysis. IEEE Intelligent Systems 31(2): 102 -107
4. Carbonell and Brown, 1988 — Carbonell J.G., Brown R.D. (1988) Anaphora resolution: a multi-strategy approach. In: Proceedings of the 12th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp 96–101
5. Castagnola, 2002 — Castagnola L. (2002) Anaphora resolution for question answering. PhD thesis, Massachusetts Institute of Technology
6. Clark and Manning, 2016a — Clark K., Manning C.D. (2016a) Deep reinforcement learning for mention-ranking coreference models. arXiv preprint arXiv:160908667
7. Clark and Manning, 2016b — Clark K., Manning C.D. (2016b) Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:160601323
8. Heim, 1982 — Heim I. (1982) The semantics of definite and indefinite nps. University of Massachusetts at Amherst dissertation
9. HuggingFace, 2017 — HuggingFace. (2017). NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks. Github. https://github.com/huggingface/neuralcoref
10. Joshi et al. 2019a — Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2019, November). *SpanBERT: Improving pre-training by representing and predicting spans*. https://arxiv.org/pdf/1907.10529.pdf
11. Joshi et al. 2019b — Joshi, M., et al. «BERT for Coreference Resolution: Baselines and Analysis». *arXiv:1908.09091 [cs]*, December 2019. *arXiv.org*, http://arxiv.org/abs/1908.09091.
12. Joshi and Kummerfeld, 2019 — Joshi, M., & Kummerfeld, J. (2019, November). *BERT and SpanBERT for coreference resolution*. Github. https://github.com/mandarjoshi90/coref
13. Kennedy and Boguraev, 1996 — Kennedy C., Boguraev B. (1996) Anaphora for everyone: pronominal anaphora resoluation without a parser. In: Proceedings of the 16th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp 113–118
14. Kuratov and Arkhipov, 2019 — Kuratov, Y., Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.
15. Lappin and Leass, 1994 — Lappin S., Leass H. J. (1994) An algorithm for pronominal anaphora resolution. Computational linguistics 20(4):535–561

16. Lee et al. 2018 — Lee, K., He, L., & Zettlemoyer, L. (2018). *Higher-order Coreference Resolution with Coarse-to-fine Inference*. Association for Computational Linguistics. https://www.aclweb.org/anthology/N18-2108.pdf

17. Luo, 2005 — X. Luo , On coreference resolution performance metrics, in: Proceedings of the Con- ference on Human Language Technology and Empirical Methods in Natural Lan- guage Processing, Association for Computational Linguistics, 2005, pp. 25–32 .

18. Mitkov, 2014 — Mitkov R. (2014) Anaphora resolution. Routledge

19. Neural Networks and Deep Learning lab, MIPT, 2018 — Neural Networks and Deep Learning lab, MIPT. (2018). *BERT in DeepPavlov*. DeepPavlov. http://docs.deeppavlov.ai/en/master/features/models/bert.html

20. Pradhan et al., 2012 — Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, July). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. Association for Computational Linguistics.

21. Preuss, 1992 — Preuss S. (1992) Anaphora resolution in machine translation. TU, Fachbereich 20, Projektgruppe KIT

22. Steinberger et al., 2007 — Steinberger J., Poesio M., Kabadjov M. A., Jezek K. (2007) Two uses of anaphora resolution in summarization. Information Processing & Management 43(6):1663 - 1680

23. Sysoev et al., 2017 — Sysoev, A. A., Andrianov, I. A., & Khadzhiiskaia, A. Y. (2017, June). Coreference Resolution in Russian: State-of-the-art Approaches Application and Evolvementhttp://www.dialog-21.ru/media/3954/sysoevaaetal.pdf. International Conference "Dialogue 2017." http://www.dialog-21.ru/media/3954/sysoevaaetal.pdf

24. Toldova et al., 2014 — Toldova, S. J., Roytberg, A., Ladygina, A. A., Vasilyeva, M. D., Azerkovich, I. L., Kurzukov, M., Sim, G., Gorshkov, D. V., Ivanova, A., Nedoluzhko, A., & Grishina, Y. (2014). *RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian*. http://www.dialog-21.ru/digests/dialog2014/materials/pdf/ToldovaSJu.pdf

25. Weischedel and Ralph et al., 2013 — Weischedel, Ralph, et al. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

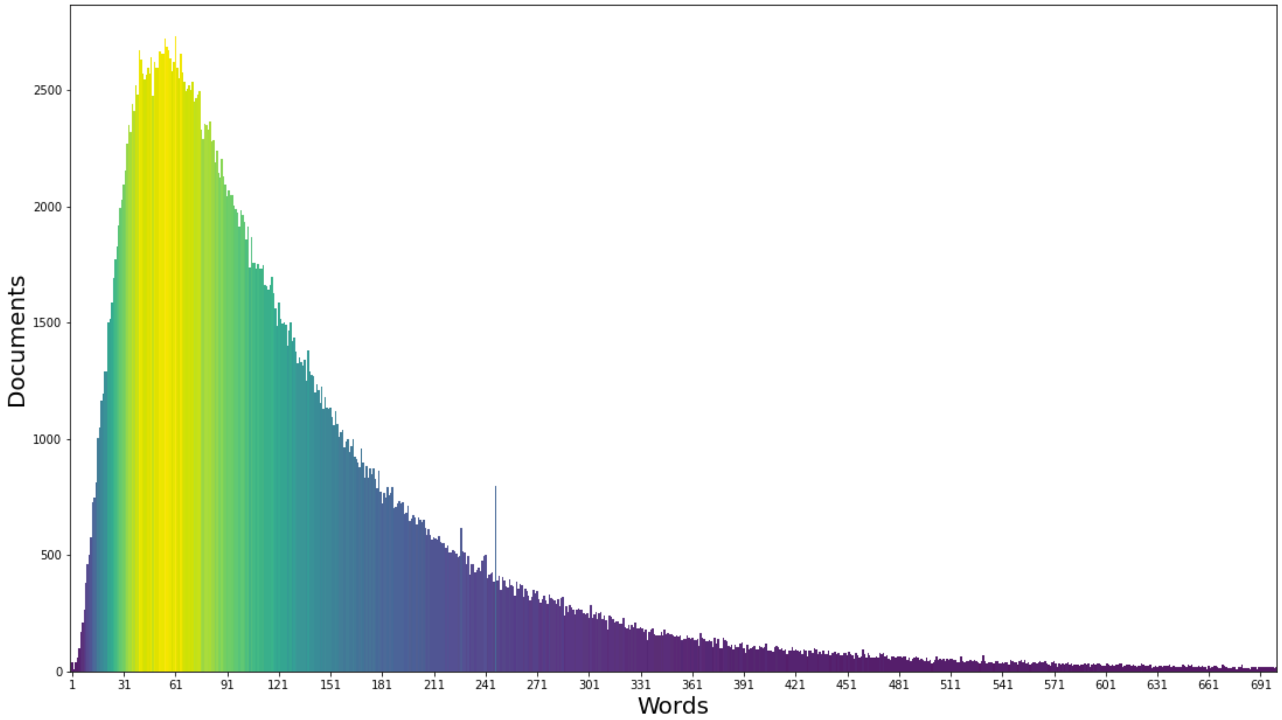
26. Wiseman et al, 2016 — Wiseman S., Rush A. M., Shieber S. M. (2016) Learning global features for coreference resolution. arXiv preprint arXiv:160403035

27. Xia and Van Durme, 2021 — Xia, Patrick, и Benjamin Van Durme. «Moving on from OntoNotes: Coreference Resolution Model Transfer». *arXiv:2104.08457 [cs]*, April 2021. *arXiv.org*, http://arxiv.org/abs/2104.08457.

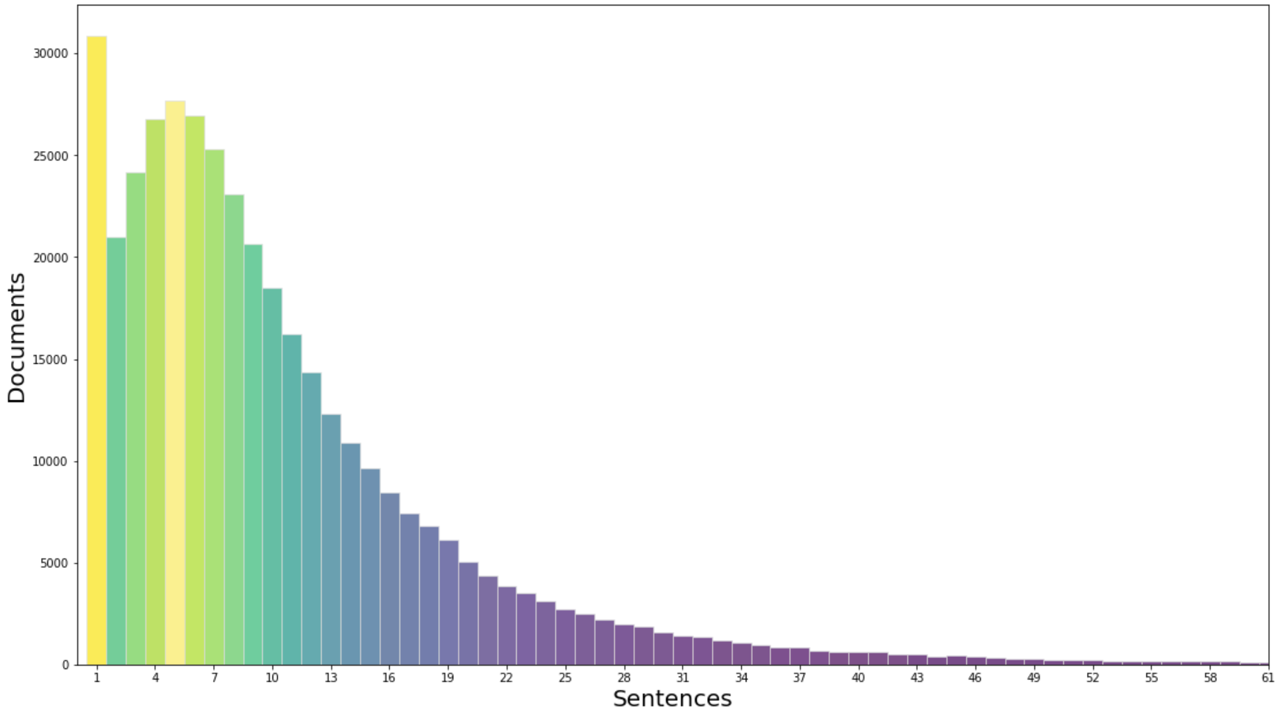
28. Young et al, 2018b — Young T., Hazarika D., Poria S., Cambria E. (2018b) Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine 13(3)

29. Yatsishin et al., 2020 — Gusev A., Kuznetsova A., Polyanskaya A., Yatsishin E., «BERT Implementation for Detecting Adverse Drug Effects Mentions in Russian». *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, Association for Computational Linguistics, 2020, cc. 46–50. *ACLWeb*, https://www.aclweb.org/anthology/2020.smm4h-1.7

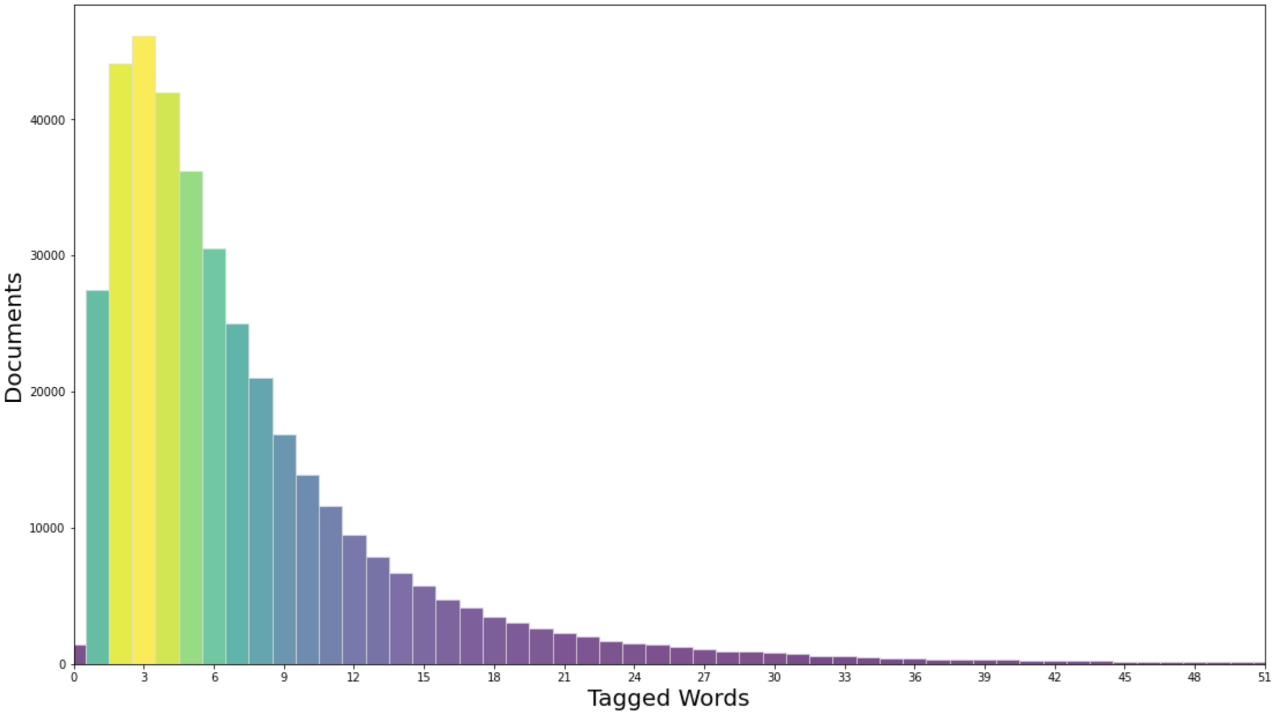
13. Appendix



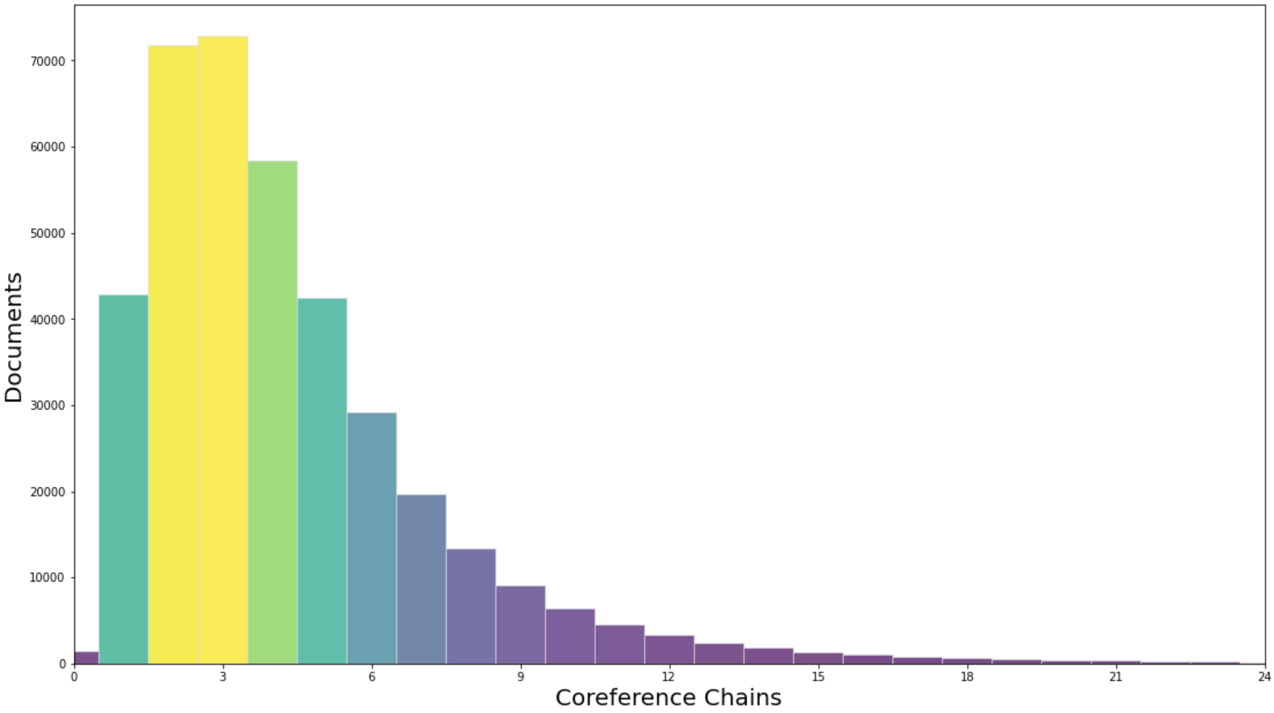
Appendix 1. The major part of the dataset are messages which contain from 40 to 80 words in their text body.



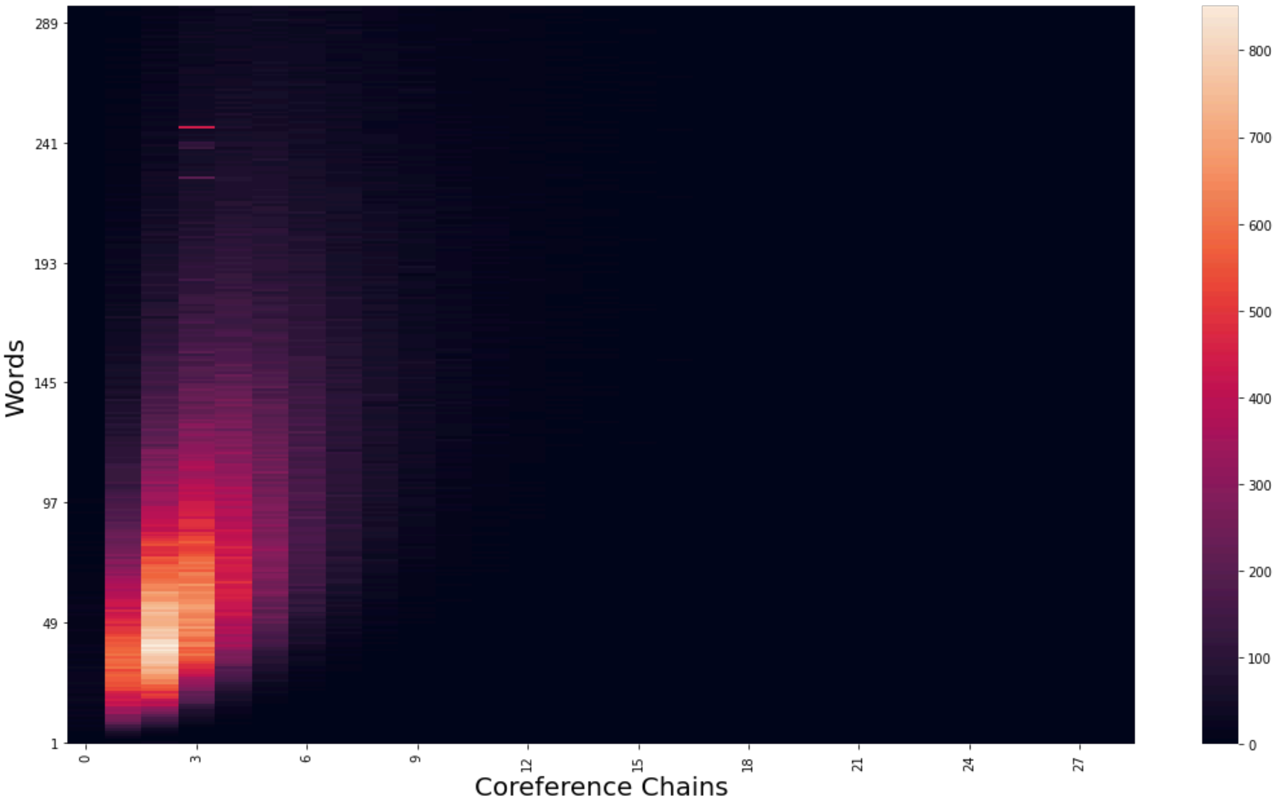
Appendix 2. Most of the messages contain less than 10 sentences in them.



Appendix 3. More than half of the messages contain from 2 to 5 words tagged for coreference.



Appendix 4. Roughly half of the documents consist of 2 to 3 coreference chains per message.



Appendix 5. Relation between words amount and coreference chains amount in a single message. The most common situation is a message consisting of 30 to 40 words with 2 coreference chains present.