AGRR-2019: A Corpus for Gapping Resolution in Russian

Maria Ponomareva[♠] **Kira Droganova**[†] **Ivan Smurov**[♠,♣] **Tatiana Shavrina**[◇,♡]
[♠]ABBYY, Moscow, Russia
[†]Charles University, Faculty of Mathematics and Physics
[♣]Moscow Institute of Physics and Technology, Moscow
[◇]Sberbank, Moscow, Russia
[♡]National Research University Higher School of Economics, Moscow
{maria.ponomareva,ivan.smurov}@abbyy.com
droganova@ufal.mff.cuni.cz
Shavrina.T.O@sberbank.ru

Abstract

This paper provides a comprehensive overview of the gapping dataset for Russian that consists of 7.5k sentences with gapping (as well as 15k relevant negative sentences) and comprises data from various genres: news, fiction, social media and technical texts. The dataset was prepared for the Automatic Gapping Resolution Shared Task for Russian (AGRR-2019) - a competition aimed at stimulating the development of NLP tools and methods for processing of ellipsis.

In this paper, we pay special attention to the gapping resolution methods that were introduced within the shared task as well as an alternative test set that illustrates that our corpus is a diverse and representative subset of Russian language gapping sufficient for effective utilization of machine learning techniques.

1 Introduction

During the last two years gapping (i.e., the omission of a repeated predicate which can be understood from context (Ross, 1970)) has received considerable attention in NLP works, both dedicated to parsing (Schuster et al., 2018; Kummerfeld and Klein, 2017) and to corpora enhancement and enrichment (Nivre et al., 2018; Droganova et al., 2018). At the same time, just a few works dealt with compiling a corpus that would represent different types of ellipsis, and almost exclusively for English. Most of these works address VP-ellipsis, which refers to the omission of a verb phrase whose meaning can be reconstructed from the context (Johnson, 2001), for instance, in “Mary loves flowers. John does too” (Hardt, 1997; Nielsen, 2005; Bos and Spenader, 2011). The research has mainly been conducted so far on rather small amounts of data, not exceeding several hundreds of sentences. In this work we aim to create a resource with a decent amount of data that would include a broad variety of genres and would rely minimally on any specific NLP frameworks and parsing systems.

This work consists of four parts. First, we describe the dataset, its features, and provide examples of Russian-specific constructions with gapping. Second, we describe an alternative test set that we have prepared to demonstrate that our corpus is representative enough. Then we briefly describe the key metrics that have been proposed to evaluate the quality of gapping resolution methods within the shared task. Finally, we provide a detailed analysis of the methods that have successfully solved the gapping resolution task as well as the results that were achieved on the alternative test.

2 Gapping

We confine ourselves to the types of elliptical constructions for Russian that involve omission of a verb, a verb phrase or a full clause.

In this work we use the following terminology for gapping elements. We call the pronounced elements of the gapped clause remnants. Parallel elements found in a full clause that are similar to remnants both semantically and syntactically are called remnant correlates. The missing material is called the gap (Coppock, 2001).

Traditionally, gapping is defined as the omission of a repeating predicate in non-initial composed and subordinate clauses where both remnants to the left and to the right remain expressed.

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 35–43,
Florence, Italy, 2 August 2019. ©2019 Association for Computational Linguistics

(1) Я принял её за итальянку, а его за
I mistook her for Italian and him for
шведа.
Swede

‘I mistook her for Italian and ~~I mistook~~ him for Swede’

However, a broader interpretation is possible (Testelets, 2011). Some features of gapping worth mentioning are listed below.

Elements remaining after predicate omission can be of different types. Consider the following examples where remnants are predicates (2), preposition phrases (3), adverbs (4), adjectives (5) potentially with their dependents.

(2) Одно может вдохновлять, а другое
one can inspire and other
вгонять в тоску.
put in melancholy

‘One thing can inspire and the other ~~can~~ put you in a melancholic mood.’

(3) Советую вам поменьше думать о
recommend you less think about
проблемах, и побольше — об их
problems and more - about their
решении.
solution

‘I recommend you to think less about problems, and ~~think~~ more about their solutions.’

(4) Вначале они играли интересно, потом
at.first they played interesting.ADV after
– прескучно.
- boring.ADV.INT

‘At first they played interestingly, then ~~they played~~ extremely dully.’

(5) Сердце ее было слишком чистым, чувства
heart her was too pure feelings
слишком искренними.
too sincere

‘Her heart was too pure and her feelings ~~were~~ too sincere.’

The set of constructions for Russian that implement stripping (Merchant, 2016) seems to be broader than for English and the difference between gapping and stripping in Russian is less clear. We encountered a wide variety of examples that go beyond the canonical examples. Examples (6) and (7) illustrate the cases when arguments/adjuncts of the elided verb do not fully correspond to the arguments/adjuncts of the pronounced verb, thus some of them (*в конце* ‘in the end’ in (6), *за 2009 год* ‘during year 2009’) do not have correlates. We consider such examples gapping with one remnant and include them in the corpus.

(6) Добавляем муку, крахмал и
add flour starch and
разрыхлитель, а **в конце** сметану.
baking.powder and in end sour.cream

‘We add flour, starch and baking powder, and at the end ~~we add~~ sour cream.’

(7) Рост цен составил 11,9 процента
growth prices amounted.to 11.9 percent
(**за 2009 год** - 4,4 процента)
in 2009 year - 4.4 percent

‘Price growth amounted to 11.9 percent (in 2009 ~~it amounted to~~ 4.4 percent)’

3 Corpus Description

Since the publicly available markup with gapping is sparse, one of our key motivations was to create a corpus that contains as many examples of gapping as possible. To the best of our knowledge, no other publicly available dataset contains a comparable amount of gapping examples.

With that in mind, we decided to base our corpus on the markup obtained with Compreno (Anisimovich et al., 2012). Compreno is a syntactic and semantic parser that contains a module for predicting null elements in the syntactic structure of a sentence. An overview of the module can be found in (Bogdanov, 2012).

While cleaning up the output of a specific system allows us to obtain markup much faster than annotating from scratch, training on the resulting corpus may yield systems that would reproduce the original system’s output instead of properly modeling the real-world natural language phenomenon. We took this risk because even if the corpus we have created contains Compreno bias, the selection is representative enough. Moreover, in order to further test for the presence of such bias, we evaluated the top systems of the shared task on an alternative test set that was created from SynTagRus (see Section 4).

The corpus is available on the shared task’s GitHub [1].

3.1 Annotation Scheme

We utilize the following labels for fully annotated sentences with gapping:

- The gap is labeled V.
- The head of the pronounced predicate corresponding to the elided predicate is labeled cV.

[1] https://github.com/dialogue-evaluation/AGRR-2019

- Remnants and their correlates are labeled Rn and cRn respectively, where n is the pair's index

For gapping annotation we use square brackets to mark all gapping elements (whole NP, VP, PP etc. for remnants and their correlates and the predicate controlling the gap), the gap is marked with *[$_V$]*. Example (8) shows an example of bracket annotation of (1).

(8) Я [$_{\text{cV}}$ принял] [$_{\text{cR1}}$ её] [$_{\text{cR2}}$ за итальянку],
I mistook her for Italian
а [$_{\text{R1}}$ его] [$_{\text{R2}}$ за шведа].
and him for Swede
'I mistook her for Italian and ~~I mistook~~ him for Swede'

Therefore, the full list of annotation labels is as follows: cV, $cR1$, $cR2$, V, $R1$, $R2$.

3.2 Obtaining the Data

In this section we provide a detailed description of the process of compiling the corpus. The bulk of the collection comprises Russian texts of various genres: news, fiction, technical texts. To our understanding, many NLP tasks that could benefit from gapping resolution are often applied to social network data. Therefore, we balanced the corpus by adding texts from the popular Russian social network VKontakte. They make up a quater of the collection.

First, all texts in the text collection were parsed with Compreno. We identified the sentences in which gapping was predicted. Using the Compreno parser, we generated bracketed annotation for each sentence (in which every gapping element X has an opening bracket $[X$ and closing bracket $]$).

Mindful of our main goal (i.e., to maximize the amount of data in the corpus), we decided to avoid fixing the annotation errors manually. Instead 11 assessors were asked to evaluate the annotation, assigning one of four classes:

0 no gapping, no markup is needed;

1 all gapping elements are annotated correctly;

2 some gapping elements are annotated incorrectly;

3 problematic example.

Each sentence was evaluated by two assessors. Table 1 shows that 41% out of 17411 sentences have correct annotation and 19% were erroneously attributed to the examples with gapping, according to both annotators.

	0	**1**	**2**	**3**
0	**3350 (19%)**	370 (2.1%)	303 (1.7%)	254 (1.5%)
1	394(2.3%)	**7201(41%)**	1163 (6.7%)	283 (1.6%)
2	288 (1.7%)	581 (3.3%)	1960 (11%)	302 (1.7 %)
3	446 (2.5 %)	230 (1.3%)	153 (0.9%)	133 (0.8 %)

Table 1: Assessment analysis for the AGRR corpus; 0, 1, 2, 3 - annotation classes.

The main application of our corpus is in machine learning, therefore the corpus has to include negative examples (i.e., sentences without gapping). We considered two types of negative examples to select more relevant sentences. The first type comprises problematic negative sentences on which the Compreno parser false positively predicted gapping (labeled 0 by both assessors). Introducing negative examples of this type (i.e. hard negatives) supposedly would allow a system to improve upon the results of the source parser. The second type comprises sentences of at least 6 words that contain a dash or a comma, and a verb. We made the negative class twice as large as the positive one.

It is worth mentioning that cases marked 2 and 3 noticeably overlap with cases of gapping from the SynTagRus gapping test set, which we use to validate our AGRR corpus (see section 4; for cases 2 and 3 examples see the official shared task report (Smurov et al., 2019)).

The test set contains ten times fewer examples than the combined training and development sets with the same distribution of genres - 75% from fiction and technical literature, 25% from social media - and the same 1:2 ratio of positive to negative classes.

		0		1		sum
dev	vk	670	2760	326	1382	20548
	other	2090		1056		
train	vk	2860	10864	1366	5542	
	other	8004		4176		
test	vk	343	1365	185	680	2045
	other	1022		495		
sum			14989		7604	**22593**

Table 2: # examples by class; vk stands for social media texts

3.3 Dataset Format

When choosing the annotation format, we aimed to minimize reliance on any specific NLP frameworks and parsers. Since tokenization is often an integral part of NLP pipelines, we decided not to provide any gold standard tokenization and thus did not choose the commonly used CoNLL-U format.

Instead, markup of each sentence contains a class label (1 if gapping is present in the sentence, 0 otherwise) and character offsets for each gapping element (no offsets if sentence does not contain the corresponding gapping element).

4 SynTagRus Gapping Test Set

In order to test how well our corpus represents the phenomenon in question, we employ an alternative test set[2] obtained from SynTagRus - the dependency treebank for Russian that provides comprehensive manually-corrected morphological and syntactic annotation (Boguslavsky et al., 2009; Dyachenko et al., 2015).

To detect and extract relevant sentences, we rely on the original SynTagRus annotation (Iomdin and Sizov, 2009), i.e., the Nodetype attribute, which, if present with the value "FANTOM", indicates an omission in surface representation.

All the sentences were manually verified and divided into three categories:

1 cases similar to the ones encountered in the AGRR corpus;

2 cases of gapping not included in the AGRR corpus;

3 cases considered other types of ellipsis rather than gapping.

Sentences from all three categories as well as the number of aooripriate negative examples (obtained from SynTagRus with simple heuristics) will be further jointly referred to as the SynTagRus gapping test set.

We expect the systems trained on the AGRR corpus to show better results for category 1, because the examples may differ stylistically and thematically but not on a structural level. High scores obtained for category 2 would demonstrate that the corpus and the top systems were transferable to a broader range of gapping cases. Additionally, we provide the results obtained by the top systems for category 3.

We further illustrate the diversity of ellipsis cases in categories 2 and 3 using examples adapted from the SynTagRus gapping test corpus.

4.1 Gapping not Included in the AGRR Corpus

In Russian, the number of remnants is limited only by the valency of the predicate and can exceed two. Consider an example (9) with three remnants.

(9) [$_{cR1}$ В Испании] [$_{cR2}$ в 1923 году] [$_{cV}$
. In Spain in 1923 year
установил] диктатуру [$_{cR3}$
established dictatorship
генерал Педро де Ривера], [$_{R1}$ в
general Pedro de Rivera, in
Польше] [$_{R2}$ в 1926-м] - [$_{R3}$ Пилсудски].
Poland in 1926 - Pilsudski

'In Spain, the dictatorship of General Pedro de Rivera was established in 1923, while in Poland the dictatorship ~~was established~~ by Pilsudski in 1926.'

The AGRR corpus does not contain examples where the order of remnants differs from the order of correlates, though the structure is possible under certain conditions (Paducheva, 1974).

(10) [$_{cR1}$ Школа и уроки] [$_{cV}$
. school and lessons
принадлежали] [$_{cR2}$ кругу мучительных
belonged.to circle painful
обязанностей], а [$_{R2}$ душевному выбору]
duties and soul.ADJ choice
- [$_{R1}$ зеленая птица с красной головой].
- green bird with red head

'School and lessons belonged to the circle of painful duties, while a green bird with a red head ~~belonged~~ to the choice of the soul.'

The cases with two independent instances of gapping are not seen by the systems trained on the AGRR corpus. In (11) the bracketed sentence has its own gapping with overt predicate имеет 'has' not connected to the first occurrence of gapping, where predicate достигает 'reaches' is elided.

(11) [$_{cR1}$ Ширина долины] [$_{cV}$ **достигает**] [$_{cR2}$
. width valley reaches
600 км], [$_{R1}$ глубина] - [$_{R2}$ 8 км] (для
600 km, depth - 8 km for
сравнения: Большой каньон [$_{cV}$ **имеет**] [$_{R1}$
comparison Grand Canyon has
ширину] [$_{R2}$ до 25 км] [$_{R1}$ и глубину] [$_{R2}$
width to 25 km and depth
1,8 км]).
1.8 km

'The width of the valley reaches 600 kilometers, the depth ~~reaches~~ 8 kilometers (for comparison: the width of the Grand Canyon is about 25 kilometers and the depth ~~is~~ 1.8 kilometers).'

In Russian, gapping is not necessarily formed by omission of a verb. See (12), where the elided predicate is a noun (отчуджение 'isolation').

[2] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3001

(12) Бюрократизм привел к [$_{cV}$
red.tape led to
отчуждению] [$_{cR1}$ трудящихся] [$_{cR2}$ от
alienation working.people from
власти], [$_{R1}$ крестьян] [$_{R2}$ от земли].
power peasants from land
'Red tape led to the alienation of working people from power, and ~~alienation~~ of peasants from the land'

The SynTagRus gapping test set contains several examples illustrating a particular type of gapping that we refer to as gapping with generalization. In this type of gapping, the correlate clause semantically generalizes over instances described in subsequent gapped clauses. Furthermore, the main clause may lack the correlates of some remnants, e.g. промышленностью 'industry', наукой 'science' in (13).

(13) [$_{cR1}$ Средства и способы] создаются
. means and methods are.created
талантливыми учеными, а [$_{cV}$
talented scientists and
реализуются]: [$_{R1}$ средства] - [$_{R2}$ **военной**
are.realized means - military
промышленностью], а [$_{R1}$ способы] - [$_{R2}$
industry and methods -
военной наукой и опытом]
military science and experience.
'Means and methods are created by talented scientists, and are realized: the means ~~are realized~~ by the military industry, and the methods ~~are realized~~ by military science and experience.'

According to (Kazenin, 2007), gapping in Russian cannot elide an intermediate node in the tree structure. However, our data shows that such elision is possible. Consider (14), where the left correlate is higher syntactically than the elided predicate.

(14) Если [$_{cR1}$ можно] [$_{cV}$ передать] [$_{cR2}$ один
if is.possible transfer.INF one
университет], то почему [$_{R1}$ нельзя] [$_{R2}$
university, then why not.possible
другие]?!
others
'If it is possible to transfer one university, then why can't others ~~be transferred~~?!'

4.2 Other Types of Ellipsis

Along with cases of gapping not included in the AGRR corpus, we categorized sentences from the SynTagRus gapping test set that contain types of ellipsis other than gapping. Below we provide frequent categories of ellipsis with illustrations.

Ellipsis in comparative constructions (Bacskai-Atkari, 2018; Kennedy and Merchant, 2000) has restrictions that differ from gapping.

(15) От сна за рулем погибает
from sleeping behind wheel die
столько же водителей, сколько от
as.many drivers how.many/as from
алкоголя
alcohol
'As many drivers die from sleeping behind the wheel, as ~~many drivers die~~ from alcohol'

Cases where the second remnant is missing and the second clause contains just one remnant are called stripping (Merchant, 2016). Canonical examples of stripping are limited to a small number of constructions (16) - (17). According to (Hankamer and Sag, 1976), who introduced the term: "Stripping is a rule that deletes everything in a clause under identity with corresponding parts of a preceding clause except for one constituent (and sometimes a clause-initial adverb or negative)."

(16) The man stole the car after midnight, **but not** the diamonds. (Merchant, 2016)

(17) Abby can speak passable Dutch, and Ben, **too**. (Wurmbrand, 2013)

Our SynTagRus gapping test corpus contains examples with more (нет in (18)) and less canonical (причем in (19)) markers, but all of them can be distinguished from gapping with one remnant by the presence of closed set markers (see Section 2).

(18) Тогда деньги стали общими, а
Then money became shared and
экономики – **нет**.
economy - not.
'Then the money became shared, but the economy did not ~~become shared~~.'

(19) В Сталинграде каждый сражается,
in Stalingrad, everyone fights
причем как мужчины, так и женщины
and both men and women.'
'In Stalingrad, everyone continuously fights, both men and women ~~fight~~.'

Another type of ellipsis encountered in the SynTagRus gapping test corpus is sluicing (Merchant, 2001). Sluicing deletes the predicate from an embedded interrogative clause with no arguments remaining.

(20) Медикам дается указание как-то
doctors are.given instructions somehow
бороться с этим явлением, а как –
cope with this phenomenon and how -
никому не известно.
no.one NEG knows
‘Doctors are instructed to somehow cope with this phenomenon, but no one knows how ~~to cope with it~~.’

Finally, in the SynTagRus gapping test set there are numerous sentences with the following type of ellipsis: the repeating predicate is elided leaving only its arguments, and there are no correlates for arguments in the full clause. In sentences of this category, the second clause adds further details to the situation mentioned in the full clause.

Consider (20), where the predicate меняются (‘they change’) has no subject in the full clause, while it is added in the elided clause with одним игроком (‘by one player’).

(21) Правила меняются по ходу игры
rules are.changed with progress game
и всегда почему-то одним
and always for.some.reason one
игроком
player.INST
‘The rules are changed as the game progresses and for some reason ~~the rules are changed~~ always by one player’

In (22) the elided clause adds the manner справкой(‘by certificate’) to the action подтвердить (‘to verify’)

(22) Студент должен подтвердить свои доходы,
student must confirm his income
причем желательно справкой.
and preferably certificate.INST
‘The student must confirm their income, and preferably ~~confirm~~ with a certificate.’

5 Shared Task

In this paper, we revisit the information about the shared task that is essential for understanding the results of this paper (for details see the shared task report (Smurov et al., 2019))

We have formulated 3 different tasks concerning gapping with increasing complexity:

1. Binary presence-absence classification - for every sentence, decide if there is a gapping construction present.
2. Gap resolution - for every sentence with gapping, predict the position of the elided predicate and the head of the pronounced predicate in the antecedent clause.
3. Full annotation - for every sentence with gapping, predict the linear position of the elided predicate and positions of its remnants in the clause with the gap, as well as the positions of remnant correlates and the head of the pronounced predicate in the antecedent clause.

Solutions of all three tasks can be utilized by researchers studying gapping. Since sentences with gapping are naturally rare, the solution of the binary classification task will help researchers to find sentences with gapping for further analysis and data enrichment. Solutions of the other two tasks can be used to facilitate gapping resolution for parsing systems as well as to verify the quality of gapping annotation in syntactic corpora.

5.1 Metrics

The main metric for the binary classification task is standard f-measure. Two other tasks were scored based on symbolwise f-measure on gapping elements relevant to the particular task (all 6 for full annotation, V and cV for gap resolution).

The following is a description of symbolwise f-measure:

- true negative samples for binary classification task do not affect total f-measure;
- for true positive samples, symbolwise f-measure is obtained for each relevant gapping element separately, thus generating 6 scores for the full annotation task and 2 scores for the gap resolution task (if the evaluated sentence is either false positive or false negative, all the generated scores are equal to 0);
- the obtained f-measures are macro-averaged over the whole corpus.

One particular feature of the described metrics is that the second and the third task scores cannot exceed the first task score and thus binary classification errors are relatively harshly penalized in all three tasks. We have deliberately chosen such metrics since ellipsis is a rare language phenomenon and thus misclassification (false positive in particular) should be treated with caution.

6 Results and Analysis

6.1 Evaluation Results

Results of the top two participants on both the AGRR-2019 and the SynTagRus gapping test set are presented in Table 3. The implemented solutions are described in detail in the next section. The full table with shared task results as well as brief description of each participating system is available in the official report.

Corpus	Team	Binary	Gap	Full
AGRR	Winner	0.96	0.90	0.89
	2nd best	0.95	0.86	0.84
SynTagRus	Winner	0.91	0.76	0.77
	2nd best	0.88	0.67	0.64

Table 3: Top systems F1 scores on AGRR-2019 and SynTagRus test set. Binary: binary classification; Gap: gap resolution; Full: full annotation.

F1 scores on the SynTagRus gapping test set are measured for the subset consisting of categories 0 and 1. While examples of categories 2 and 3 cannot be reliably measured with the shared task metrics, we have calculated the number of examples of each category classified by the top systems as gapping. These results are shown in Table 4.

Cat	Total	Team	positives	positives, %
0	1166	Winner	8	0.7%
		2nd best	30	2.6%
1	507	Winner	433	85.4%
		2nd best	420	82.8%
2	75	Winner	26	35%
		2nd best	37	49%
3	100	Winner	6	6%
		2nd best	13	13%

Table 4: Number of sentences classified as gapping for each category of SynTagRus gapping test set.

Table 3 demonstrates that the AGRR-2019 corpus contains enough data for effective utilization of machine learning techniques. The results on the SynTagRus gapping test set in particular show that systems trained on the AGRR-2019 corpus are able to yield reasonably good results on a dataset obtained without any usage of the Compreno parser. While both systems experience a performance drop relative to scores on the AGRR-2019 test set, this can be attributed to domain shift (as two corpora have different genre composition etc.). In our opinion these results provide enough evidence to state that while the AGRR-2019 corpus has some inherent restrictions (see Section 4), it reflects a real-world linguistic phenomenon rather than the output of the Compreno system.

Performance on category 0 examples, as is shown in Table 4, demonstrates that high-precision systems can be trained on the AGRR-2019 corpus[3].

Performance on category 2 examples demonstrates that such systems can potentially recognize gapping examples of types completely unrepresented in the training set (obviously, performance on such sentences could be improved if similar examples were be added to the training set).

Performance on category 3 examples, by contrast, demonstrates that such systems can differentiate gapping from other types of ellipsis (including rather similar ones such as stripping and sluicing).

6.2 General Analysis

Most participants, including all top systems, treated gap resolution and full annotation tasks as sequence labeling tasks. The most popular approaches were to enhance the standard BLSTM-CRF architecture (Lample et al., 2016; Ma and Hovy, 2016), to pretrain an LSTM-based language model or to use transformer-based solutions (Vaswani et al., 2017; Devlin et al., 2018).

Most participating systems did not use any token-level features other than word embeddings, character-level embeddings, or language model embeddings (Peters et al., 2018; Devlin et al., 2018; Howard and Ruder, 2018). Of particular note is that neither of the 2 top-scoring systems used morphological or syntactic features. While it may be theorized that using such features could yield some improvements, we presume that language model embeddings (especially when coupled with self-attention as in the top two systems) contain most syntactic information relevant to ellipsis resolution.

6.3 Top Systems Analysis

The top two systems share several important elements: language model embeddings, self-attention (the winner as part of BERT, the second best team solution directly), and the part of the system designed to choose sound label

[3] It can be argued that the second best system has high false positive rate relative to the frequency of gapping in natural language. However one should keep in mind that classes 0 and 1 had 2:1 distribution in the training set. Changing this balance in favour of negative examples may potentially increase the precision of the systems. Moreover, manual analysis of these false positives shows that some of these examples do in fact contain gapping while many others are borderline.

chains (FSA-based postprocessor for the winner, NCRF++ for the second best team; (Yang and Zhang, 2018)). The third element is necessary when solving the task as sequence labeling (and more task-specific FSA-postprocessing yields better results). We can assume the first two elements combined contain most syntactic and semantic information relevant to ellipsis resolution.

The top two systems share one additional feature that most other systems lack: both are joined models that simultaneously learn the sentence-level gapping class and token-level gapping element labels.

We assume that this feature is relevant because it allows systems to minimize false positive examples for the gap resolution and full annotation tasks. Since false positive examples receive a rather harsh score penalty, joint training could potentially offer a substantial score improvement for the whole system.

7 Conclusion

We have presented the AGRR-2019 gapping corpus for Russian. Our corpus contains 22.5k sentences, including 7.5k sentences with gapping and 15k relevant negative sentences. The corpus is multi-genre and social media texts form a quarter of it.

It should be noted that to the best of our knowledge no other publicly available corpus for any language contains a comparable number of gapping examples. We believe that theoretical studies may also benefit from this data.

We have developed an annotation scheme that identifies gapping elements - parts of the sentence most relevant for gapping resolution from the theoretical point of view (see analysis in section 2). Our annotation scheme allows for successful solution of gapping resolution tasks by modifying standard sequence labeling techniques.

An important property of the AGRR-2019 corpus is that the systems trained on this corpus yield low number of false positives. Given the fact that gapping is a naturally rare phenomenon, this feature is extremely important.

While our corpus has some inherent limitations (see Section 4), the evaluation of the top system on the SynTagRus gapping test set demonstrates that the AGRR-2019 corpus is not an artificial creation of Compreno parser, but rather covers a large subset of Russian language gapping (see Section 6.1).

We hope that the size and diversity of our corpus will provide researchers interested in gapping with a valuable source of information that could bring the community closer to resolving ellipsis.

The corpus described in this paper can be utilized to improve parsing quality, possibly not only for Russian but for other Slavic languages as well.

Acknowledgments

We are grateful to Alexey Bogdanov for his help with linguistic analysis and useful advice. We are thankful to Ekaterina Lyutikova for her insightful comments on the SynTagRus gapping test set. We thank the anonymous reviewers for their valuable feedback.

The work was partially supported by the GA UK grant 794417.

References

Konstantin Anisimovich, Konstantin Druzhkin, Filipp Minlos, Maria Petrova, Vladimir Selegey, and Konstantin Zuev. 2012. Syntactic and semantic parser based on abbyy compreno linguistic technologies. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, volume 2, pages 90–103, Bekasovo, Russia.

Julia Bacskai-Atkari. 2018. *Deletion phenomena in comparative constructions: English comparatives in a cross-linguistic perspective*. Language Science Press, Berlin.

Alexey Bogdanov. 2012. Description of gapping in a system of automatic translation. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, volume 2, pages 61–70, Bekasovo, Russia.

Igor Boguslavsky, Leonid Iomdin, Svetlana Timoshenko, and Tatiana Frolova. 2009. Development of the russian tagged corpus with lexical and functional annotation. In *Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia*, pages 83–90.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Elizabeth Coppock. 2001. Gapping: In defense of deletion. In *Proceedings of the Chicago Linguistics Society*, volume 37, pages 133–148.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.0480*.

Kira Droganova, Filip Ginter, Jenna Kanerva, and Daniel Zeman. 2018. Mind the gap: Data enrichment in dependency parsing of elliptical constructions. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 47–54, Bruxelles, Belgium. Association for Computational Linguistics.

Pavel Dyachenko, Leonid Iomdin, Alexander Lazursky, Leonid Mityushin, Olga Podlesskaya, Victor Sizov, Tatiana Frolova, and Leonid Tsinman. 2015. Sovremennoe sostoyanie gluboko annotirovannogo korpusa tekstov russkogo yazyka (syntagrus). [the current state of the deeply annotated corpus of russian texts (syntagrus)]. *Trudy Instituta Russkogo Yazyka im. V. V. Vinogradova*, (6):272–300.

Jorge Hankamer and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry*, 7(3):391–428.

Daniel Hardt. 1997. An empirical approach to vp ellipsis. *Computational Linguistics*, 23(4):525–541.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. Association for Computational Linguistics.

Leonid Iomdin and Victor Sizov. 2009. Structure editor: a powerful environment for tagged corpora. *Research Infrastructure for Digital Lexicography*, page 1.

Kyle Johnson. 2001. *What VP ellipsis can do, and what it can't, but not why*. Citeseer.

Konstantin Kazenin. 2007. O nekotoryh ogranicheniyah na ellipsis v russkov yazyke. [on some restrictions on ellipsis for russian]. *Voprosy Yazykoznaniya*, (2):92–107.

Chris Kennedy and Jason Merchant. 2000. Attributive comparative deletion. *Natural Language and Linguistic Theory*, 18.

Jonathan K Kummerfeld and Dan Klein. 2017. Parsing with traces: An o (n 4) algorithm and a structural representation. *Transactions of the Association for Computational Linguistics*, 5:441–454.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL-HLT*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*.

Jason Merchant. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford University Press, Oxford.

Jason Merchant. 2016. *Ellipsis: A survey of analytical approaches*. University of Chicago, Chicago, IL.

Leif Arda Nielsen. 2005. *A corpus-based study of Verb Phrase Ellipsis Identification and Resolution*. Ph.D. thesis, King's College London.

Joakim Nivre, Paola Marongiu, Filip Ginter, Jenna Kanerva, Simonetta Montemagni, Sebastian Schuster, and Maria Simi. 2018. Enhancing universal dependency treebanks: A case study. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 102–107.

Elena Paducheva. 1974. *O semantike sintaksisa. Materialy k transformacionnoj grammatike russkogo jazyka. [On the Semantics of Syntax: Materials toward the Transformational Grammar of Russian]*. Nauka, Moscow.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*.

John Robert Ross. 1970. Gapping and the order of constituents. In *Manfred Bierwisch and Karl Erich Heidolph, editors, Progress in Linguistics, De Gruyter*, pages 249–259.

Sebastian Schuster, Joakim Nivre, and Christopher D Manning. 2018. Sentences with gapping: Parsing and reconstructing elided predicates. *arXiv preprint arXiv:1804.06922*.

Ivan Smurov, Maria Ponomareva, Tatiana Shavrina, and Kira Droganova. 2019. Agrr-2019: Automatic gapping resolution for russian. In *Computational Linguistics and Intellectual Technologies*, pages 561–575, Moscow, Russia. nakl. RGGU.

Yakov Testelets. 2011. Ellipsis in russian: theoretical and descriptive approaches. In *Tipologiya morfosintaksicheskih parametrov*, M.: MGGU.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Susi Wurmbrand. 2013. Stripping and topless complements. *Ms., University of Connecticut*.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.