ANTIPLAGIAT

Automatic Generation of Verbatim and Paraphrased Plagiarism Corpus

Andrey Khazov, Rita Kuznetsova
JSC Anti-Plagiat

Problem statement

Purpose

Generate corpus for PlagEvalRus-2017 competition.

Problem

There are no large Russian paraphrased plagiarism corpora, and manual generation is time-consuming.

Problem statement

- There is a set of candidate documents in Russian $T = \{t_i\}_{i=1}^{n}$ and source documents $W = \{w_j\}_{j=1}^{m}$, $T \cap W = \varnothing$ for corpus generation.
- For each t_i some part of text should be replaced with random text fragments from source documents.
- Replacing fragments can be paraphrased.

Algorithm description

- $\forall t_i$ we choose from one to K sources $\{w_k\}_{k=1}^{K} \subset W$.
- Random subset of sentences $\{c_{il}\}_{l=1}^{Z} \subset C_i$ is chosen from candidate document
- Each of c_{il} changed by one or more randomly chosen consecutive source sentences $c_l \rightarrow (s_{w_k}, s_{w_k+1}, \ldots, s_{w_k+v})$, $v \in \{1, \ldots, V\}$, where s_{w_k} — sentence from w_k source.
- Each of s_{w_k} can be paraphrased or not.

Examples

- Synonym replacement
  ***Скорее всего**, это пациент клиники для душевнобольных.*
  ***Должно быть**, это пациент клиники для душевнобольных.*
- Adding and removing synonym chains
  *Малому следовало **отдать** должное.*
  *Малому следовало **вернуть, возвратить или отдать** должное.*
- Adding and removing diminutives
  *Вчера утром Карл-Хайнц притащил три **бутылки** вина и галеты.*
  *Вчера утром Карл-Хайнц притащил три **бутыли** вина и галеты.*

Examples

- Singular/plural replacement
  В конце приведено графическое приложение *в формате А3 «Геоизотермы западной части Ново-Грозненского месторождения».*
  В концах приведены графические приложения *в формате А3 «Геоизотермы западной части Ново-Грозненского месторождения».*
- Abbreviation and disabbreviation
  Положение о взыскании налогов и неналоговых платежей, утвержденных постановлением ***ЦИК*** *и СНК СССР от 17 сентября 1932 года.*
  Положение о взыскании налогов и неналоговых платежей, утвержденных постановлением ***Центральная избирательная комиссия*** *и СНК СССР от 17 сентября 1932 года.*

Evaluation metrics

For evaluation we use adapted PAN assessed paraphrasing quality task metrics:

- **cosine similarity** between the original sentence and the paraphrased sentence;
- **integrity** — preservation of the original sentence's sense in the paraphrased, peer reviewed using a three-point scale, where 0 is the lowest value of metrics, 2 is the highest;
- **coherence** — correct morphological structure of the paraphrased sentence, measured using a three-point scale of expert assessment as well.

Paraphrasing quality

Paraphrase type	**Cosine similarity**	**Integrity**	**Coherence**
Synonym replacement	0.916	1.17	1.83
Adding and removing synonym chains	0.929	0.85	1.55
Abbreviation and disabbreviation	0.912	1.32	1.43
Adding and removing diminutive forms	0.942	1.07	1.87
Singular/plural form replacement	0.734	1.78	1.47

Results

- Proposed algorithm of generation of a corpus for verbatim and borrowing plagiarism.
- Prepared and tested list of paraphrasing types suitable for automatic generation in the Russian language.
- Developed quality metrics system .
- Generated corpus for the PlagEvalRus competition at the Dialogue 2017 conference.

Thank you for your attention!

Andrey Khazov, hazov@ap-team.ru
Rita Kuznetsova, kuznetsova@ap-team.ru