ANTIPLAGIAT
RESEARCH

Machine-Translated Text Detection in a Collection of Russian Scientific Papers

Alexey Romanov, Rita Kuznetsova,
Oleg Bakhteev, Anton Khritankov

research@antiplagiat.ru

Machine-Translated Text: Why So Serious?

- Recent advances in the field of statistical machine translation (SMT) lead to high availability of SMT systems on the Web.
- Student papers lack proper analysis by their tutors.
- It is very tempting to find relevant information in English, automatically translate it into Russian, and paste it into the paper “as is”!
- 1’500’000 bachelor and master theses are posted every year.
  - and even more reports and term works
- We need a fast automated way of spotting machine-translated text portions in huge collections of papers.

Machine-Translated Text: Why So Serious?

- SMT often contains grammatical errors or inappropriate words:
  - *First individuals in the system take the maximum number of contacts for any parameter combination.*
  - *Первые лица в системе взять максимальное количество контактов для любой комбинации параметров.*
- We focus on detection of these SMT shortcomings and consider two groups of them:
  - word salad
  - phrase salad
- The algorithm must:
  - work at sentence level
  - be fast (no deep analysis)

Word Salad

- **Word salad** is a “confused or unintelligible mixture of seemingly random words”
  - often attributed to mental diseases: logorrhea, schizophasia etc.
  - may be produced by low-quality SMT systems
- Examples in English:
  - *In worlds with pencils, schools page drink slime.*
  - *Take sharpness filling soda cans.*
  - *Run desk making dinner sunglasses menu.*
- Grammatical features:
  - semantic inconsistency
  - unexpected word combinations

Phrase Salad

- **Phrase salad:** grammatically correct phrases combined together in an improper way
  - occurs in output of SMT systems and text generators
- Examples in Russian:
  - *Все это имеет прямое, а также косвенное влияние на экономическую деятельность и производственных мощностей.*
  - *На практике не существует широкое признание процесс выбора параметра геометрическое распределение.*
  - *Масштаб и положение можно принимать любые значения, совместимых с областью временного ряда.*
- Grammatical features:
  - disagreement in sentence parts

Related Work

- MT quality estimation
  - [Gamon et al., 2005] — multiple syntactic features, French
  - [Avramidis et al., 2011] — PCFG features, English
  - [Aharoni et al., 2014] — presence / absence of POS N-grams, English
- Eliminating low-quality content from parallel corpora
  - [Antonova & Misyurev, 2011] — comparison to another decoder output, Russian + English
- Web spam detection
  - [Grechnikov et al., 2009] — word co-occurrence analysis, Russian
  - [Pavlov & Dobrov, 2011] — Markov chain generator output detection
- Word salad detection
  - [Arase et al., 2013] — language model likelihoods, English + Japanese

Formal Problem Statement

Formal Problem Statement

- $D = \{(x^i, y^i)\}_{i=1}^{m}$ — labeled set of sentences
- $x^i = f(s^i)$ — vector representation of a sentence in $\mathbb{R}^n$
- $s^i = (w_1^i, \dots, w_{k_i}^i)$ — ordered sequence of sentence words
- $y^i \in Y = \{0, 1\}$ — class label
  - 0 for authentic sentences, 1 for machine-translated sentences
- **Problem:** find optimal classification model g: $\mathbb{R}^n \to Y$:
  - with $D = D_L \sqcup D_T$, D_L being a training set, D_T — a test set

$$\hat{g} = \arg\min_{g} \frac{1}{|D_T|} \sum_{i=1}^{|D_T|} [g(s_i) \neq y_i]$$

Solution Design

Solution Design

- Let's estimate the likelihood that a sentence is machine-translated, according to several language models (LMs)…
  - Lexical 2,3-gram LMs trained on authentic texts
  - Lexical 2,3-gram LMs trained on machine-translated texts
  - POS tag 2,3-gram LMs trained on authentic texts
  - POS tag 2,3-gram LMs trained on machine-translated texts
  - word2vec (skip-gram and CBOW) models trained on authentic texts
- … and use these estimates as features for classification task.
  - 2 * 4 + 2 = 10 features in total
- N-gram models with N > 3 have negligible effect on classification performance
  - according to [Arase et al., 2013]

Lexical N-Gram Language Models

Estimated sentence likelihood:

$$\hat{p}_{LM}(s) = \hat{p}_{LM}(w_1, \dots, w_k) = \prod_{i=1}^{k} \hat{p}_{LM}(w_i \mid w_{i-N+1}, \dots, w_{i-1})$$

E.g., for N = 3:

$$\hat{p}_{LM}(w_i \mid w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}w_{i-1}w_i) + 1}{count(w_{i-2}w_{i-1}) + |V_3|}$$

V_3 — set of all corpus 3-grams

Lexical model sentence score:

$$score_{LM}(s) = \frac{1}{k} \log \hat{p}_{LM}(s)$$

POS Tag N-Gram Language Models

- POS tags contain morphological information:
  - part of speech
  - gender, case and number for nouns
  - gender, case and number for adjectives and participles
  - person, case and number for personal pronouns
  - person (or an indicator of infinitive form) for verbs

$$\hat{p}_{POS}(s) = \hat{p}_{LM}\big(h(w_1), \dots, h(w_k)\big)$$

$$h(w) — \text{POS tag for } w$$

$$score_{POS}(s) = \frac{1}{k}\log \hat{p}_{POS}(s)$$

- Lexical LM scores are aimed at *word salad* detection
- POS LM scores are aimed at *phrase salad* detection

Word2vec Likelihood Scores

- Skip-gram model sentence likelihood:

$$score_{SG}(s) = \frac{\log \hat{p}_{SG}(s)}{k} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbb{I}_{[1 \le |i-j| \le 5]} \log \hat{p}_{SG}(w_j | w_i)$$

- Continuous-bag-of-words model sentence likelihood:

$$score_{CBOW}(s) = \frac{1}{k} \sum_{j=1}^{k} \log \hat{p}_{CBOW}(w_j | s_{-j})$$

$$s_{-j} = (w_{j-5}, \dots, w_{j-1}, w_{j+1}, \dots, w_{j+5})$$

- $\hat{p}_{SG}$ and $\hat{p}_{CBOW}$ are obtained directly from pre-trained word2vec models

Experiments

Data Preparation

- 300K human-written sentences in Russian
  - jurisprudential and sociological papers
- 300K English sentences machine-translated into Russian
  - papers of the same domain
  - translated into Russian with one of popular online SMT systems
- 1M articles from Russian Wikipedia
  - short articles filtered out
  - used for word2vec training only
- Separate language model and classifier training
  - 200K + 200K sentences for language model training
  - 100K + 100K sentences for classifier training and validation

Experiment Setting

- Filtering and preprocessing
  - Russian sentences containing non-Cyrillic words filtered out
  - sentences split into tokens
  - words lowercased, punctuation removed
  - special token for numbers
  - unknown words filtered out
- *pymorphy2* for POS tags
- *gensim* for word2vec training
- *scikit-learn* for classification
  - best performance with RandomForestClassifier on 100 trees

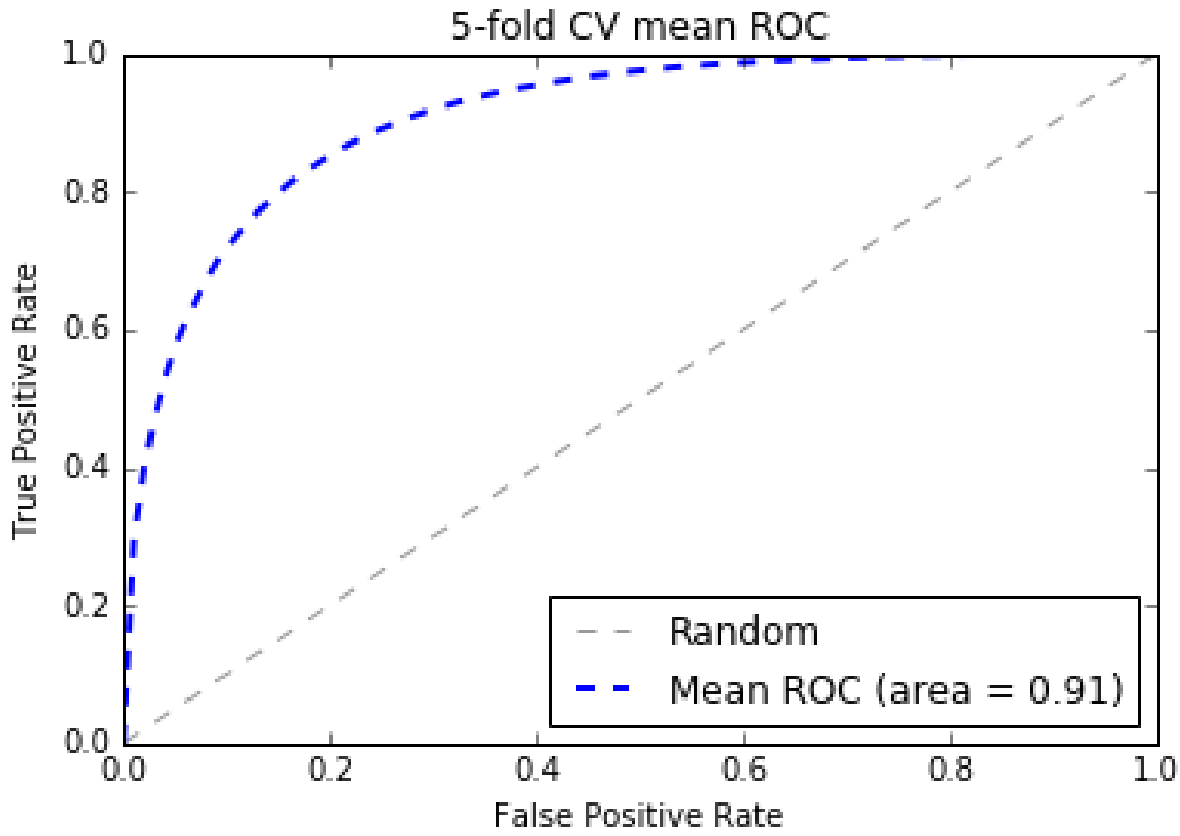
Experiment Results: Overall Performance

- Validation technique:
  - 5-fold cross-validation on the training set
- Performance measures:
  - precision, recall, F1 and AUC ROC

Features	F1	AUC ROC
Lexical 2,3-gram LMs (baseline)	0.754	0.816
POS tag 2,3-gram LMs	0.727	0.804
word2vec LMs	0.643	0.673
Lexical 2,3-gram + POS tag 2,3-gram	0.826	0.907
Lexical 2,3-gram + POS tag 2,3-gram + word2vec	**0.832**	**0.912**

Features	Prec	Rec
Lexical 2,3-gram + POS tag 2,3-gram + word2vec	**0.836**	**0.828**

Experiment Results: Overall Performance



Experiment Results: Feature Importance

- Scores obtained from the trained classifier

Feature	Importance ratio, %
Lexical 2-gram score (MT texts)	26.8
POS tag 3-gram score (authentic texts)	13.4
POS tag 3-gram score (MT texts)	11.5
POS tag 2-gram score (MT texts)	8.0
POS tag 2-gram score (authentic texts)	7.5
CBOW score	7.5
Lexical 3-gram score (MT texts)	6.9
Skip-gram score	6.4
Lexical 2-gram score (authentic texts)	6.2
Lexical 3-gram score (authentic texts)	5.9

Experiment Results: False Positive Errors

- Examples:
  - *Сопоставление с результатами натурного эксперимента.*
  - *При всей своей строгости и лаконичности эта модель обладает существенным недостатком — она является существенно эксплицитной.*
  - *Так, мысль Раскольникова, что, убив ростовщицу, он уничтожает только «вошь», паразита и, таким образом, совершает не столько преступление, сколько благодеяние, опровергается рядом обстоятельств.*
- Most common causes:
  - words and combinations with no occurrences in the training corpus
  - out-of-domain words
  - personal names

Experiment Results: False Negative Errors

- Examples:
  - *В среднем работал по 10 часов в день и 20 процентов работали по 12 часов в день.*
  - *Этот курс был запущен полностью практически между иностранными группами по 4-6 человек.*
  - *Преимущества РТС, в частности, темпы и уровни рынка они приводят.*
- Most common causes:
  - low rate of inappropriate translation
  - low rate of phrase salad
  - inconsistency that may be detected only by syntactic parsing

Conclusion & Future Plans

Conclusion & Future Plans

- The method is feasible for detection of machine-translated text at sentence level.
- Different language models can catch specific linguistic phenomena occurring in the output of SMT systems.
- Future plans:
  - LM training on corpora of higher quality and richness
  - accurate tuning of the method parameters
  - experiments on the output of rule-based MT systems
  - adaptation of the approach for the task of machine-generated text detection (output of context-free-grammar generators etc.)

References

- Aharoni R., Koppel M., Goldberg Y. (2014), Automatic Detection of Machine Translated Text and Translation Quality Estimation, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 289–295.
- Antonova A., Misyurev A. (2011), Building a Web-based parallel corpus and filtering out machine-translated text, Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, Portland, pp. 136–144.
- Arase Y., Zhou M. (2013), Machine Translation Detection from Monolingual WebText, ACL (1), Sofia, pp. 1597–1607.
- Avramidis E., Popovic M., Vilar D., Burchardt A. (2011), Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features, Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, pp. 65–70.
- Gamon M., Aue A., Smets M. (2005), Sentence-level MT evaluation without reference translations: Beyond language modeling, Proceedings of EAMT, Budapest, pp. 103–111.
- Grechnikov E. A., Gusev G. G., Kustarev A. A., Raigorodsky A. M. (2009), Detection of Artificial Texts [Poisk neestestvennykh tekstov], Proc. 11th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies” [Trudy XI Vserossiyskoy nauchnoy konferentsii “Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii”], Petrozavodsk, pp. 306–308.
- Pavlov A. S., Dobrov B. V. (2011), Detecting Mass-Generated Unnatural Texts through Topical Diversity Analysis [Metody obnaruzheniya massovo porozhdennykh neestestvennykh tekstov na osnove analiza raznoobraziya tematicheskoy struktury tekstov], Proc. 13th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies” [Trudy XIII Vserossiyskoy nauchnoy konferentsii “Elektronnye biblioteki: perspektivnye metody I tekhnologii, elektronnye kollektsii”], Voronezh, pp. 210–218.

ANTIPLAGIAT
RESEARCH

Thanks for your attention

Questions / Comments?