Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2016”

Moscow, June 1–4, 2016

ВЫЯВЛЕНИЕ МАШИННО-ПЕРЕВЕДЁННЫХ ТЕКСТОВ В КОЛЛЕКЦИИ НАУЧНЫХ ДОКУМЕНТОВ НА РУССКОМ ЯЗЫКЕ

Романов А. В. (romanov@ap-team.ru),
Кузнецова М. В. (kuznetsova@ap-team.ru),
Бахтеев О. Ю. (bahteev@ap-team.ru),
Хританков А. С. (khritankov@ap-team.ru)

Antiplagiat.Research, Москва, Россия;
Московский физико-технический институт
(государственный университет), Москва, Россия

Ключевые слова: обработка естественного языка, статистический машинный перевод, выявление машинного перевода, статистические языковые модели, word2vec

MACHINE-TRANSLATED TEXT DETECTION IN A COLLECTION OF RUSSIAN SCIENTIFIC PAPERS

Romanov A. V. (romanov@ap-team.ru),
Kuznetsova M. V. (kuznetsova@ap-team.ru),
Bakhteev O. Yu. (bahteev@ap-team.ru),
Khritankov A. S. (khritankov@ap-team.ru)

Antiplagiat.Research, Moscow, Russia; Moscow Institute
of Physics and Technology (MIPT), Moscow, Russia

In this paper, we propose a method of machine-translated text detection. By ‘machine-translated’ texts, we mean, principally, the output of statistical

machine translation systems. We focus on syntactic correctness and semantic consistency of sentences that constitute a text. More specifically, we make an attempt of detecting a certain phenomenon typically occurring in machine-translated documents. This phenomenon comprises the cases when small parts of the sentence, correctly translated, are combined together in an improper way. The proposed method is based on a supervised approach with a number of handcrafted features. First, we construct N-gram language models on a set of authentic scientific papers and on a set of machine-generated texts and assess the probability of each sentence according to these models. In addition, we propose N-gram language models on part-of-speech tag sequences corresponding to the texts given. Furthermore, we explore the effectiveness of features obtained from two trained word2vec (CBOW and skip-gram) models. We assess quality of the method on a sample of Russian scientific papers, and English scientific documents machine-translated into Russian. Preliminary results demonstrate feasibility of the approach.

Key words: natural language processing, statistical machine translation, machine translation detection, statistical language models, word2vec

1. Introduction

Recent advances in the field of statistical machine translation (SMT) and online translation services built upon it have enabled a massive increase in the volume of machine-translated texts available on the Web. Such a technology can be useful for providing information in languages with 'low density' on the Web. For example, automatically translated software documentation [15] becomes more readily available to speakers of these languages. On the other hand, prompt availability of translation systems leads to their potential misuse. Our experience with analysis of student home assignments, course projects and research papers shows that these often contain translated text fragments. While academic papers may contain text either translated by a human or translated with a translation system and post-processed by a human afterwards, student works frequently include fragments translated with one of public SMT systems and left "as is". Such works often lack detailed analysis by a human expert due to large amounts of them or due to the expert's negligence. Moreover, only a portion of text may be machine-translated, thus the problem of detection descends to the sentence level.

Development of automatic methods of machine-translated text detection may help to discover such misuse and is, therefore, of great importance for today's academic community.

Our approach is to develop an algorithm that would be able to classify sentences into two classes: human-written and machine-translated.

In this paper, we focus on detecting *word salad* and *phrase salad* phenomena, which are often produced by machine translation systems. The word salad is commonly defined as a "confused or unintelligible mixture of seemingly random words and phrases; the words may or may not be grammatically correct, but are semantically confused to the point" [22]. The phrase salad denotes text segments in which

small parts of sentences are semantically consistent, grammatically and syntactically correct, but are combined together in an improper way, so that the entire sentences become incorrect or absurd.

We propose features obtained from statistical language models (LM) such as N-gram LMs. We construct such LMs on words of the sentences and on sequences of part-of-speech (POS) tags corresponding to them. In addition, we investigate the applicability of the features obtained from trained word2vec (skip-gram and CBOW) models. We conduct a series of experiments on human-written Russian scientific papers, and English papers machine-translated into Russian in order to evaluate the quality of our method.

2. Related work

The problem of translated text detection has already been an active research topic for several years. Early works on this topic [6], [9] aimed to develop a method of automatic evaluation of machine translation systems (in order to track improvements of the system over time or to compare machine-translated texts with reference translations performed by human experts). The authors develop classifiers for distinguishing human-translated sentences from machine-translated ones and propose several groups of features, including language model likelihood scores for sentences, density of function words, features that capture syntactic relationships between words, and others. The topic of automatic SMT quality assessment is further developed in [4], where the authors focus on translation output ranking on a basis of the estimates of sentence grammaticality. In [1] the authors detect output of several SMT systems and conclude the dependence of detection quality and machine translation quality. The majority of works in this area rely heavily on human 'gold-standard' translation availability for source sentences, which is not the case in our task setting.

The development of corpus linguistics has drawn attention to automatic preparation of bilingual corpora. Several methods involve automatic detection of machine-translated text for eliminating low-quality content from parallel corpora. In [2] the authors try to detect the output of phrase-based SMT systems by exploiting their limited potential of word reordering within a sentence. Another approach [20] determines the likelihood of bilingual sentence pairs to be machine-translated using such features as character length ratio, token length ratio etc.

An area related to machine-translated text detection is detection of *translationese*, i.e. detection of overly literal translation performed without taking language features and idioms into consideration, in a collection of authentic documents written by native speakers. The proposed methods [5], [13], [21] are, in principle, related to aforementioned approaches in terms of variety of features used, but the problem setting is different.

Another related field is automatic web spam detection. The problem is similar to the one being considered, as it handles machine-generated texts lacking grammatical correctness. The authors of [11] analyze bigram co-occurrences in order to detect word salad. In [17] and [18] the authors focus on detection of texts generated

by specific models (Markov chain generators etc.) These works handle less complicated models of text generation, while SMT systems produce more diversified documents.

Our approach builds upon and extends the method presented in [3], which was designed for English-Japanese pair of languages. In addition to POS tag sequences and LMs, our approach also takes into account some specific characteristics of Russian words (e.g. the case of a noun, the tense of a verb etc.), which enable detection of disagreement in machine-translated sequences.

Moreover, we use features calculated with word2vec CBOW and skip-gram models [14] to capture statistical irregularities of machine-translated texts compared to authentic texts.

3. Method description

3.1. Formal problem statement

Our method is aimed at detection of machine-translated sentences in a mixed sample of authentic human-written sentences in Russian and sentences originally written in another language and machine-translated into Russian. In other words, our task is to determine whether each sentence of a document is machine-translated or not.

Let $D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{m}$ be a labeled set of pairs *(object, answer)*, where $\boldsymbol{x_i} = f(x_i)$ is a vector representation of sentence x_i in a feature space $\mathbb{R}^n$ $x_i = \left(x_i^1, \ldots, x_i^{l_i}\right)$; is an ordered sequence of sentence words; $y_i \in Y = \{0, 1\}$ is a class label, where 0 corresponds to the class of authentic sentences and 1 corresponds to the class of machine-translated sentences.

Let $D = D_L \sqcup D_T$ be a partition into a training set D_L and a test set D_T. We set up a problem of finding a classification model $g\colon \mathbb{R}^n \to Y$ that minimizes the empirical loss, which is the aggregated value of loss function $S\colon Y \times Y \to \mathbb{R}$, $S(y_1, y_2) = [y_1 \neq y_2]$, over test set D_T:

$$\hat{g} = \arg\min_{\mathrm{g}} S(g(x_i), y_i \,|D_T) \; = \arg\min_{\mathrm{g}} \frac{1}{|D_T|} \sum_{i=1}^{|D_T|} [g(x_i) \neq y_i]$$

3.2. Detection word and phrase salad phenomenon

SMT systems are known to generate texts that may lack semantic consistency and grammatical correctness. Basically, this effect takes place for several reasons. Modern SMT systems face many challenges, including homonymy and polysemy disambiguation, sentence structure transformation for languages that are not closely related, and others. Most of these systems handle segments of a sentence in a source language in order to produce a number of phrases in a target language, which are to be combined in a resulting sentence afterwards. Errors in the latter stage lead to a phenomenon known as *phrase salad*. This phenomenon comprises the cases when

grammatically correct phrases are combined together in an improper way. Several examples of phrase salad sentences in Russian produced by machine translation are:

(1) *Все это имеет прямое, как 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.

This phenomenon is a natural extension of the concept of *word salad*, which describes the cases when randomly chosen words or words from different domains constitute a sentence, which becomes a nonsensical set of words. This may frequently be attributed to the output of text generators that use Markov chains or context-free-grammars in their work.

The latter case is easier to detect using statistical language models since N-gram models trained on a set of authentic documents are sensitive to "unlikely" sequences of words. Phrase salad, however, is more subtle as these language models are not able to detect nonsensical phrases of length greater than N.

Based on these observations, we propose features to capture both word salad and phrase salad by estimating the likelihood of sentences according to previously trained statistical language models. These features capture grammatical correctness of the sentence and its lexical integrity. We attempt to build separate models for authentic human-generated text and for machine-translated text since each class of texts may contain particular hidden properties. By contrasting these two sets of features, we can effectively determine whether the sentence is machine-translated or not.

3.3. Lexical features

We build N-gram models for authentic texts and for machine-translated texts in order to compute likelihood of a sentence according to these models:

$$\widehat{p_{LM}}(x) = \widehat{p_{LM}}(x^1, \dots, x^l) = \prod_{i=1}^{l} \widehat{p_{LM}}(x^i | x^{i-1}, \dots, x^{i-N+1}),$$

where N-gram probabilities are estimated from text corpora. We also propose a small positive probability constant estimate for the N-grams that do not occur in the training corpus. A score of the sentence is log likelihood normalized by the length of the sentence.

We train 2- and 3-gram models on two different sets and thus get four different features from them. The length of N-gram is chosen not to be greater than 3 for several reasons. First, we restrict size of our language models for convenience and performance. Second, we take into account the findings of [3], where the authors confirm a negligible effect of extending models to 4-gram and higher.

3.4. Part-of-speech features

Part-of-speech features are also derived from statistical analysis of language models. Computation of this set of features is similar to that of features described in the previous section, but instead of word sequences we use POS tag sequences. Therefore, probability of a POS N-gram is given by:

$$\widehat{p_{POS}}(x) = \widehat{p_{LM}}\left(h(x^1), \dots, h(x^l)\right) = \prod_{i=1}^{l} \widehat{p_{LM}}(h(x^i)|h(x^{i-1}), \dots, h(x^{i-N+1})),$$

where $h: W \to H$ is a part-of-speech tagging function.

We use the following word tags:

- part of speech for all words;
- gender, case and number for nouns;
- gender, case and number for adjectives and participles;
- grammatical person, case and number for personal pronouns;
- grammatical person (or an indicator of infinitive form) for verbs.

We collect this information to discover syntactical disagreement among sentence words, for example, between a noun and a dependent adjective. This approach, as compared to full syntactic parsing, is a trade-off between accuracy of detection and computation time. Computational complexity of POS tagging is linear to the length of the input while polynomial for full syntactic parsing [8], [23]. For instance, when we see an adjective and an adjacent noun in a Russian sentence, this may not be the case of a noun phrase as they may refer to two different adjacent noun phrases. We make an assumption here that this case is less frequent in real sentences, than a potential disagreement between words in machine-translated sentences. We expect this set of features to be helpful in capturing phrase salad phenomenon.

We encode the entire portion of information obtained from POS tagger in a single token, which is used afterwards in construction of a language model. Like in the case of LM lexical features, we use 2- and 3-gram LMs.

3.5. Word2vec features

We train two different word2vec (skip-gram and CBOW) models on a larger separate sample of authentic Russian sentences. According to these models, sentence scores are computed with respect to the log-likelihood of the word appearing in a certain context, and vice versa. Thus, two more features are added to the feature set.

3.6. Classification

The proposed method associates each Russian sentence x with a feature vector $\boldsymbol{x} = f(x) \in \mathbb{R}^{10}$, We apply a random forest classifier to the two-class classification task in the feature space. The classifier choice is justified by the following. First,

we have a limited amount of features, which can be processed by tree composition algorithms with high accuracy. Second, it showed the best performance among several classifiers (including linear models and boosting methods) during our experiments.

4. Experiments

4.1. Data preparation

We prepare a collection of human-generated and machine-translated sentences extracted from scientific papers in a specific domain. This corpus is later used for both training of statistical language models and classifier training.

We attempt to confine sentences from both classes to use similar vocabulary for the sake of purity of the experiment.

As a source of human-written texts, we adopt a sample of jurisprudential and sociological papers available online [7] with open access. We obtain machine-translated sentences from a set of English papers of Munich Personal RePEc Archive [16], and translate them into Russian with an online translation service [10]. We focus on a single translation system since the majority of modern SMT systems produce similar output for the Russian language as a target language. We do not consider rule-based translation systems in this task, as they are likely to generate texts with different language phenomena. We also use a sample of 1M articles from the Russian Wikipedia for word2vec model training.

In human-written texts, we select only sentences that do not contain words in English or other non-Cyrillic words. Such words may or may not refer to machine translation errors and can bias our language models. As a result, we prepared the dataset of *300K* authentic sentences and *300K* machine-translated sentences. We used 2/3 of this collection for training of language models and the remaining part for classifier evaluation.

4.2. Experiment setting

We lowercase all the sentences and split them into tokens. We also remove punctuation and replace all numbers with the single token. We discard sentences that contain fewer than 4 tokens because short machine-translated sentences are known to be barely distinguishable from human-generated ones [3].

We use Python pymorphy2 package [12] to retrieve POS tags and relevant information and scikit-learn [19] RandomForestClassifier implementation. We use 5-fold cross-validation technique to train the classifier and to assess quality of the method on the prepared sample. We tune appropriate parameters of the classifier with grid search.

We evaluate the method with F1-measure, as this metrics aggregates the overall quality of two-class classification. Taking into account ability of the random forest classifier to predict probability scores, we also calculate AUC ROC metrics to obtain more stable performance characteristics of our algorithm.

4.3. Experiment results

4.3.1. Overall performance

We conduct a series of experiments with various subsets of the proposed features. Table 1 shows the results of this study. We consider the case when the features are restricted to LM lexical features as the baseline, since a method using these features has been proposed earlier by various authors. The results show the feasibility of our approach and the effectiveness of use of multiple LM-based features.

Table 1. Performance of different feature combinations

Features	F1	AUC ROC
LM lexical (baseline)	0.754	0.816
LM POS	0.727	0.804
word2vec	0.643	0.673
LM lexical + LM POS	0.826	0.907
LM lexical + LM POS + word2vec	**0.832**	**0.912**

4.3.2. Feature importance

We use feature importance estimates of the random forest classifier to understand whether all of the features are useful in the classification task or not. The numerical values are provided in the Table 2. They suggest that the proposed features of all types contribute to the decision made by the classifier and, therefore, are feasible for our task. Moreover, these findings confirm the benefits from constructing models not only on authentic texts, but also on machine-translated texts.

Table 2. Feature importance ratio

Feature	Importance ratio, %
LM lexical 2-gram score on machine-translated texts	26.8
LM POS 3-gram score on authentic texts	13.4
LM POS 3-gram score on machine-translated texts	11.5
LM POS 2-gram score on machine-translated texts	8.0
LM POS 2-gram score on authentic texts	7.5
CBOW score	7.5
LM lexical 3-gram score on machine-translated texts	6.9
Skip-gram score	6.4
LM lexical 2-gram score on authentic texts	6.2
LM lexical 3-gram score on authentic texts	5.9

4.3.3. Error analysis

We make a random subsample of 50 false positive errors and 50 false negative errors of classification in order to analyze the characteristics of misclassified sentences.

Example sentences of both classes are:

(2) *Сопоставление с результатами натурного эксперимента.* (false positive)
При всей своей строгости и лаконичности эта модель обладает существенным недостатком — она является существенно эксплицитной. (false positive)
Так, мысль Раскольникова, что, убив ростовщицу, он уничтожает только «вошь», паразита и, таким образом, совершает не столько преступление, сколько благодеяние, опровергается рядом обстоятельств. (false positive)

В среднем работал по 10 часов в день и 20 процентов работали по 12 часов в день. (false negative)
Этот курс был запущен полностью практически между иностранными группами по 4–6 человек. (false negative)
Преимущества PTC, в частности, темпы и уровни рынка они приводят. (false negative)

Overall, the most common causes of false positive classifications are:
- use of words and word combinations that do not occur in the training corpus (*натурного эксперимента, эксплицитной*);
- use of out-of-domain constructions and personal names (*Раскольникова, «вошь»*).

The most common causes of false negative classifications are:
- low rate of grammatical errors in translated texts;
- low rate of phrase salad.

These findings suggest that the quality of the method may be improved if a larger corpus of higher quality is used for language model training.

5. Conclusion

We propose a method of machine-translated text detection in a collection of scientific papers. The method is based on the supervised approach and operates individual sentences. We train statistical language models and use likelihood estimates of sentences as classification features. Preliminary experiments show feasibility of LM lexical and LM POS features and achieve decent results on a dataset of documents from a specific domain.

As the future work, we plan to improve the quality of our approach by tuning the parameters of the language models we construct. The experiment shows that different language models could catch specific language features that may occur in the output of SMT systems. Therefore, accurate tuning of the parameters is one of the appropriate tasks for future research. Another challenging problem is applicability of our approach to the more general task of machine-generated text detection, especially to detection of the output of context-free-grammar (CFG) based text generators.

References

1. *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.
2. *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.
3. *Arase Y., Zhou M.* (2013), Machine Translation Detection from Monolingual Web-Text, ACL (1), Sofia, pp. 1597–1607.
4. *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.
5. *Baroni M., Bernardini S.* (2006), A new approach to the study of translationese: Machine-learning the difference between original and translated text, Literary and Linguistic Computing, Vol. 21(3), pp. 259–274.
6. *Corston-Oliver S., Gamon M., Brockett C.* (2001), A machine learning approach to the automatic evaluation of machine translation, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, pp. 148–155.
7. *Cyberleninka.ru,* available at: http://cyberleninka.ru/
8. *Earley J.* (1970), An efficient context-free parsing algorithm, Communications of the ACM, Vol. 13(2), pp. 94–102.
9. *Gamon M., Aue A., Smets M.* (2005), Sentence-level MT evaluation without reference translations: Beyond language modeling, Proceedings of EAMT, Budapest, pp. 103–111.
10. *Google Translate,* available at: http://translate.google.com/
11. *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.
12. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages, Analysis of Images, Social Networks and Texts, Yekaterinburg, pp. 320–332.
13. *Kurokawa D., Goutte C., Isabelle P.* (2009), Automatic detection of translated text and its impact on machine translation, Proceedings. MT Summit XII, The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas, Ottawa.
14. *Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.* (2013), Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, pp. 3111–3119.
15. *MSDN*, available at: http://msdn.microsoft.com/
16. *Munich Personal RePEc Archive,* available at: http://mpra.repec.org/

17. *Pavlov A. S., Dobrov B. V.* (2009), Detecting Web Spam Created with Markov Chains Text Generators [Metod obnaruzheniya poiskovogo spama, porozhdennogo s pomoschyu tsepey Markova], 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. 311–317.
18. *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.
19. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., ..., Vanderplas J.* (2011), Scikit-learn: Machine learning in Python, The Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
20. *Rarrick S., Quirk C., Lewis W.* (2011), MT detection in web-scraped parallel corpora, Proceedings of the Machine Translation Summit (MT Summit XIII), Xiamen.
21. *Twitto-Shmuel N., Ordan N., Wintner S.* (2015), Statistical Machine Translation with Automatic Identification of Translationese, EMNLP 2015, Lisbon, p. 47.
22. *Word salad*, available at: http://en.wikipedia.org/wiki/Word_salad
23. *Younger D. H.* (1967), Recognition and parsing of context-free languages in time n 3, Information and control, Vol. 10(2), pp. 189–208.