Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

PART-OF-SPEECH TAGGING: THE POWER OF THE LINEAR SVM-BASED FILTRATION METHOD FOR RUSSIAN LANGUAGE

Kazennikov A. O. (kazennikov@iqmen.ru)

IQMen LLC, Moscow, Russia

We present our approach to Part-of-Speech tagging and lemmatization tasks for Russian language in the context of MorphoRuEval-2017 Shared Task. The approach ranked second on the closed track and on several test subsets it ranked first.

We proposed a filtration-based method which seamlessly integrates a classical morphological analyzer approach with machine learning based filtering. The method addresses both tasks in a unified fashion. Our method consists of two stages. On the first stage we generate a set of candidate substitutions which simultaneously recovers the normal form and provides all necessary morphological information. We select an optimal substitution for the current word given its context on the second stage.

The filtration stage of the presented method is based on Linear SVMs extended with hash kernel. The extension reduces the size of our model by an order of magnitude and allows to easily tune the tradeoff between the precision and the model size.

Keywords: POS Tagging, Morphological analysis, SVM, Hashing trick

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ: ФИЛЬТРАЦИОННЫЙ МЕТОД НА ОСНОВЕ SVM ДЛЯ РУССКОГО ЯЗЫКА

Казенников А. О. (kazennikov@iqmen.ru)

ЗАО «Айкумен-ИБС», Москва, Россия

В настоящей статье представлен метод снятия морфологической омонимии занявший второе место в общей таблице на закрытой дорожке соревнования MorphoRuEval-2017. Предлагаемый метод сочетает классический морфологический анализ и позволяет одновременно решать задачи лемматизации и восстановления морфологических признаков. Предлагаемый метод состоит из двух стадий: генерации возможных вариантов анализа словоформы и выбора корректной из списка возможных вариантов.

Первая стадия основана на анализе по словарю, состоящего из различных источников: конвертированного словаря АОТ, словаря, составленного по корпусу и предиктивного модуля. Вторая стадия реализована с помощью классификации на основе линейной SVM, дополненной алгоритмами хеширования. Это позволяет сократить модель признаков машинного обучения на порядок без какой-либо потери в качестве и в дальнейшем гибко настраивать соотношение между точностью снятия омонимии и размером модели.

Ключевые слова: Морфологический анализ, Снятие морфологической омонимии, SVM, Хеширование

1. Introduction

Morphological analysis plays an important role in almost any NLP pipeline, especially for morphologically-rich languages such as Russian language. It is usually one of the early stages of the pipeline, and the overall performance heavily depends on the quality of these first stages.

There exists a slight ambiguity in the formulation of the part-of-speech tagging problem. Early research on the problem was done mainly for the English language which has a relatively simple morphology, if compared, for example to the Russian language. So the term “part-of-speech” for English usually refers to an extended atomic tagset, rather than a strict part-of-speech tags such as “noun”, “verb”, or “adjective”. The distinction between strict POS tags and the extended atomic tagset is much higher for Russian, which has about 10 strict part-of-speech tags, whereas the full morphological model contains about 10 additional categories witch totals to about 40–60 morphological features (those numbers depend on the used morphological model), and results to over 300 atomic part-of-speech tags. This leads to severe precision penalties when successful approaches for English atomic POS tags are transferred to Russian language without modifications.

The second goal of the Shared Task, the lemmatization, is the task of reconstruction of the normal form of a word and is tightly coupled with the task of POS-tagging. This problem is more significant for the Russian language, because it is the highly inflected language. For example, the Zaliznyak’s dictionary[1] used in AOT project[2] contains about 120k word records which produce on expansion over 4.5M wordforms. That ratio is an order of magnitude higher if compared with English language.

Through this paper we will refer to “POS-tagging” as the task of recovery of a full set of morphological features for a word, and to “morphological analysis” as the joint task of POS-tagging and lemmatization.

The rest of the paper is organized as follows. Section 2 presents related work to the Shared Task. Section 3 describes the MorphoRuEval Shared Task setting. Section 4 introduces our approach to the MorphoRuEval POS-tagging and lemmatization tasks. Section 5 provides the evaluation results. Finally, we provide some concluding remarks in the last Section.

2. Related work

We identify three areas of research related to the MorphoRuEval Shared Task. The first area is the theoretic area of research of the tagset structure that could represent the linguistic properties of the Russian language. In this area we want to note the tagset of AOT project [2], the RusCorpora tagset [3], the SynTagRus tagset [4], the positional tagset [5], and the Universal Dependencies tagset [6].

The second area of research focuses on practical aspects of morphological analysis—the implementation of morphological analyzers. The approach of [7] is based on two separate finite state automata (FSA) for stems and endings, AOT project [2] uses a single automaton for storing the dictionary, the ETAP-3 NLP Processor [8] uses the idea of two-level Finite State Transducer for storing data for both analysis and morphological generation in a single FST. This area includes the research on predictive morphological analysis of unknown words. There we want to note the work of [2] which uses reverse endings to build a guesser FSA to deal with unknown words and [9] that introduces the normalizing substitution concept and presents some heuristics to lexical disambiguation.

The third area related to the Shared Task is the area of POS-tagging and disambiguation. The state-of-the art approaches are based on machine learning techniques. The notable approaches are the transformation-based approach of Brill tagger [10], the decision tree approach of TreeTagger [11], the classical approach based on HMMs of TnT tagger [12] and SVM-based approach of SVMTool [13], further elaborated in [14]. The recent research focuses on deep-learning approaches and various architectures [15, 16, 17].

3. MorphoRuEval Shared Task setting

All participants of the MorphoRuEval Shared Task were provided with several resources to train their models. Some of these resources were annotated and some were plain-text. We will focus on annotated resources only. They included:

1. GICR corpus, 1M tokens.
2. RNC corpus (the open part), 1.2M tokens.
3. SynTagRus corpus, 900k tokens.
4. OpenCorpora corpus, 400k tokens.

All corpora were converted to a simplified variant of Universal Dependencies morphological tagset format [6] (Table 1). The morphological model used in the Shared Task consisted of 12 POS tags and 12 feature categories. A valid parse contains at most one feature from each category. This totals in 40 features (of which 12 were POS tags). All corpora were semi-automatically converted to the Shared Task tagset format. This resulted in some inconsistencies between corpora. However, there were explicitly stated that all inconsistencies should be resolved in the favor of the GICR annotation flavor. Thus, the GICR corpus could be viewed as a gold-standard corpus, and the others as a source of potentially unreliable auxiliary information.

Table 1 sums up the morphological tagset used through the Shared Task. We should note that punctuation marks were treated as words too.

Table 1. Morphological model of the MorphoRuEval-2017.
Features skipped from the evaluation are marked with '*'

#	Category	Features
1	POS	NOUN, PROPN (same as NOUN), ADJ, PRON, NUM, VERB, ADV, DET, CONJ*, ADP, PART*, H*, INTJ*, PUNCT
2	Case	Nom, Gen, Dat, Acc, Loc, Ins
3	Number	Sing, Plur
4	Gender	Masc, Fem, Neut
5	Animacy	Anim*, Inan*
6	Tense	Past, Notpast
7	Person	1, 2, 3
8	VerbForm	Inf, Fin, Conv
9	Mood	Ind, Imp
10	Variant	Short/Brev
11	Degree	Pos, Cmp
12	NumForm	Digit

Table 2 presents some statistical properties of the provided corpora. It shows significant annotation inconsistencies between corpora used in the Shared Task.

Table 2. Training corpora statistics

Corpus	Tokens	Unique lemmas	Unique feature sets	Unique words
GICR	1M	43k	303	115k
SynTagRus	0.9M	43k	250	104k
RNC	1.2M	53k	557	127k
OpenCorpora	0.4M	42.5k	337	79k

The MorphoRuEval Shared Task had a strong focus on the evaluation of morphological aspects limiting the possible error sources. Both training and testing were done on pre-tokenized data, discarding any errors that could happen due to tokenization differences.

4. Proposed method

Our method integrates a classical dictionary-based morphological analysis pipeline with machine learning based disambiguation techniques.

The overall tagging procedure is straightforward and proceeds in greedy manner. It consists of the following steps:

1. Generate all parse candidates for each token of the sentence
2. Scan the sentence in the left-to-right manner.
    1. Score each parse candidate with respect to the sentence context
    2. Select the best parse

3. Assign it to the current token
4. Proceed to the next token

4.1. Candidate generation stage

The first stage of our model generates parse candidates for the given word. We used the normalizing substitution concept from [9] to represent a single parse candidate. A substitution is a triple of:

- The wordform ending
- The Normal form ending
- The full set of associated morphological features, including the POS tag.

This representation simultaneously provides candidate solutions for both goals of the Shared Task: it recovers morphological features of the word as well as the normal form.

A substitution is applied to the word in a trivial manner:

1. The ending is stripped from the word form,
2. The ending of normal form is appended,
3. All morphological features of the substitution are assigned to the parse.

For example, a substitution of

> (wfEnd="e", nfEnd="a", feats=NOUN, Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing)

transforms the word "*руке*" into "*рука*" and assigns respective features to the parse.

We used several data sources to build this module:

- A dictionary collected from the provided corpora, as it is the gold standard for features and lemmatization (after some experiments we used GICR corpus only).
- A partial transformation of the dictionary of AOT project [2] to the Shared Task tagset (the substitution mapping was performed on GICR joined with SynTagRus).
- A guesser for treating unrecognized words (we used GICR only again).
- Some simple heuristics for parsing special kinds of tokens (numbers, for example).
- A hand-crafted dictionary of frequent incorrectly parsed words (~50 wordforms total).

We collected the corpus-based dictionary at the first step. So we got a mapping from each wordform to a set of possible normalizing substitutions.

The conversion of the AOT dictionary posed some challenges. The Shared Task tagset doesn't maps one-to-one to any existing machine-readable dictionary of Russian. We designed a conversion procedure that maps normalizing substitutions of the corpus dictionary to the substitutions of AOT dictionary. We assume that if a corpus substitution perfectly matches an AOT dictionary substitution, then we could safely assign this corpus substitution to other AOT dictionary wordforms that derive this substitution. To do this, we used some transformations of AOT tagset to obtain a partially converted dictionary. Those transformations included:

- Rule-based direct feature mapping. For example "С" → "Noun"

- Splitting verb paradigms to verbs and participles (as they were treated as adjectives through the Shared Task)
- Conversion of short forms of adjectives to adverbs
- Post-processing of the immutable words like *кофе*.

To recover the full mapping we filtered the corpus dictionary through that partially-converted dictionary. We kept only substitution mappings that didn't produce any ambiguities. That led to a conversion of about a third of AOT dictionary substitutions and totaled in 1.6M converted wordforms.

Finally, we implemented a morphological guesser to get viable parses for out-of-dictionary words. The guesser was designed under assumptions of that: a) all irregular words are contained in the dictionary; b) unknown words are relatively long; c) all unknown words are derived from high-frequency word paradigms. The main idea of the guesser was inspired by [7]. We built two finite-state automata. One for the reversed endings of the word and another one for the reversed stems (prepended with the ending). For example wordform "*руке*" will be split into "е" as ending (id=42) and "кур" as reversed stem. The guess procedure is:

- Reverse the unknown word
- Traverse the endings FSA to find all possible endings
- For each ending, traverse the stems FSA and collect all possible substitutions
- Filter out unreliable parses (for example, if the recognized part is shorter than 3 characters)

At last we added small hand-crafted dictionary of frequent incorrectly-parsed words from other Shared Task corpora as they could appear in the test set. That included some words from Shared Task tagging rules (for example, tagging "*нет*" as a verb), and high frequency adverb/adjective ambiguities missing from AOT dictionary.

4.2. Filtering stage

The filtering stage selects a single parse from a set of generated parse candidates at the previous stage. The overall architecture was inspired by the SVMTool [12] and was further elaborated in [13]. The filtering algorithm is quite trivial: score each parse of the word against the context and choose the highest-scoring one.

The Shared Task tagset contains over 300 different combinations of morphological features. Using a 300-class classifier seemed highly impractical as it doesn't take advantage of the tagset structure and secondly, that the provided datasets were relatively small and highly imbalanced for this approach.

We trained a separate classifier for each group of features separately. This resulted in 12 multiclass classifiers instead of 300 binary ones.

The following tagging procedure was used:

1. Collect all morphological features from each parse candidate. This step reduces the number of classifier evaluations.
2. Score each feature against the context of the word.
3. Select a parse that:
    1. Has the highest ranking part-of-speech
    2. Has the maximal sum of feature scores.

The selection procedure was split into two parts to prevent the case when the sum of feature scores outweights the part-of-speech score. It is undesirable as we found out that part-of-speech classifier has a negligible error rate (about 0,8%).

4.3. Feature group classifier and the context feature model

We used a modified SVM multiclass classifier of LIBLINEAR [17] to score a single feature group. It uses one-against-all classification scheme and the training algorithm optimizes all classes simultaneously. We modified the original implementation by replacing a weight vector for each class by a shared vector by means of the hashing trick [18]. The basic idea is the replacement of the dot-product function:

$$dot(w, x, i) = \sum_j w[i][j]x[j]$$

by:

$$dot(w, x, i) = \sum_j w[hash(i, j)]x[j]$$

where w is a weight vector, x is a feature vector, i is the class we are scoring against, and $hash(i, j)$ is a hash function that maps its inputs to an integer value from a predefined range (regarded as effective feature count). Our system used MurmurHash3[19] as the hashing function and 2M as effective feature count. The effective feature count is independent of the number of classes, so the per-class effective feature count is a fraction of the total feature space. For example, if the effective feature count is 1M and the number of classes is 10 then the effective feature count per class is just 100k.

The hashing trick allows to easily tune the resulting feature space size. Another benefit of the hashing trick is that we discarded the feature mapping table of the one-hot encoding procedure and significantly reduced memory requirements for our method.

The drawback of the hashing trick is in its lossy compression scheme. And if the chosen effective feature count is too small for the problem, the hash function collisions could significantly reduce the model performance.

We estimated the total number of distinct features of our model and used it as an initial effective feature count. We tried to double the number of effective features and haven't seen any significant performance improvement. After that, we tried to halve the effective feature count and observed some performance loss. So we used the original estimation of the effective feature count through all experiments.

Our feature model produces about 3M distinct features. The hashing technique reduced the effective per-class feature count by an order of magnitude without significant performance loss.

The model uses mostly context features. We used a context window of size 7 (±3 words around of the main one). The context window was divided into two parts: the tagged part (words before the current one), and untagged one (words starting with the current one). All parses in the tagged part were already resolved and we could use all available information (such as case, number gender features) from them.

The features used for the tagged part of the context were inapplicable because words in the untagged part don't have a resolved parse yet. To overcome this we used a concept of *ambiguity class* over the morphological category. It is a sorted set of the possible morphological features of that category collected from candidate parses of the word. For example, for wordform "человека" the ambiguity class over the "Case" category would be "genitive/accusative", because we don't know the correct case for the word yet, but we can narrow it to two options instead of six.

For each word of the full context we use following features:

- word prefixes of length 2, 3 and 4
- word suffixes of length 2, 3 and 4
- wordform itself
- lowercased wordform
- For each word of the tagged part of the context:
- POS tag of the word
- POS tag + suffix
- POS tag + suffix of the main word
- Number, Gender, Case morphological categories of the word (and their combinations)
- Stem and the Ending

For each word in untagged part of the context (starting from the main one):

- Ambiguity classes for POS, Number, Gender and Case categories
- Ambiguity classes for POS, Number, Gender and Case categories coupled with suffix of the main word

Finally, for the main word we used some additional features:

- A flag for the main word is at start of the sentence
- Capitalization of the main word

5. Experiments

We conducted several experiments on the different combinations of training/test data during the development of our system. The results are presented in Table 3.

Table 3. Evaluation of our model on different training/test set combinations

Training/Test pair	POS-only,	POS-full	Lemma	Lemma+POS
GICR/GICR (9:1)	99,23%	94,52%	98,59%	76,28%
Syntagrus/Syntagrus (9:1)	97,85%	91,78%	97,73%	58,22%
RNC/RNC (9:1)	96,64%	70,28%	94,08%	25,33%
OpenCorpora/OpenCorpora (9:1)	98,17%	57,29%	98,51%	14,53%
GICR/Syntagrus	96,24%	88,85%	97,26%	48,81%
GICR/RNC	95,18%	68,64%	93,67%	23,77%
GICR/Opencorpora	97,11%	55,91%	97,93%	13,61%

Table 3 shows a significant loss of precision when the model was trained on one corpus and tested on a different one. The Shared Task organizers explicitly stated that the GICR annotation could be viewed as a reference and all inconsistencies should be resolved in favor of GICR annotation. As a result, we tuned our model to the GICR annotation.

Table 4. Effect of using partially-converted AOT dictionary, GICR corpus

Training/Test pair	POS-only,	POS-full	Lemma	Lemma+POS
Guesser only	98.22%	91.99%	86.02%	45.45%
Guesser + Corpus Dict	99.04%	94.37%	98.96%	76.67%
Guesser + Corpus dict + AOT Dict	99.23%	94.52%	98.59%	76.28%

We note high sensitivity of the proposed model to the quality of the generation stage (Table 4). The “guesser only” mode generates all parse candidates guesser only. The “guesser + corpus dict” experiment show synthetic results when the parse candidates of the GICR corpus dictionary were complemented heuristically by the guesser results (to handle the situation when there is a potential parse of the word that didn’t occur in the corpus). Our final model (Guesser + Corpus dict + AOT Dict in the table) shows significant improvement from the proposed corpus-dictionary mapping procedure.

Our final model was trained on the GICR corpus. Our final results on the closed track of the Shared Task are presented in Tables 5–8. Our results are marked with bold, the best ones are marked by ‘*’.

Table 5. Precision on News subset of the test set

Team ID	Tags, by word	Tags, by sentence	Lemmas, by word	Lemmas, by sentence
O	**93.99%***	**63.13%**	**92.96%**	**54.62%***
A	93.83%	61.45%	93.01%*	54.19%
C	93.71%	64.8%*	—	—
H	93.35%	55.03%	81.6%	17.04%

Table 6. Precision on Vkontakte subset of the test set

Team	Tags, by word	Tags, by sentence	Lemmas, by word	Lemmas, by sentence
H	92.42%*	63.59%	82.8%	35.39%
O	**92.39%**	**64.08%**	**91.69%***	**61.09%***
C	92.29%	65.85%*	—	—
A	91.49%	61.44%	90.97%	60.21%

Table 7. Precision on Fiction subset of the test set

Team	Tags, by word	Tags, by sentence	Lemmas, by word	Lemmas, by sentence
C	94.16%*	65.23%*	—	—
O	**92.87%**	**60.91%**	**92.01%***	**57.11%***
A	92.4%	60.15%	91.46%	55.08%
H	92.16%	56.6%	77.78%	22.08%

Table 8. Precision on full test set

Team	Tags, by word	Tags, by sentence	Lemmas, by word	Lemmas, by sentence
C	93.39%*	65.29%*	—	—
O	**93.08%**	**62.71%**	**92.22%***	**58.21%***
H	92.64%	58.4%	80.71%	25.01%
A	92.57%	61.01%	91.98%	56.49%

Our approach ranked second, losing to the top system slightly more than 0.3% on POS-tagging task. On News subset our system showed top precision. On the Vkontakte subset our system lost about 0.04% to the top one. The result tables show that our method is strongly consistent and robust across different text sources types.

On the lemmatization task, our approach ranked top, seconding only in the News subset with the gap of only 0,05%. The lemmatization performance was also consistent across different text sources types.

6. Conclusions

We presented an approach to the part-of-speech tagging and lemmatization that is closely related to classical morphological analysis frameworks. The two-stage scheme showed high precision and robustness. That allowed our model to get a consistent second rank on the POS-tagging task of the closed track of the MorphoRuEval-2017 Shared Task, even ranking first on several test subsets. Our method ranked first on the full test set of the lemmatization task, ranking second only on News subset with the gap of 0,05% to the top system.

Experiments showed that the model performance significantly depends on the consistency of the corpus annotation and for this level of precision corpora-to-corpora differences are critical to the model performance.

The application of the converted AOT dictionary significantly improved the overall performance of our method. The consistency of morphological information between the generation phase of our model and gold standard corpus also was critical to the success of our method.

We believe that the performance of the presented method could be improved by further efforts on dictionary-to-corpus matching.

The source code for all experiments is available at: https://github.com/kzn/morphoRuEval.

Acknowledgements

We want to thank MorphoRuEval organizers team for their work in organizing this Shared Task competition.

References

1. *Zaliznyak, A. A.* (1980). Russian grammatical dictionary [Grammaticheskyi slovar' russkogo yazyka]. Russkij Jazyk, Moskva.
2. *Sokirko A. V.* (2004) Morphological Modules on AOT.RU [Morfologicheskie moduli na saite AOT.RU]. Dialogue conference proc.Moscow, pp. 559–564.
3. *Lyashevskaya O. N., Plungian V. A., Sichinava D. V.* (2005) On the morphological standard for Russian National Corpus for Russian Language [O morphologicheskom standarte Natsional'nogo korpusa russkogo yazyka]. Natsional'nyi Korpus Russkogo Yazyka, 2005(2), pp. 111–135.
4. *Apresyan Yu. D., Boguslavsky I. M., Iomdin B. L., Iomdin L. L., et al.* (2005). Syntatic and semantic annotated corpus for Russian language: state-of-the-art and perspectives [Sintaksichesky i semantichesky annotirovannyi korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy]. Natsyonalny corpus russkogo yazyka. 2005, pp. 193–214.
5. *Hana, Jirka, and Anna Feldman* (2010). "A Positional Tagset for Russian." In LREC.
6. *Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning Ch. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D.* (2016), Universal Dependencies v1: A Multilingual Treebank Collection, Proc. of LREC 2016, Portoroz, Slovenia, pp. 1659–1666.
7. *Segalovich, I.,* (2003). A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In MLMTA (pp. 273–280).
8. *Kazennikov A.O* (2008). Using Finite Automata for Morphological Analysis and Synthesis Based on the Dictionaries of the ETAP-3 System, [Ispol'zovanie konechnyi avtomatov dlya morphologicheskogo analiza i sinteza na osnove slovarei sistemy ETAP]. Sb. tr. konf. molodykh uchenykh i spetsialistov ITIS (Proc. Conf. Young Scientists and Specialists of ITIS), pp. 201–205
9. *Zelenkov Yu., Segalovich I., Titov V.* (2005) Probabilistic Model for Morphological Disambiguation based on Normalizing Substitutions and Nearest Word Positions [Veroyatnostnaya model' snyatya morfologicheskoy omonimii na osnove normalizuyuyschikh podstanovok i pozitsiy sosednikh slov]. Dialogue conference proc., Moscow pp. 188–197.
10. *Brill, E.* (1992, February) A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language (pp. 112–116). Association for Computational Linguistics.
11. *Schmid, H.* (1995) Treetagger| a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, p.28.
12. *Brants, T.* (2000) TnT—A Statistical Part-of-Speech Tagger. "6th Applied Natural Language Processing Conference".

13. *Gimenez, J. and Marquez, L.,* (2004) SVMTool: A General POS Tagger Generator Based on Support Vector Machines, Proc. 4 Int. Conf. Language Resourc. Evaluat. (LREC'04), Lisbon, Portugal, pp. 43–46.
14. *Petrochenkov V. V., Kazennikov A. O.* (2013) A Statistical Tagger for Morphological Tagging of Russian Language Texts. Automation and Remote Control, Vol. 74, No. 10, pp. 1724–1732
15. *Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P.* (2011) Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), pp. 2493–2537
16. *Huang, Z., Xu, W. and Yu, K.* (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.
17. *Plank, B., Søgaard, A. and Goldberg, Y.* (2016) Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. arXiv preprint arXiv:1604.05529.
18. *Fan, R.-E., Chang, K.-W., Hsieh, C.-J., et al.* (2008) LIBLINEAR: A Library for Large Linear Classification, J. Machine Learning Res. vol. 9, pp. 1871–1874.
19. *Shi, Q., Petterson, J., Dror, G., et al.* (2009) Hash Kernels for Structured Data, J. Machine Learning, vol. 10, pp. 2615–2637.
20. *Appleby A.* (2008) Murmurhash 3.0. https://github.com/aappleby/smhasher.