ПРИМЕНЕНИЕ МОДЕЛИ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ МОРФОЛОГИЧЕСКИХ ХАРАКТЕРИСТИК СЛОВ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

Музычка С. А. (s.muzychka@samsung.com)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия;
ООО «Исследовательский центр Самсунг», Москва, Россия

Романенко А. А. (a.romanenko@samsung.com)

Московский физико-технический институт
(государственный университет), Москва, Россия;
ООО «Исследовательский центр Самсунг», Москва, Россия

Пионтковская И. И. (p.irina@samsung.com)

ООО «Исследовательский центр Самсунг», Москва, Россия

В статье рассматривается проблема снятия морфологической омонимии для русского языка с помощью статистических методов, а именно аппарата условных случайных полей (англ. *Conditional Random Fields,* CRF). Предлагается модифицированная модель CRF, дающая результаты, соответствующие state-of-the-art.

Также рассматривается применение CRF для нормализации цифровой записи числительных. Приводятся результаты вычислительного эксперимента.

Ключевые слова: снятие морфологической омонимии, условное случайное поле, CRF, нормализация текста, NLP

CONDITIONAL RANDOM FIELD FOR MORPHOLOGICAL DISAMBIGUATION IN RUSSIAN

Muzychka S. A. (s.muzychka@samsung.com)
Lomonosov Moscow State University, Moscow, Russia;
Samsung R&D Institute Rus, Moscow, Russia

Romanenko A. A. (a.romanenko@samsung.com)
Moscow Institute of Physics and Technology, Moscow, Russia;
Samsung R&D Institute Rus, Moscow, Russia

Piontkovskaja I. I. (p.irina@samsung.com)
Samsung R&D Institute Rus, Moscow, Russia

We consider the problem of morphological disambiguation in Russian using statistical methods; specifically, we apply conditional random field (CRF). We propose a new modified model of linear chain CRF, which demonstrates results close to the state-of-the-art. We also propose a new statistical approach to text normalization problem using CRF. Namely, we solve the problem of normalization of numerals written as digits. Our approach allows for the consideration of both cardinal and ordinal numbers.

In order to train and test our models we used Russian text corpora. For morphological disambiguation, we used data from OpenCorpora and the SynTagRus linguistic corpus. For number normalization we used the Russian National Corpora (RusCorpora).

A brief overview of the CRF model is given, followed by a detailed description of the applied algorithm, assumptions on the training and test set, and a description of features for each particular issue.

Key words: morphological disambiguation, conditional random field, text normalization, NLP

1. Introduction

One of the key problems in text processing is morphological disambiguation. The results of this analysis can be applied to another natural language processing (NLP) problems: extracting named entities, syntactic parsing, sentiment analysis, etc.

While it is considered this problem to be solved for English language, there are some difficulties for languages with rich morphology, in particular for russian language. To solve the problem for Russian language both rule-based and statistical approaches are applied. The main obstacles in the solution of the problem are, firstly, the lack of a single common morphological tagset and, secondly, absence of common training and test corpora for verification of implemented algorithms.

CRF algorithm presents state-of-the-art results in many NLP problems related to sequence labeling. In this articles we consider an application of CRF algorithm for the solution of two problems: morphology disambiguation and number normalization. To solve this problem we propose modified CRF models that enable to reduce learning time.

The rest of the paper is structured as follows. Section 2 introduces basic concepts of CRF models, section 3 describes its application for morphological disambiguation, and section 4 presents an algorithm for number normalization. Each section contains description of used data and results.

2. Conditional Random Fields

This section describes basic concepts of CRF models introdused in [9].

2.1. General CRF Model

Definition 1. Let $G=(V,E)$ be an undirected graph. The set of random variables $\{\xi_v\}, v\in V$ form a Markov random field (MRF) with respect to *G* if they satisfy the local Markov properties:

1. **Pairwise Markov property**: any two non-adjacent variables ξ_x and ξ_y are conditionally independent given all other variables $\{\xi_v\}_{v\in V\setminus\{x,y\}}$.
2. **Local Markov property:** each variable ξ_x is conditionally independent of all other variables given its neighbors $\{\xi_v\}_{v:\{x,v\}\in E}$.
3. **Global Markov property:** any two subsets of variables $\{\xi_v\}_{v\in A}, \{\xi_v\}_{v\in B}$, $A,B\subset V$, $A\cap B=\emptyset$, are conditionally independent given a separating subset ($S\subset V$ the set of nodes is called separating for two non-intersecting subsets if each path from the first subset to the second passes through it).

Definition 2. The maximal clique of the graph *G* is called any maximal fully connected subgraph of *G*.

Theorem (*Hammersely-Clifford*) *The set of random variables $\xi=\{\xi_v\}, v\in V$ is a MRF corrsesponding to the graph G if and only if its distribution $p(\xi)$ is factorized by cliques of the graph, that is*

$$p(\xi)=\frac{1}{Z}\prod_{c\in C}\Psi_c(\xi),$$

where C is the set of all maximal cliques of G, Ψ_c are some functions depending on $\xi=\{\xi_v\}, v\in c$ only, and Z is a normalization factor called partition function.

Definition 3. CRF is an MRF such that the set of its nodes is decomposed into two noninteracting subsets $V=X\cup Y$ where *X* and *Y* are the sets of observed and hidden variables correspondingly.

Everywhere below we use the following notations $x=\{\xi_v : v\in X\}$ and $y=\{\xi_v : v\in Y\}$. Also we suppose that the random variables from *x* and *y* belong to some arbitrary discrete spaces *X* and *Y* correspondingly.

The inference problem is to predict the optimal values of *y*, given the observations *x*. According to Hammersely-Clifford theorem we should optimize

$$p(y\,|\,x)=\frac{1}{Z(x)}\prod_{c\in C}\Psi_c(x,y),\ \textit{where}\ Z(x)=\sum_{y'}\prod_{c\in C}\Psi_c(x,y')\ \textit{is a partition function.}$$

Usually Ψ_c are chosen as an exponent of the linear combination of some features with coefficients that should be determined during training. As a rule these coefficients depend on the structure of the clique only and given training set $Tr=\{(x^j,y^j)\}$ are fitted by maximizing the log-likelihood probability

$$\ln p(Tr)=\sum_j \ln p(y^j\,|\,x^j).$$

2.2. Linear Chain CRF

The structure of the model is depicted on Figure 1.

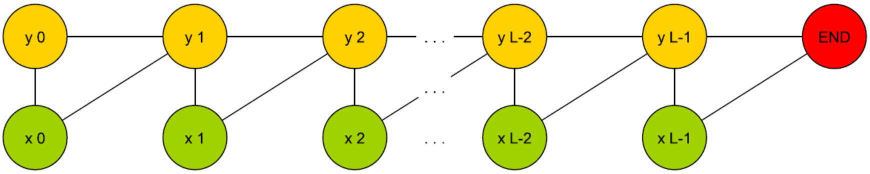

Fig. 1. Linear chain CRF

The nodes $y_0, y_1,\ldots, y_{L-1}$ correspond to our hidden parameters *y*, and $x_0, x_1,\ldots, x_{L-1}$ correspond to the observations *x*. The last node END is terminal and its value without loss of generalty can be set to any fixed element which we for simplicity denote by 'end'. Every clique in the graph consists of two consequitive hidden nodes y_k, y_{k+1} and one observable node x_k. Consequently, the functions Ψ_c inroduced above have the form $\Psi(x,y',y'')$. The observations x_k are usually represented as a vector of binary features

$$x_k=(f_k^1, f_k^2,\ldots, f_k^F), f_k^i\in\{0,1\}.$$

Finally, $\Psi(x,y',y'')$ are chosen in the form

$$\Psi(x,y',y'')=\exp\left(\sum_{i=1}^{F}\lambda(i,y',y'')f_k^i+b(y',y'')\right)$$

where $\lambda(i,y',y'')$ and $b(y',y'')$ are parameters of the model that should be tuned during training.

In order to optimize maximum likelihood function any gradient descent method can be used.

3. Morphological Disambiguation

3.1. Previous Results

Currently, there are a large number of papers devoted to the definition of parts of speech (POS-tagging). For example, the papers [1,5] describe classifiers with accuracy 95–97%. We consider the problem of full morphological disambiguation (POS+MORPH), that is for each token in the sentence we should assign the corresponding morphological labels: part of speech, gender, animacy, etc.

Note that is not always possible to quantitatively compare two different systems of analysis of the Russian language since often they use different tagsets. Also sometimes punctuation tokens are used for calculating accuracy of the algorithm. Also it is important which morphological dictionary is used: some systems use external resources (for example, Zaliznyak dictionary); the others extract dictionary from training data [5].

In paper [6], evaluation of morphological analysis on full tagset is done. Their tagset consists of 829 tags. The presented result is 94,46%. But, as far as it can be understood from the article, there is no unknown (out-of-vocabulary) words in their test set.

In [5], the result 95.25% is performed on universal MTE tagset.

Finally, there are corresponding disambiguation algorithms with high accuracy of classification for such morphologically rich languages as Czech, Hungarian, etc. The following table represents results for them reported in [2]:

Language	**arabic**	**czech**	**spanish**	**german**	**Hungarian**
Number of morphological labels	516	1811	303	681	1071
Accuarcy	90.32	92.94	97.93	88.58	96.34

3.2. Application of CRF for Morphological Disambiguation

For most languages, the standard CRF model is well suited to solve this problem and provides a practically significant results. However, for such morphologically rich languages like Russian and Czech, where the total number of morphological markers of several hundred, the direct application of the linear model of CRF may cause technical problems.

Firstly, the complexity of gradient computation in the model described above is a quadratic function of the total number of hidden states of the model, and therefore the algorithm converges too slowly. Secondly, the total number of parameters of the models is quadratic in the number of hidden states Во-вторых, количество свободных параметров модели растет квадратично с ростом числа скрытых параметров, which, in turn, increases the complexity of the model. Finally, there is a possibility of attributing any morphological tags to any token in a sentence (for example, there is a possibility to assign label «adjective» to the token «кошка»). At the same time, as a rule, for each token, we have to choose from 3–4 parsing options. To overcome the described difficulties, we consider the classic model of linear CRF in a modified form.

3.3. Description of the Data

To train and test our model, we use the Syntagrus corpus [3]. The size of this corpora is significantly lower than the corresponding analogs – OpenCorpora and RusCorpora (National corpora of Russian language) (total number of sentences is 53439, total number of tokens is 774368). Nevertheless, Syntagrus has certain advantages. Firstly, it has uniqely defined labels. For comparison, OpenCorpora provides only possible morphological labels for each token and therefore can not be used for learning. Secondly, Syntagrus provides marking for syntax trees, which makes it possible to apply the developed algorithm in subsequent parsing.

3.4. Assumptions

During training and testing we made the following simplifications:

1. Labels that contained НЕСТАНД, МЕТА and НЕПРАВ were replaced by UNKNOWN.
2. All tokens with latinic symbols were marked as NID.
3. The difference between the comparative adjectives and comparative adverbs is not considered.
4. Morphological labels СТРАД, СЛ, ПРЕВ are not considered.

To conclude the total number of labels is 353.

3.5. Morphological Dictionary

Since Syntagrus is a rather small corpora we can't use data only from it in order to compose an acceptable dictionary. Consequently, we use dictionary from OpenCorpora project (based on Zalizn'yak grammar dictionary), and then converts marks into SynTagRus format for convinience.

3.6. Description of the Algorithm

The structure of our model is depicted on the following figure

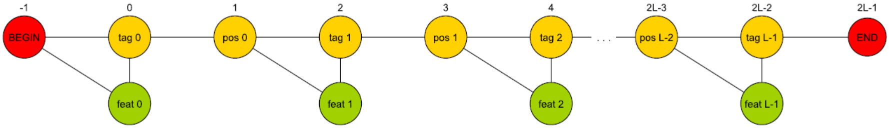

Fig. 2. CRF model, used for morphological disambiguation

The set of hidden elements is decomposed into union of two non-intersecting subsets $Y = labels \cup pos$, where *labels* is the set of all morphological labels, and *pos* is the set of all possible parts of speech. On Figure 2 the nodes $y_{2k} = l_k \in labels$ corresponds to the full morphological label of the *k*-th token of the sentence and $y_{2k+1} = p_k \in pos$ to its POS. The nodes $y_{-1} = begin$ and $y_L = end$ are initial and terminal nodes. Their content can be set to any arbitrary values and we without loss of generality prefer to denote them 'begin' and 'end' correspondingly. We see that the graph contains two types of cliques. The cliques of the first type consist of one *label*-node y_{2k}, one *pos*-node y_{2k-1} and one feature node x_k. Therefore this type of cliques can be described by function $\Psi(x, y', y'')$. The cliques of the second type correspond to the transition from *label*-node y_{2k} and one *pos*-node y_{2k+1} and are described by function $\Psi(y', y'')$. We assume that this transition is deterministic i.e. we set

$$\Psi(y', y'') = I\ \{POS\ of\ y'and\ y''coincide\}.$$

Note that the imposed restrictions allows to solve solve the two first problems described above, namely, the number of free parameters of the model is now proportionally to $|\,labels\,| \cdot |\,pos\,|$ that is significantly slower than $|\,labels\,|^2$. Finally for each node y_k we store a list of possible morphological labels $list_k$. During inference and gradient descent computations we can take into considerations only that label sequences that satisfy the imposed restrictions. This simultaneously helps to prevent very serious mistakes in the prediction of hidden labels, and reduce training time.

3.7. Features

The basic features used in experiments are possible morphological labels derived from OpenCorpora dictionary. Also we the most used tokens (conjunctions, particles, etc.) and token endings were added to the feature set.

3.8. Results

We used first 45,000 sentences for training, and the rest 8439 sentences for testing. The number of tokens in the test set is 121 968. The accuracy of the algorithm is 91,06% on the test set. This result is worth than the similar results from papers [5,6]. But distinctive feature of our algorithm is that it doesn't use lexical information at all. In [3] it is shown that lexical features could improve overall accuracy up to 9%. Our experiments show that even without lexical information statistical approach can be applied to the problem of morphological disambiguation and it performs satisfactory results.

The following table shows the distribution of errors on the part of speech.

	a	adv	com	conj	intj	nid	num	part	pr	S	unknown	v
a	1494	204	5	2	0	36	94	5	3	283	5	140
adv	198	117	1	119	1	20	20	72	4	165	3	15
com	1	0	0	0	0	6	0	0	0	3	0	0

	a	adv	com	conj	intj	nid	num	part	pr	S	unknown	v
conj	8	144	0	0	0	0	0	176	0	135	0	0
intj	0	0	0	0	0	5	0	0	2	0	0	0
nid	5	4	0	0	0	0	0	0	0	67	0	2
num	52	1	0	0	0	8	131	0	0	17	0	0
part	42	38	0	263	2	2	0	0	0	92	0	13
pr	6	74	1	0	0	2	0	0	0	28	0	0
s	304	169	7	72	5	292	38	22	10	4645	23	84
unknown	3	4	0	3	0	2	0	3	0	22	0	0
v	203	28	1	7	2	13	8	13	3	69	0	507

Here rows of the table correspond to the real parts of speech, and columns to the predicted. The diagonal elements contain the number of errors associated with incorrect additional morphological labels (case, gender, etc.) in particular, we see that the greatest number of errors occur within the group of nouns (S) and adjectives (A). The latter is due to the fact that for the distinction of the nominative and accusative cases it is necessary to consider further links in a sentence that is not guaranteed by a linear model CRF. Finally, the accuracy of the classifier on the reduced tagset containing only parts of speech is 96,7% that is consistent with the existing analogues.

4. Normalization of Russian Numerals

The one common task, which appears in Text-To-Speech System development process, is normalization of non-standard words such as abbreviations, acronyms or numerals written in digits [4, 7]. TTS system should be correctly pronoun non-standard phrase, so it should can decode word and set them to proper grammar form.

In this section we give an example of application of linear-chain CRF to detection of grammar form of Russian numerals written in digits. Previous work [8] describes variant of solution of this task based only on frequency features, not on grammar features of contexts of numerals. Also solution described in [8] makes one strict assumption: all numerals written in digits are cardinal that often breaks down in practice. We avoid this assumption and propose to use grammar features of words from context of numerals.

4.1. Dataset and Feature Generation

We used subset of National Russian corpora. Phrases containing numerals with grammar features were selected and all numerals were converted to digit form.

Features that we generate for tokens can divide into 5 parts.

1. Grammar features of words (not numerals). In practice we use disambiguation tool described above.
2. Features indicated that word is specific for cardinal or ordinal numerals. For instance, names of currency usually take place with cardinal numerals.

3. Features indicate prepositions, because they help determine case of nominals.
4. Features that indicate spelling characteristics of numerals: length in symbols, terminal digit, etc.
5. Features from group 1–4 for neighbor tokens

4.2. CRF Model Details

We use modified model of linear-chain CRF in the same manner as in morphological disambiguation algorithm. This model is shown below on figure 3.

We maximize probability of labels sequence "POS of previous token, type and grammar form of numerals, POS of next token" when we have observed sequence "features of previous token Feat$_{prev}$, features of numerals Feat$_{num}$, features of next token Feat$_{next}$. In addition we present type and grammar form of numerals like sequence of five labels: TYPE (cardinal or ordinal), CASE, GEN (masculine, feminine, neuter or unknown), SNGL (singular, plural or unknown) and ANIM (animacy, inanimacy or unknown). We include in features of token features of 7 right and 7 left neighbor tokens.

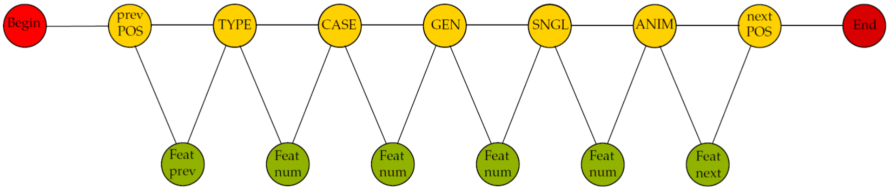

Fig. 3. Model CRF used for detection type and grammar form of numerals

4.3. Experiments

As noticed above we use for training and testing algorithm subset of National Russian corpora (10268 phrases with numerals). We split all data set on training (8251 phrases) and testing (2017) parts.

Accuracy of result model on test set is Also in table below we show result of detection type and grammar form of numerals averaged by category of labels. We evaluate P precision, R recall and F$_1$-measure as quality measures.

Quality measure	TYPE, %	CASE, %	GEN, %	SNGL, %	ANIM, %
P	97.21	91.33	89.77	82.39	87.66
R	97.21	92.93	90.74	85.97	95.05
F_1	97.21	92.10	90.24	84.05	91.11

Moreover, we evaluate quality of model with 5-fold cross-validation procedure. Result of evaluation is so applying model has high generalization ability.

4.4. Remarks and Result Analysis

Type and grammar form of numerals which are predicted by model described above allow convert numerals to form of words with help of finite-state automaton. This finite-state automaton and model CRF described above give a system of numerals normalization.

It is needed to be noticed that actual accuracy of system should be higher. Firstly, numerals with different grammar features in form of words sometimes match together, and CRF model mistakes usually in these cases. For example, model confuses on nominative and accusative cases. Secondly, when system detects definite gender, animacy or number instead of label unknown it is not mistake. Second note considered accuracy of algorithm rises to 94.53%.

Nevertheless, using in text numerals in form of digit or in form of words are depends on many conditions: kind of text, narrative style of author, etc. But we use as training data phrases with numerals largely in form of words. So, the issue about data quality and quality of model remains open.

5. Conclusions

In article we give some examples of applications of conditional random fields to important tasks of natural language processing. The accuracy of disambiguating algorithm based on CRF is 91.06%, for task of normalization of numerals accuracy is 94.53%. As result of paper we mark up that quality of machine learning approach to NLP tasks for Russian is at a level of quality of rule-based analogues but statistic approach needs less human resources.

References

1. *Antonova A. Ju., Solovyev A. N.* (2013), Conditional random field models for the processing of russian [Ispol'zovanie metoda uslovnyh sluchajnyh polej dlja obrabotki tekstov na russkom jazyke], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013"], Bekasovo, pp. 39–52.
2. *Muller T., Schmid H., Schutze H.* (2013), Efficient Higher-Order CRFs for Morphological Tagging, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, p. 322–332.
3. *Nivre J., Boguslavsky M., Iomdin L.* (2008), Parsing the SynTagRus treebank of Russian, Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 641–648.
4. *Olinsky C., Black A. W.* (2000), Non-standard word and homograph resolution for asian language text analysis, In proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), ISCA, pp. 733–736

5. *Sharoff S., Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011", Bekasovo, pp. 591–605.
6. *Sokirko, A. V., Toldova, S. Y.* (2004). Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian. Proc. of Corpus Linguistics–2004, Saint-Petersburg. [In Russian].
7. *Sproat R., Black A. W., Chen S. F. and Kumar Sh., Ostendorf M., Richards C.* et al (2001), Normalization of Non-Standard Words, Computer Speech & Language, Vol. 15, pp. 287–333
8. *Sproat R.* (2010), Lightly supervised learning of text normalization: Russian number names, Workshop on Language Spoken Technology (SLT), IEEE, pp. 436–441
9. *Sutton C., McCallum A.* (2006), An Introduction to Conditional Random Fields for Relational Learning, MIT Press.