АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ПРАВИЛ ДЛЯ СНЯТИЯ МОРФОЛОГИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ

Протопопова Е. В. (protoev@gmail.com),
Бочаров В. В. (victor.bocharov@gmail.com)

Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

Ключевые слова: омонимия, морфологическая разметка, русский язык, неконтролируемое обучение

UNSUPERVISED LEARNING OF PART-OF-SPEECH DISAMBIGUATION RULES

Protopopova E. V. (protoev@gmail.com),
Bocharov V. V. (victor.bocharov@gmail.com)

Saint Petersburg State University, Saint Petersburg, Russia

Morphological disambiguation is one of the key aims of part-of-speech tagging. The task is considered to be solved, though all the tools for disambiguation use a lot of manually created data. This paper describes an attempt to disambiguate Russian corpus without manually annotated data. The method used was proposed about twenty years ago but has not been applied to synthetic languages yet. The main idea of our approach is to derive disambiguation rules automatically from a corpus with ambiguous annotations using only a few statistical data. It can be done in a simple way by means of unsupervised learning. The results are quite high and can be compared to results of existing systems. We also tried to measure the size of the corpus necessary to produce a reasonable set of disambiguation rules and showed that it can be comparable in size with the corpora used to train statistical disambiguation models.

Keywords: ambiguity, Russian language, morphological annotation, unsupervised learning

1. Introduction

Although we have observed a great improvement in POS-tagging for English and other European languages and works on this topic became quite rare during past ten years, the number of works devoted to POS-tagging of Russian language is rather low. The existing systems [Zelenkov et al. 2005, Sokirko, Toldova 2005, Sharoff, Nivre 2011] use machine learning methods which require a lot of manually annotated data. In this work we have tried to apply an unsupervised algorithm described in [Brill 1995]. This method has several advantages:

- It is based on automatically annotated corpus and the manually created annotation is necessary only for evaluation.
- The output of the system is a list of rules which can be understood and explained.

In this paper we are going to examine, what accuracy we can achieve using this unsupervised model, what corpus do we need to train the model effectively and how the size of corpus affects the resulting rules.

Our approach is based on that described in [Brill 1995]. We should mention that we did not take into account morphological categories other than part-of-speech, though case ambiguity, for example, is very common in Russian. The core idea of Brill's approach is to gather statistical information about POS tags and their distribution. The rule to transform tags A_B into tag B can be obtained if the system has seen that tag B is more frequent than tag A in the observed context.

We conducted a number of experiments to assess the size of training corpus. The algorithm was applied to corpora of various size — from 1K to 170K sentences and we obtained different lists of rules thereby. We present and compare the results in Section 4.

2. Related work

All known approaches to disambiguating Russian corpora are trained on manually tagged part of Russian National Corpus. The first such algorithm presented in [Zelenkov et al. 2004] is based on statistical knowledge about tags and their contexts and its accuracy is more than 97%. The probability of each possible tag is computed as sum of probabilities of this tag in the context multiplied by their scores ("influence" on its environment).

A tagger based on Hidden Markov Model is described in [Sokirko, Toldova 2005] and achieves up to 98% accuracy. The algorithm takes into account trigram probability for tags (tag0 was seen after tag1 and tag2) and bigram probability for words (word1 was tagged as tag1 after tag0). Manually annotated part of RNC (5M words) was used as training data.

Later research in [Sharoff, Nivre 2009] showed that HMM with some improvements (guessing inknown words by their ending) can be successfully applied for POS tagging and achieves a competitive result of about 97% on a reduced tagset. The authors mention common problems of statistical POS tagging stating however that "a completely automatic machine learning procedure can quickly produce a fast and reliable NLP component".

3. Methods and data used

3.1. The original algorithm

The idea was in short described above and here we want to mention some details. We did not change the original algorithm but there are a few details in implementation.

First, a text is annotated by an initial-state annotator, which can assign a word either a random structure or an output of a manually-created dictionary. Then a learner is given transformation templates such as following:

Change tag from x to Y in context C.

where x is set of tags (i.e. morphological hypothesis) assigned to a word, $Y \in x$. C is one context feature: one word or tags to the left or to the right are considered as context features.

Unlike supervised learner, the unsupervised one cannot measure the accuracy of the transformation, hence a special scoring function is used to find more reliable disambiguation contexts. In each learning iteration the score is based on the current tagging of a corpus.

Computing the score for the transformation above includes three steps:

1) For each tag $Z \in x$, $Z \neq Y$ compute

$$\frac{freq\,(Y)}{freq\,(Z)} \cdot incontext(Z, C)$$

2) Let

$$R = argmax_z \frac{freq\,(Y)}{freq\,(Z)} \cdot incontext(Z, C)$$

3) Then score for this transformation is

$$score = incontext(Y, C) - \frac{freq\,(Y)}{freq\,(R)} \cdot incontext(R, C)$$

where *freq(X)* is number of occurences of words unambiguously tagged with tag X and *incontext(X,C)* is number of occurences of words unambiguously tagged with tag X in context C.

On each interation the algorithm searches the transformation which maximizes the scoring function and the learning stops when no positive scoring transformations can be found.

In [Brill 1995] the algorithm was applied to tagging English corpus and resulted in 95.1% accuracy on 200K words part of Penn Treebank and 96.0% accuracy on 350K words part of Brown corpus.

In our implementation we took into account only nearest right and left context including punctuation and sentence borders. If several rules have the same score and it is the best score on this iteration, they are applied in descending order of the chosen tag frequency.

The rules were obtained from 10 random corpora of each size and are written in the following way:

ADJF NOUN → NOUN | 1:tag=PNCT

that is

Change tag from ADJF NOUN to NOUN if next tag is PNCT.

We also store the following information for each rule: its score, number of rule applications on this iteration, number of occurences of an ambiguous tag.

3.2. Differences between English and Russian tagset

The tagset for inflective languages such as Russian is bigger and different in its structure from one for English because it includes not only part of speech tags but also grammatical categories such as case, number, gender, tense etc. For one word form a morphological hypothesis is a set of tags that can be considered as a set of key-value pairs where key is a grammatical category: part-of-speech — noun, number — single, case — nominative etc. Each morphological interpretation is a set of morphological hypothesis (i.e. set of tag sets). The following example includes three hypothesis (one hypothesis per line) for form “чай” that can be both imperative mood of verb “чаять” (to hope, to expect) and nominative or accusative case of noun “чай” (tea):

чай VERB impf sing excl tran impr

чай NOUN masc sing inan nomn

чай NOUN masc sing inan accs

According to the morphological dictionary we used there are 4369 different tag sets (lines in the example above) and 1678 of them are assigned to more than 10 forms. It makes a little sense to use such a big tagset in machine learning because most of items are very rare. The obvious solution is to split disambiguation task into several steps: one step per grammatical category (as in [Acedanski, Gołuchowski, 2009], the approach described below). In this paper we describe the first step where only part of speech tags are considered. The example with form “чай” now looks much simpler and the step-wide tagset is reduced to only part-of-speech tags:

чай VERB
чай NOUN

Our annotator assigns unknown words tags UNKN (unknown sequence of cyrillic characters), LATN (unknown sequence of latin characters), NUMR (numeric characters) and PNCT (punctuation). The algorithm also uses tags SBEG and SEND as sentence borders context.

A similar but supervised approach was applied to Polish POS-tagging [Acedanski Gołuchowski, 2009]. Tagging was performed in two phases, first of all POS, case and person were disambiguated and then other categories.

3.3. Training and test corpora

We used articles from http://www.chaskor.ru/ as training data. 15M tokens were annotated using OpenCorpora morphological dictionary and ten disjoint random corpora of each size (from 1K to 170K sentences) were derived from it. The annotation is represented in a simple way:

48,501 Согласно	328,254 согласно	328,255 согласно	328,258 согласен
	ADVB	PREP	ADJS Qual neut sing

Each hypothesis include lemma's id, lemma and morphological annotation itself.

For the test set we have taken a random sample of manually disambiguated sentences from OpenCorpora project.

4. Results

Our first aim was to find out what corpus size is required to produce a reasonable set of rules. Since there is no straightforward way to measure it, we made a number of experiments. 325 lists of rules obtained from different training sets were compared and the results are described below.

4.1. Disambiguation rules

First of all, the lists of rules were compared according to their size and content. The results (fig. 1) show that the number of rules increases if we increase the size of training corpus because the number of contexts increases respectively. The number of unique rules increases less but it did not become absolutely stable on big corpora (more than 3M words).

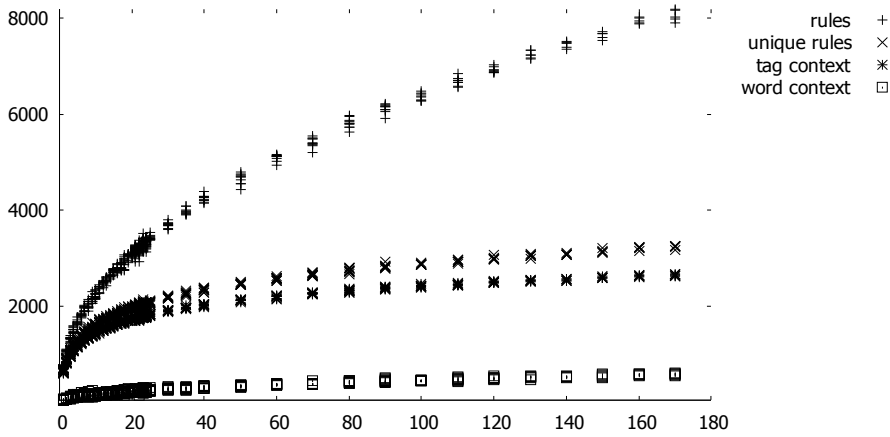

Fig. 1. Number of disambiguation rules obtained from different corpora

The similarity between lists of rules was also measured as Spearman rank correlation coefficient, which is higher for bigger training corpora (fig.2). The increasing correlation coefficient shows that same rules are ranked in almost the same order in lists obtained on big corpora.

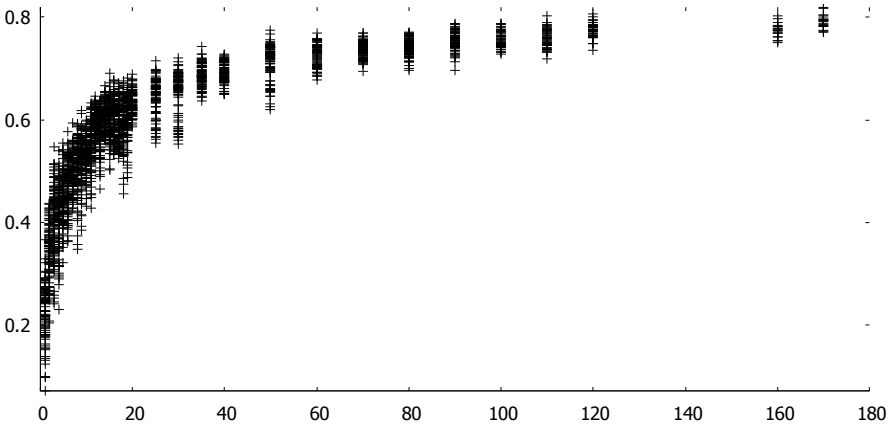

Fig. 2. Spearman rank correlation between each two sets of rules

4.2. Unsupervised annotation results

Another way to check if we got enough rules for tagging is to examine the difference in annotation after applying different rule lists. A test set containing 1K sentences was disambiguated using various sets of rules. First of all, we compared the resulting annotations with each other. We see that the difference decreases if we increase size of the training corpus (fig. 3).

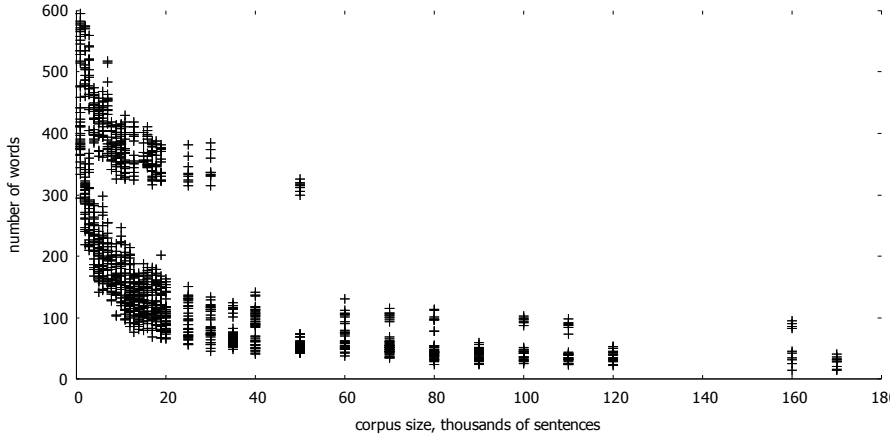

Fig. 3. Number of words tagged differently depending on the size of the training corpus

The increase in training corpus size causes the increasing number of context features hence the number of words with changed annotation grows (from 13 to 15%) and the number of ambiguous annotations left decreases (fig. 4). We should mention that 40% of words in the corpus were ambiguous.s

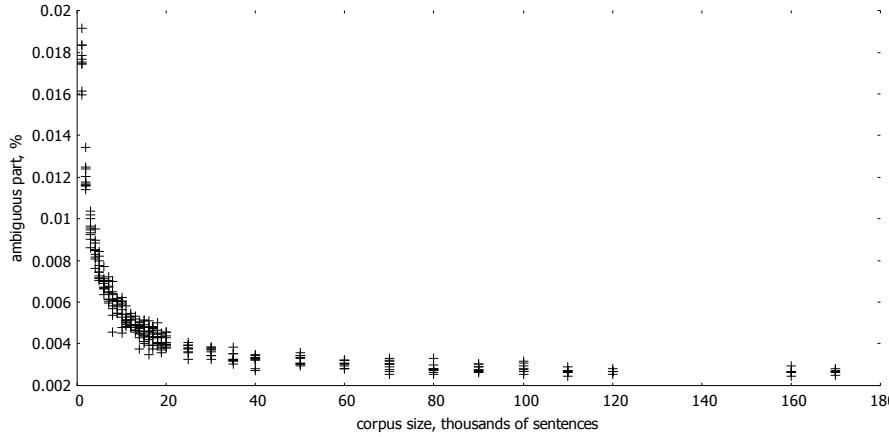

Fig. 4. Number of ambiguous annotations left

However, the growing number of context features does not always correspond to the increasing accuracy in disambiguation as it is shown below. We also compute the recall of our algorithm as fraction of unambiguous annotations in the test corpus. This metric shows the same results as those described above.

Most of ambiguous tags left are tags for related parts-of-speech (CONJ, INTJ, PRCL) which can appear in any context and tags which include these function words classes and are used for only several words such as *uzhe 'already'* — ADVB/COMP/NOUN/PRCL.

4.3. Accuracy

Tagging accuracy was measured on manually disambiguated 100 sentences selected from OpenCorpora project corpus. The test corpus includes sentences of different genres and syntactic structure. They were disambiguated with sets of rules described above and this annotation was compared with manually created one. Number of tagging mistakes decreases according to size of training corpora.

For each mistake an initial ambiguous tag is regarded as a mistake types. We took three most frequent mistakes for each list of rules and rank them according to the average number of their occurences. These major types of mistakes are shown below. Some of them are just mistakes in tagging one word (as *как* — ADVB/CONJ/NPRO), others are mistakes in tagging words that can appear almost in any context (CONJ/PRCL: *и*). Such cases as ADJF/NOUN are due to the limited set of context features: the rule *"ADJF NOUN -> ADJF | 1:tag=NOUN"* does not cover the situation when two adjectives are followed by a noun.

Table 1. Most frequent tagging mistakes

ADJF/NPRO	7.49417852523
CONJ/PRCL	6.26098191214
ADVB/CONJ	6.24516129032
PRCL/CONJ	6.17464424321

One of the major mistakes — tagging conjunction as adverb — is caused by tagging parenthesis as a conjunction, so that this type of ambiguity is heterogenous: out tagger mixes the adverbs used as parentheses with the words which really can be tagged either as conjunction or as an adverb (*kogda 'when', kak 'how', tak 'so'*).

Tagging accuracy is computed as a fraction of correct tags in test corpus (fig. 5). We suppose that the dispersion of results is due to the genre peculiarities of the training corpora and to the size of test corpus itself. However, we can see that stable reasonable results were obtained on the corpora bigger than 50K sentences. The highest average precision is observed on training corpora of 19–20 thousands of sentences. It should be noticed that tagger which chooses a random tag for each ambiguous word achieves accuracy about 93% in our case.

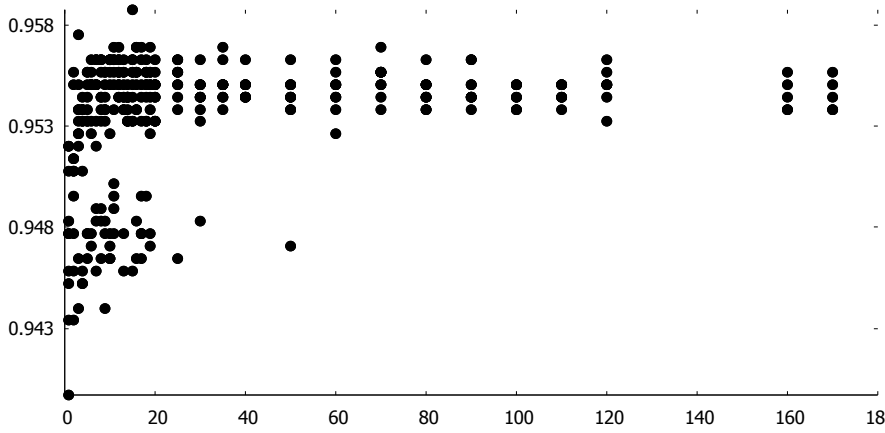

Fig. 5. Tagging accuracy

5. Conclusion

Taking into account the simplicity of the algorithm, we can say it has achieved quite reasonable accuracy and the performance of the system can be improved. We have shown that the morphological disambiguation task for Russian language can be solved almost without any special linguistic work using only corpora and morphological dictionary and our results are practically the same as those obtained in Brill's work. We have not come to a definite conclusion about the sufficient size of the training corpus, though several evaluations show that systems trained on corpora of 60K sentences and bigger achieve quite high results and produce practically the same number of rules.

There are several ways to improve our system. First of all, in this work we have used few context features. The context can be extended to four words (two left and two right neighbours) and the learner can take into account some more features including some lexical and grammatical categories. The algorithm should also be tuned to solve the ambiguity of word-forms (such as case ambiguity) as its supervised version was adapted for Polish morphological disambiguation. Another way for improvements is to study the influence of training corpus genre on the resulting set of rules. These improvements may require more linguistic knowledge which nevertheless cannot be compared to the task of creating fully manually annotated corpus.

References

1. *Acedański S. and Gołuchowski K.* A Morphosyntactic Rule-Based Brill Tagger for Polish. Recent Advances in Intelligent Information Systems, Kraków, Poland, 2009, pp. 67–76.
2. *Brill E.* Unsupervised Learning Of Disambiguation Rules For Part Of Speech Tagging. Proceedings of the Third Workshop on Very Large Corpora. Cambridge, Massachusetts, USA, 1995.
3. *Sharoff S., Nivre J.* 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, 2011.
4. *Sokirko A., Toldova S.* Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian,
available at http://aot.ru/docs/RusCorporaHMM.htm
5. *Zelenkov J., Segalovich I., Titov V.* Probabilistic model for morphological disambiguation based on normalising substitutions and adjacent words positions. [Verojatnostnaja model' cn'atija morfologicheskoj neodnoznachnosti na osnove normalizujushchih podstanovok i pozicij sosednih slov]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005". Zvenigorod, 2005.