МОРФОЛОГИЧЕСКАЯ РАЗМЕТКА КОРПУСА СИЛАМИ ВОЛОНТЁРОВ

Бочаров В. В. (bocharov@opencorpora.org),
Алексеева С. В. (alexeeva@opencorpora.org),
Грановский Д. В. (granovsky@opencorpora.org),
Протопопова Е. В. (protoev@gmail.com),
Степанова М. Е. (mariarusia@gmail.com),
Суриков А. В. (ksurent@opencorpora.org)

Проект «Открытый Корпус», OpenCorpora.org

Ключевые слова: корпус, разметка, морфологическая омонимия, краудсорсинг

CROWDSOURCING MORPHOLOGICAL ANNOTATION

Bocharov V. V. (bocharov@opencorpora.org),
Alexeeva S. V. (alexeeva@opencorpora.org),
Granovsky D. V. (granovsky@opencorpora.org),
Protopopova E. V. (protoev@gmail.com),
Stepanova M. E. (mariarusia@gmail.com),
Surikov A. V. (ksurent@opencorpora.org)

OpenCorpora.org

Manually annotated corpora are very important and very expensive resources: the annotation process requires a lot of time and skills. In OpenCorpora project we are trying to involve into annotation works native speakers with no special linguistic knowledge. In this paper we describe the way we organize our processes in order to maintain high quality of annotation and report on our preliminary results.

Key words: corpora, annotation, part of speech tagging, ambiguity, crowd-sourcing

1. Introduction

Corpora with manual annotation are required for testing and development of text analysis tools. In the OpenCorpora project we have already created a 1 million of words corpus[1] of Russian texts with human-verified words, sentences and paragraphs boundaries [2]. Morphology is the next level of annotation we are working on. We do this work in two steps: first of all each word gets all possible morphological hypothesis according to dictionary and later all wrong hypothesis are removed by human annotator. Handwork of linguists experts is expensive and we are trying to use native speakers with no linguistic knowledge as much as possible while maintaining high quality of annotation. It has been demonstrated that crowd-sourcing is a suitable method for obtaining linguistic data and "the quality is comparable to controlled laboratory experiments, and in some cases superior" [4]. We have involved several thousands of volunteers into annotation works by providing them with simple annotation questions. In each question we are asking about one grammatical category of one word within a sentence context. We have collected more than 1.1 million of answers[2]. In order to annotate 1 million of words about 4 millions of questions are to be asked.

2. Morphological annotation process

As we have stated before we use morphological dictionary (taken from AOT project [5] with some modifications in the tag set and complex cases' interpretations [1] to find all possible hypothesis for each word. No postprocessing or heuristics is applied to the set of hypothesis so we accept even very rare interpretations such as ИЗА (noun, feminine, plural, genitive case, personal name) for the word ИЗ. An example of our dictionary-based annotation is shown in Figure 1 (this way to display annotation is described in [1] and [3]).

Fig. 1. Dictionary-based annotation

[1] Statistics on corpus size is always up to date on page http://opencorpora.org/?page=genre_stats

[2] Statistics on contribution to corpus annotation is located on page http://opencorpora.org/?page=stats

The final goal is to have only one interpretation for each word for sentences with no syntactic or semantic ambiguity (as show in Figure 2). For ambiguous sentences several interpretations are allowed.

Мама	мыла	раму
v мама x	v мыть x	v рама x
СУЩ, од, жр, ед, им	ГЛ, несов, перех, жр, ед, прош, изъяв	СУЩ, неод, жр, ед, вн

Fig. 2. Unambiguous annotation

In OpenCorpora project the choice of the right morphological interpretation is done by hand by volunteers. In order to simplify this work we have splitted the annotation of each word into a set of simple annotation questions. Each question is asked about one grammatical category of one single word within a sentence context. In our example “Мама мыла раму” according to the set of hypothesis for the word МЫЛА following questions can be asked:

- is МЫЛА a verb or a noun?
- is МЫЛА singular or plural form of noun?
- is МЫЛА in nominative or accusative case?

This questions form a decision tree (Figure 3) where the next question is asked only in case it is meaningful after the previous answer. For the word МЫЛА in our example the correct answer for the first question is VERB and no other questions will be asked.

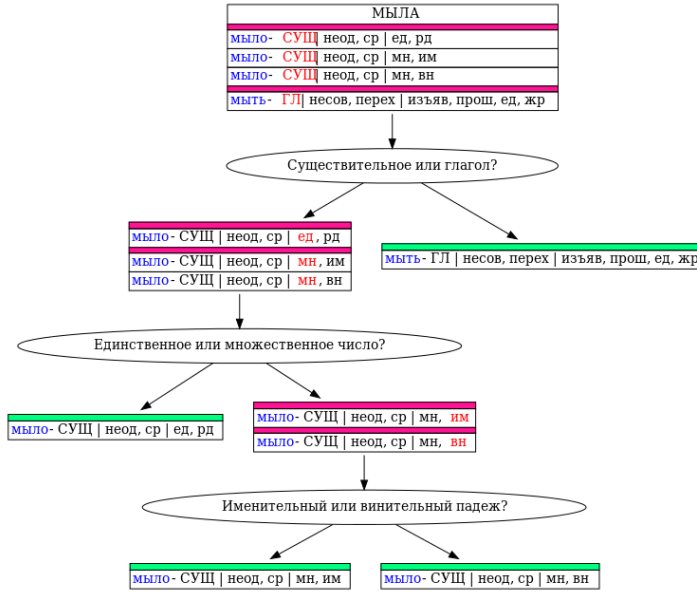

Fig. 3. Annotation decision tree for word МЫЛА

Annotation questions are grouped by type (i.e. “VERB or NOUN”, “singular or plural”, …) and volunteers can choose the specific type of questions they want. Each question type has its own instruction where general guidelines and the explanation of tricky cases are provided. The purpose of guidelines is to refresh background knowledge of linguistic categories and to specify issues which need interpretation different from one given in the secondary or high school. We don’t expect volunteers to know grammar perfectly. Instead we always ask them (both in guidelines and on web pages) to skip questions they don’t understand instead of making doubtful contribution.

3. Annotation quality estimation

Each question is answered by several (three or four) people and then it goes to moderation for approval. Moderators have a good linguistic background and they are able to make a correct decision. It will be very time-consuming to review and approve all answers. At first we have decided to do manual approval only for answers where there is some disagreement between volunteers or comments added. This decision was based on following calculations: let’s assume that all volunteers make random mistakes in 10% of answers (this is high error rate to simple questions like “singular or plural?”). Thus the probability of the event “all three annotators are wrong” is $0.1^3 = 0.001$ i.e. one annotation mistake per 1,000 words (0.1%) will be automatically approved if moderators will review only examples with disagreement.

In practice it turned out differently: an error rate for questions “is noun singular or plural?” is between 0.5% and 10% for most of volunteers and we have found 2% cases where all annotators were wrong. This means that our initial assumption of random error distribution isn’t true and the probability of an annotation error depends on the annotated word itself and on its context.

In order to find features that cause annotation errors we have splitted contexts into a set of simple context features. A context feature consists of position (0 is a position of word being annotated, −1 is one word to the left, +1 — one word to the right) and a word at that position. For each feature we have calculated the number of annotation disagreement events in examples with this feature. Following table includes top features ordered by percentage of disagreement events for questions of type “is noun singular or plural?”. In the rightmost column we show the expected error probability assuming that in case of disagreement between three annotators two of them are wrong (i.e. the worst case).

Table 1. Disagreement statistics for singular vs. plural disambiguation

Context feature	Position	Total samples	Samples with disagreement	Samples without disagreement	Disagreement rate	Expected error probability
word = четыре	−1	64	47	17	73.44%	48.96%
word = две	−1	136	89	47	65.44%	43.63%

Context feature	Position	Total samples	Samples with disagreement	Samples without disagreement	Disagreement rate	Expected error probability
word=три	−1	115	75	40	65.22%	43.48%
word = два	−1	93	60	33	64.52%	43.01%
word = две	−2	58	36	22	62.07%	41.38%
word = одна	4	13	8	5	61.54%	41.03%
word = две	0	226	135	91	59.73%	39.82%
word = копейки	0	17	10	7	58.82%	39.22%
word = четыре	–	95	55	40	57.89%	38.60%

This statistics reflect the norm of Russian grammar stating that the noun after the numeral ending in 1, 2, 3 or 4 must be in the singular. This is counterintuitive and most of people without linguistic knowledge make mistakes.

With these results we have decided to include into manual approval list for moderators all examples with context features provoking errors. The final list of such features will influence the overall precision of the annotation. In order to illustrate this we have plotted all context features occurring in questions of "is noun singular or plural?" type in the 2d space (Figure 4). The estimated error probability is on X-axis and the total number of examples is on Y-axis (logarithmic scale).

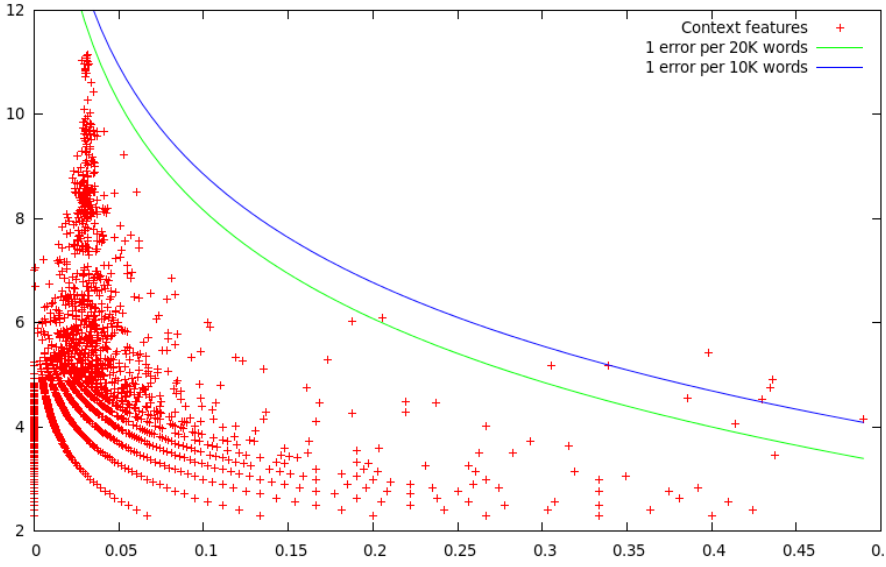

Fig. 4. Context features for "is noun singular or plural?" question type

Each dot on the plot corresponds to one context feature. Lines represent quality goals: the blue one — one error per 10K questions of this type (i.e. words), the green line — one error per 20K of question. Examples with context features above the line are to be included into manual approval list in order to meet quality goal chosen.

The feature with highest frequency is the pseudo-feature that is available in 100% of examples. The quality goal line that intersects with this feature denotes the highest possible annotation precision achievable with partial manual approval process. Better annotation requires all examples to be reviewed by people with expert knowledge in linguistics.

4. Conclusion

In this paper we have described our experience of crowd-sourcing morphological annotation in OpenCorpora project: the way annotation process is organized, our preliminary results and quality estimations technique based on disagreement rate between several annotators.

During the annotation process we collect not only annotation results but also the information about participants' interaction with user interface including timestamps of clicks on buttons. These data allow deeper analysis of both annotation and text understanding process. All the data we have collected are provided in the Download section on http://opencorpora.org and are licensed under the terms of Creative Commons CC-BY.

References

1. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M.* Quality assurance tools in the OpenCorpora project. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 101–109
2. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M. Surikov A.* Text segmentation in the OpenCorpora project. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]
3. *Bocharov V., Granovsky D.* Software tools for collaborative morphological markup [Programmnoje obespechenije dlja kollektivnoj taboty nad morfologicheskoj razmetkoj korpusa]. Trudy mezhdunarodnoj konferentsii "Korpusnaja lingvistika — 2011" [Corpus Linguistics — 2011: Proceedings of the International Conference]. Saint-Petersburg, 2011
4. *Munro R., Bethard S., Kuperman V.* et al. Crowdsourcing and language studies: the new generation of linguistic data // Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. — 2010.
5. *AOT*, available at http://www.aot.ru.