Learning Computational Linguistics through NLP Evaluation Events: the experience of Russian evaluation initiative

Anastasia Bonch-Osmolovskaya
National Research University
Higher School of Economics
101000, Myasnickaya, 20
Moscow, Russia
abonch@gmail.com

Olga Lyashevskaya
National Research University
Higher School of Economics,
101000, Myasnickaya, 20
Moscow, Russia
olesar@gmail.com

Svetlana Toldova
Moscow State University,
Faculty of Philology
119991, Leninskie Gory,
1 Hum. Bldg., Moscow, Russia
toldova@yandex.ru

Abstract

We present in the paper our experience of involving the students of the department of theoretical and computational linguistics of the Moscow State University into full-cycle activities of preparing and evaluating the results of the NLP Evaluation forums, held in 2010 and 2012 in Russia. The forum of 2010 started as a new initiative and was the first independent evaluation of morphology parsers for Russian in Russia. At the same time the forum campaign has been a source of a successful academic course which resulted in a close-knit student team, strong enough to implement the two-year research for the second forum on syntax, held in 2012. The new forum of anaphora (to be held in 2014) is now prepared mostly by students.

1 Introduction

Russian computational linguistics counts more than 50 years history, started with the first MT research in 1955 (Bar-Hilel 1960). Still up to the first decade of the 21 century all the research groups – those, inheriting the Soviet tradition, as well as the new commercial industry labs - existed in a disjoined mode. The absence of the state-of-the-art investigation on the performance of parsers for Russian as well as on the effect of different computational methods for Russian rich morphology impeded the teaching of computational linguistics, making it dilettantish. It's not surprising that the first initiative of the Evaluation forum emerged in the academy. The academic status of the initiative also guaranteed its independence. The complete cycle of the forum in 2010 on morphology, starting with mark-up scheme of the Gold Standard and ending the final paper preparation has served as a basis for a course in computational linguistics with excellent set of tasks for students to carry out. The problem of the first year experience was insufficient communication with all the participants during the forum preparation. This is very important for the pioneer status of the forum and also the educational perspective of the initiative. That's why the two year period of forum preparation has been chosen. The task of the first year is to prepare and hold a round-table open to all the potential participants where the basic decisions on the test collections, tasks, mark-up and evaluation process are made. The task of the second year is the evaluation forum itself and the preparation of an overview paper. Below we will focus on the educational process: we will describe student tasks during the complete cycle of the evaluation forum preparation. The consistent practical aim of the course distinguishes in from most of the courses in computational linguistics (Hearst, 2005; Liddy and McCracken, 2005; Baldridge and Erk, 2008). This is a course in NLP evaluation which, as we believe, gives students very useful theoretical and practical skills of making sound and deliberate decisions during linguistic data analyses. The main idea of the course is to involve students into solving "real-life" expert tasks, and to show them multiple approaches to mark-up and data analysis. We would like to underline that the practical value of the course: students not only do the routine assessment procedure, but analyze the best practices and create the design of the forum. The course is organized as follows: students complete tasks at home and discuss the results at class with two

Proceedings of the Fourth Workshop on Teaching Natural Language Processing, pages 61–65,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

or three instructors. The experienced students may act as instructors also. The class ends by collective presentation at the conference. Students work in small teams of 2 or 3 persons, each team doing its piece of work. All the students have strong background in theoretical linguistics and math, some students have good programming skills. The main stages of the first year are: 1) getting theoretical background 2) first mark-up experience and proto-gold standard 3) feedback from the participants 4) round-table preparation. The second year consists of the following stage: 5) preparing Gold Standard 6) results evaluation 7) final paper preparation. These stages correspond to the four semesters of special courses on NLP, home task activities, hands-on student activities and practice in academic writing. Each of the stage will be discussed below. The corresponding teaching methods are described in a separate section.

2 Background task

The first task students have to complete is to study theoretical background which consists of a) actual evaluation practices b) state of art of Russian NLP systems that can potentially participate in the forum. Primarily students study reports of the main evaluation forums that have been held on the current task. The topics to be discussed in class are: the types of system running the competitions (statistical, rule-based, hybrid), their theoretical linguistic basis: for example, HPSG parsers or dependency parsers for syntax; the test collections, their sources, size and mark-up scheme; the tasks and their metrics; the performance rate. The students have to find the answers on all this questions making their way through exhaustive reports, they have to draw out some common grounds to be compared and analyzed. For example for the syntax forum (Gareyshina et al., 2012) tree-banks of different languages and structure types has been analyzed and compared. The very important point of this stage is that it results in collective determining some ideal scenario of the future forum which is to be inevitably corrected by performing the second investigation – examining all the information about the potential participants, such as collecting and reading all the related papers, testing demos or installing the open-source resources. For example, the main problem for the morphology forum was to determine a mark-up scheme that would be convenient for all the participants (Lyasevskaya et al., 2010). This problem is crucial because of Russian rich morphology and the variety of theoretical traditions different systems rest upon. The investigation of syntactic parsing (all the systems that took part in the forum, use dependency parsing) revealed the impossibility to compare the types of syntactic relations specified by different systems. The fact is not surprising bearing in mind that there is no open tree-bank such as Penn tree bank to be trained on for Russian. The workshop devoted to comparing different syntactic parsing outputs has been exhausting but fruitful: we arrived to a decision that the main task of the forum should include only evaluating what syntactic heads were to be marked by the participants. Correctness of parsing the whole sentence was decided to count as irrelevant. Only the choice of the head was evaluated. We would like to underline that the design and the scenario of the forums are always determined as a result of individual work of student groups together with collective analysis and summing-up conclusions. Finally the last but not the least object of this task is to juxtapose theoretical and computational linguistics: students have to analyze the scope of underlining linguistic phenomena and to compare them with applied realizations in NLP. The more sophisticated linguistic task is in focus, the more interesting topics are raised in class. For example, the examination of different principles of anaphoric resolution this year showed the limits of applied tasks and solutions (particularly in discourse anaphora resolution and identifying lexical coherence determined extralinguistically), and revealed the perspectives of future development in NLP and artificial intelligence. The analysis is then partly fulfilled in Gold Standard mark-up. The scheme is always broader then it has to be for the evaluation task. The important additional outcome of such corpus mark-up is to prepare some new open resource that can serve also for corpus linguistic and theoretical linguistic research.

3 First mark-up experience and first feedback

As it has been noted earlier the theoretical stage of the course results in the forum scenario and the mark-up scheme for the Gold Standard. At the next stage students begin by making mark-up on a few selected texts. Each text is marked-up with several students and all the cases of interannotator discrepancy have to be analyzed and discussed in class. The discussion leads to formulating more distinct mark-up criteria as well as to determining the cases which should not be evaluated. The mark-up is made by special tool programmed by the students with good programming skills. The specification of requirements for the tool is also the task to be performed by students. The first mark-up staging is all in one testing the mark-up scheme, elaboration of the evaluation framework and metrics as well as technical testing of the tool. As a result some small (usually 100 sentences) "pre-gold" standard is made. Then these sentences (both a non-marked and a marked-up variant) are sent to the participants who had by this time made a claim on their participation in the forum. The idea is to get preliminary feedback to control all the previous decisions that have been made about the forum during the theoretical stage of the course. The participants have the possibility to estimate the mark-up scheme and the assessment scheme and present some on-going results of this first small test.

When we receive the first feedback from the participants, we turn to the analysis of the system possible mistakes. Our aim at this stage is not to evaluate the systems but to exclude all cases which are either theoretically unclear (i.e. the head of the conjunction group) or cannot be resolved by the system (a "boy sees the girl with the telescope" problem) or too difficult to unify (i.e. choice of the basic infinitive for Russian aspectual verbal pairs).

All this activities need special clarification: Russian is a so called "poor resource" language. The forum cannot use existing corpora as a training set. This can violate the independence of evaluation results: some of the system had been trained on these corpora while others had not. So the main practice of our evaluation forums is to conduct assessment on a Gold Standard subcorpus which normally includes about 800 randomly selected sentences that have been manually tagged. Meanwhile the routine of manual tagging serves as an important practical exercise for students.

4 The round-table

The closing event of the first year is a round-table, held at the annual conference on computational linguistics "Dialogue" (www.dialogue-21.ru). The presentation is prepared and done mostly by students and contains all the topics that had been worked on during the previous period: all important background, proposals on the forum scenario and the result of the first evaluation experiment. Usually most of the participants take active part in the round-table. This is besides all an exciting experience for students that have an opportunity to make acquaintance with researches from academy and industry, the opportunity that can have far-reaching effect for their future career. After the round table the work on the second part – the evaluation itself begins.

5 The Gold Standard mark-up stage

The Gold Standard preparation stage includes: the final version of annotator instruction work-out, the tool for Gold Standard mark-up choice or creation, Gold Standard annotators disagreement cases discussion, the final version of Gold Standard creation.

For the Syntax and Anaphora forum the special tools were created for Gold Standard Mark-up. These tools are suitable for annotators decision comparison (Gareyshina et al., 2012). The design of the tool was a special issue for discussion during the class.

The Gold Standard is tagged manually using the worked-out tagging tool. Each item (word, sentence, text (coreference chain)) is independently tagged by two experts-students, then divergences are discussed, if any, and the common decision is made. Each pair of students is responsible for the common decision in case of discrepancy. The discrepancies in pairs are written out in a special document. The students finally work out the list of problematic cases for the corresponding NLP tasks both from the point of view of theory and practical decisions, e.g. the typical morphological ambiguity cases such as Verbal Adjective vs. Participle for Russian or problems of Syntactic relation direction in case of Numeral-Noun syntactic relation, etc. The cases are discussed during seminars. Thus the annotator instruction is improved. Then the annotation is checked by the third expert (one of the tutors). Such procedure allowed us to achieve three aims. It helped to work out the algorithm for semi-automate annotators' mistakes detection procedure. Then, we wanted to avoid 'overfitting': getting the experts used to common error of the specific system and omitting errors by not noticing them. And last, tagging is supposed to give the experts the basic knowledge about difficult cases and to help them form criteria for evaluating mismatches.

6 The evaluation procedure

The stage of evaluation includes the creation a special tool for systems responses comparison with Gold Standard, the comparison of the output of the parsers to the Gold Standard.

The test sets usually are based on a Treebank used for the development of the parsers. In our case there was no Gold Standard Treebank for Russian and there is no Gold Standard Corpora with coreference mark-up. Moreover each system has its own theoretical and practical decisions due to the final purposes of the system.

The students' activity during this stage includes: the automatic comparison tool creation (this is a task for a "programming-oriented" students), the special editor for system responses comparison creation, the manual procedure of system mismatches with Gold Standard analysis.

The latter is an essential stage for Evaluation. As it was mentioned above there are systems' mismatches that should not be treated as mistakes. Thus this procedure includes the collective decision for a repertory of marks used by the annotators for differentiating cases of mismatches, the mismatches discussion during joint seminars, the mismatches manual assessment. All teams of assessors (two students and a tutor) have their own piece of a Gold Standard Corpora to check. Thus every team faces all kinds of difficulties; this principle provides the united consistent approach to all the types of discrepancies.

7 Teaching Methods and Schedule

The Forum cycle takes one and a half of academic years. Thus we have a series of three Special seminars in one of the NLP fields. Students could take part in all the stages of a Forum or only in one of them. The first part is mainly theoretical. They deepen their knowledge in theoretical approaches to linguistic analysis; get acquainted with the approaches to the corresponding NLP task. The other useful activities is a NLP software testing, the real systems discrepancy analysis. The course is also good opportunity to train academic reading skill. The comparison of systems outputs and the work out of Forum parameters are good hands-on tasks. This course is also a challenge for students to learn out how the theoretical principles interact with practical system requirements.

The second course is a practical one. Its primary aim is to work out and annotate the Gold Standard Corpus. Thus this activity could be treated as a series of hands-on in classroom together with exhaustive home-tasks. The course is a project work in a team where IT-oriented students and linguistically-oriented students work together. The practical result is an opened resource such as Syntax Treebank consisting of 800 sentences manually tagged. One of the important educational outputs of the seminar is the acquaintance with the repertory of the problematic cases in a certain NLP field of study.

The Third course is also practical one. Besides the practical tasks of Systems mismatches evaluation this course also allows students to improve their Academic writing skills. The output of this course is not only the Systems evaluation as it is but a scientific article describing the whole Forum procedure as well.

8 Conclusions

The described above students activity as the organizers of the Evaluation Forum, annotators and assessors has challenges for NLP education the enumerated below.

The «outputs» for theoretical stage are the following:

- the high-targeted, and thus highly motivated and deep acquaintance with the approaches to the NLP tasks, existing resources in other languages, methods of evaluation;
- the academic reading skills in NLP research field;
- the acquaintance with the different principle of adaptation the linguistic theory to the NLP task implementation.

The practical-skill training output:

- the annotation skill
- the academic reading and writing skill
- the NLP evaluation skill
- the inter-discipline team-working.

As it has been mentioned, ironically, the resource poverty is a challenge for NLP education with Russian language in focus. At start the procedure of a particular NLP evaluation task for Russian is a terra incognita. Before the Forum starts the number and entry list of participants (and thus the competing technologies) are not predictable. Doing something, that nobody has done before, is always a superb motivation for student involvement.

References

Baldridge, Jason, and Katrin Erk. 2008. Teaching computational linguistics to a large, diverse student body: courses, tools, and interdepartmental interaction. *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics.* P. 1-8. Association for Computational Linguistics Stroudsburg, PA, USA.

Bar-Hillel, Yehoshua. 1960. The present status of automatic translation of languages. Advances in computers 1, no. 1 P. 91-163..

Hearst, Marti. 2005. Teaching applied natural language processing: Triumphs and tribulations. Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. P. 1-8. Ann Arbor, Michigan, June. Association for Computational Linguistics.

Liddy, Elizabeth D., and Nancy J. McCracken. 2005. Hands-on NLP for an interdisciplinary audience. *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.* P. 62-28. Association for Computational Linguistics.

Gareyshina Anastasia, Ionov Maxim, Lyashevskaya Olga, Privoznov Dmitry, Sokolova Elena, Toldova Svetlana. 2012. *RU-EVAL-2012: Evaluating Dependency Parsers for Russian.* Proceedings of COLING 2012: Posters. P. 349-360. URL: http://www.aclweb.org/anthology/C12-2035.

Lasevskaya Olga, Astaf'eva Irina, Bonch-Osmolovska ya Anastasia, Gareyshina Anastasia, Grishina Julia, D'jachkov Vadim, Ionov Maxim, Koroleva Anna, Kudrinsky Maxim, Lityagina Anna, Luchina Elena, Sidorova Evgenia, Toldova Svetlana, Savchuk Svetlana., Koval' Sergej. 2010. Evaluation of the automated text analysis: POS-tagging for Russian. [Morphological Ananlysis Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka.] *Proceedings of the International Conference on Computational Linguistics Dialogue-2010.* P. 318-327.