Learning from the Linguists: Combining Annotation Tools and Machine Learning

Łukasz Kobyliński
Institute of Computer Science, Polish Academy of Sciences
L.Kobylinski@ipipan.waw.pl

Introduction

Supervised methods in Machine Learning (ML) are known to consistently perform better in most applications than unsupervised and semi-supervised approaches. Large sets of representative training data are however needed for such methods to be useful in a general case.

The usual approach to ML in linguistic engineering is to train supervised methods on the basis of annotated textual resources, such as language corpora.

What is missed in such a scheme of training automated methods is the process of annotation itself and the context in which a linguist makes her or his decisions.

We propose to train ML methods using not only the annotation layer, which is the final result of the work of a linguist, but also taking into account the meta-information from the annotation tool.

1. Learning based on Annotation

The most common way of using supervised learning approaches to linguistic problems is to train such methods on previously annotated text resources. Such resources are prepared using annotation tools by qualified linguists.

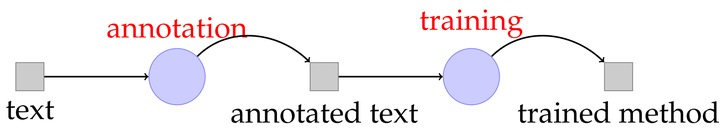

Figure 1. The usual approach to creating resources for supervised learning methods in linguistic engineering.

Such gold-standard resources are usually prepared by performing a parallel annotation of the resource by at least two linguists. The **Inter-Annotator Agreement (ITA)** may be calculated afterwards, providing information about the inherent difficulty of the problem.

Tools Following that scheme, we have developed a tool called **AnotEk** during the course of work on the task of **Word Sense Disambiguation** in Polish language texts. It is a web-based tool, allowing for simultaneous work of many linguists on the same (or different) parts of a textual resource (e.g. a corpus).

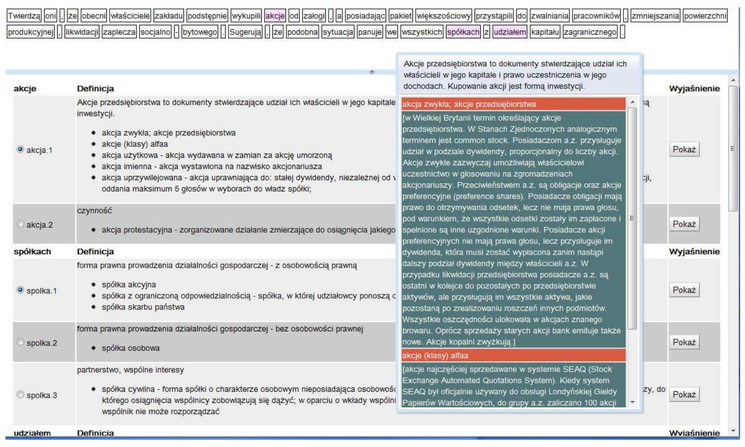

Figure 2. AnotEk: A web-based tool for creating the sense annotation layer of a text corpus.

The results of annotation, conflict resolution and additional comments are maintained in a database and may be exported to a desired format (e.g. TEI XML format).

Resources We have collected a Wikipedia-based corpus from the domain of economy during the work on the WSD task. **Word-sense annotation** layer has been manually created for this corpus and two other previously existing corpora: a collection of stock market reports and a subcorpus of the National Corpus of Polish (NCP).

Corpus	Segments	Annotated segments
NCP subcorpus	87 816	3 821
Stock market	282 366	18 719
Wiki-econo	408 221	23 269
Overall	778 403	45 809

2. Learning based on the Process of Annotation

We argue that there is valuable information not only in the result of the work of human experts, but also in the meta-data collected during their work. We aim to supply machine learning methods with additional data, which may be used to differentiate difficult cases. Based on the example of a project concerned with **correcting corpus annotation errors**, the types of additional information that may influence the training phase of ML methods include:

- search query used to find an occurrence of an annotation mistake,
- number of corrections made to the segment,
- additional comments attached to the segment or the enclosing paragraph,
- the name of the annotator.

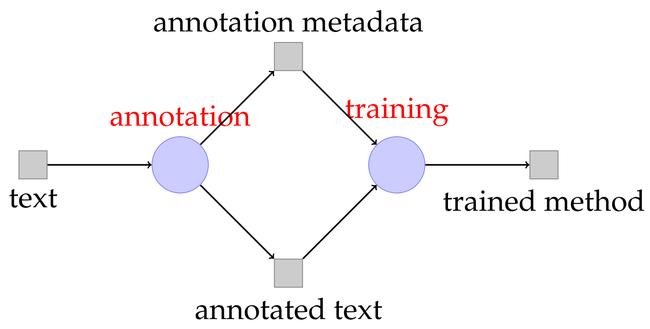

Figure 3. Including annotation meta-data in training supervised learning methods.

Tools We have developed **CorpCor**, a tool designed for correcting annotation errors in a text corpus. This web-based application may be used simultaneously by multiple linguists and allows them to search for any text fragment (using Poliqarp notation) and modify the annotation of particular segments.

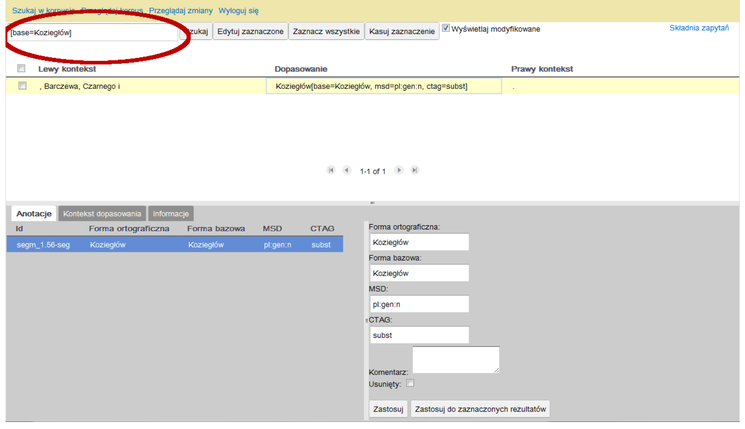

Figure 4. CorpCor: A web-based tool for editing corpus annotation. A linguist performs a query to find possible mistakes in the morphosyntactic layer of annotation.

Each modification to the annotation layer is saved in a history of edits and includes such meta-data as the search query used to find the corrected mistake and additional comments provided by the lignuist.

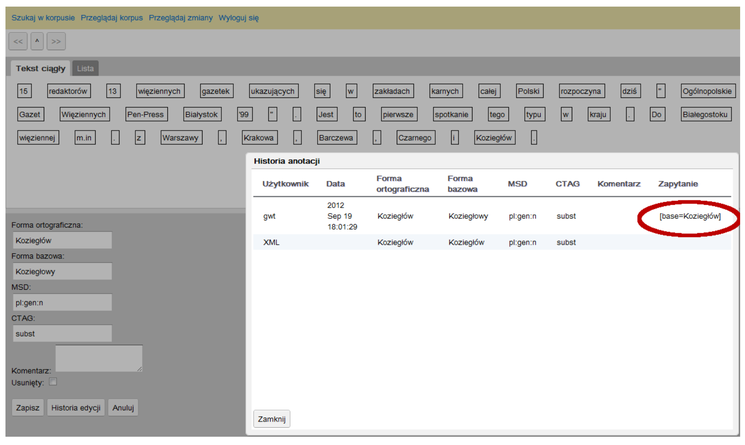

Figure 5. CorpCor: A history of edits is maintained in the system and each edit is associated with a search query performed by the linguist to find the corrected mistake.

Resources A corrected version of National Corpus of Polish will be the final result of project.

3. Using Machine Learning for Annotator Cues

Future Work In future, we would like to integrate the machine learning approach to correcting corpus annotation with the tools developed for manual work. In such a scenario the suggestions provided by the automated method may help human annotators when manually correcting mistakes in the annotation. There are at least two distinct possibilities of such an integration:

- provide an interface between the web-based tool and a machine learning method trained on a previously annotated subcorpus,
- use a machine learning method, which can be trained incrementally and modify its behavior on-line.

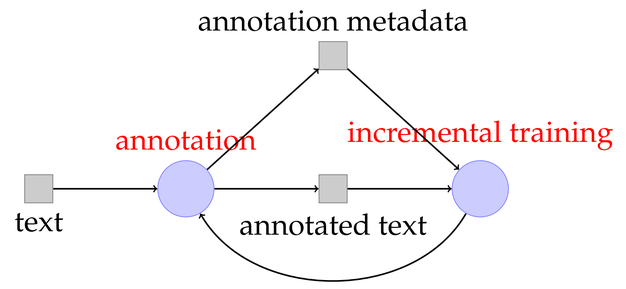

Figure 6. Providing suggestions for human annotators using predictions based on previously annotated text fragments.

The interface of the web-based tool may be easily modified to suggest a most probable selection among possible annotation values.

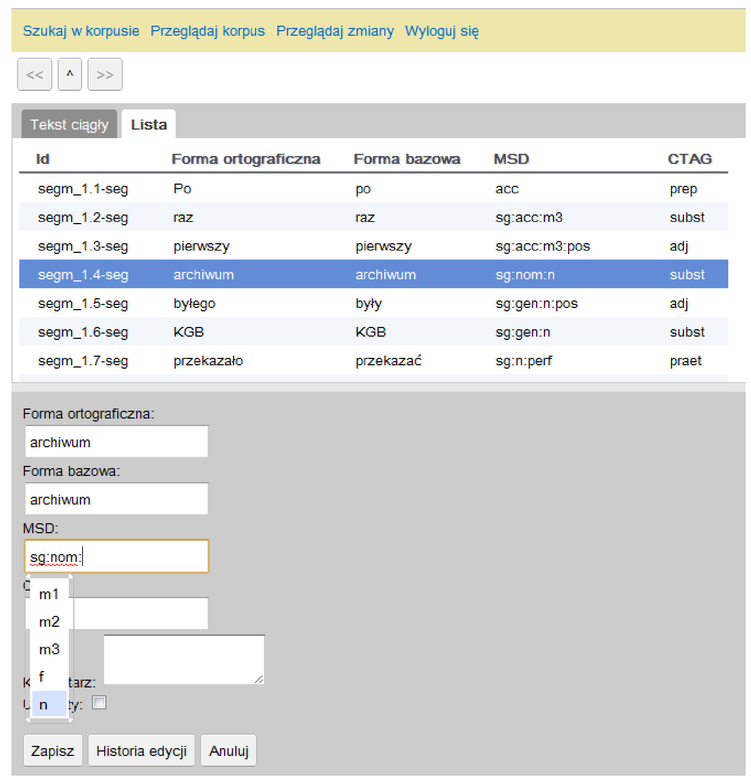

Figure 7. Planned in CorpCor: Recommended annotation may be suggested on the basis of earlier annotations.

Projects Involved

- CESAR - CEntral and South-east europeAn Resources,
- NEKST (An adaptive system to support problem-solving on the basis of document collections in the Internet), a national Ministry of Science and Higher Education Innovative Economy Operational Programme (PO IG) grant,
- Automatic detection and correction of annotation errors in Polish language corpora, a National Science Centre research grant.