

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.

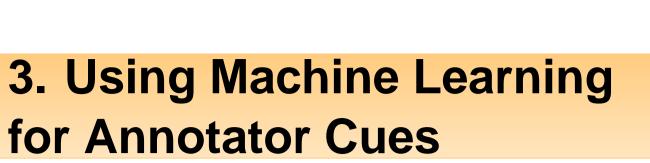
The usual approach to ML in linguistic engineering is to train supervised methods on the basis of annotated textual resources, such as language corpora.

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 metainformation from the annotation tool.

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,



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



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.

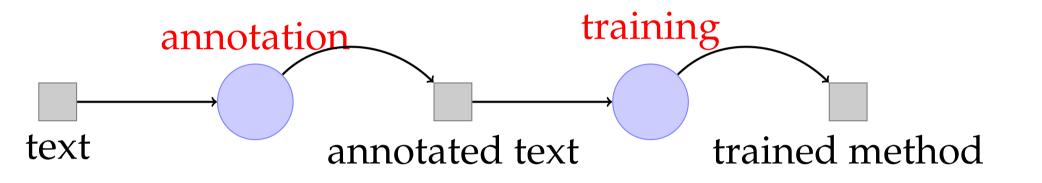


Figure 1. The usual approach to creating resources for supervised learning methods in linguistic engineering.

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.

• number of corrections made to the segment, • additional comments attached to the segment or the enclosing paragraph, • the name of the annotator.

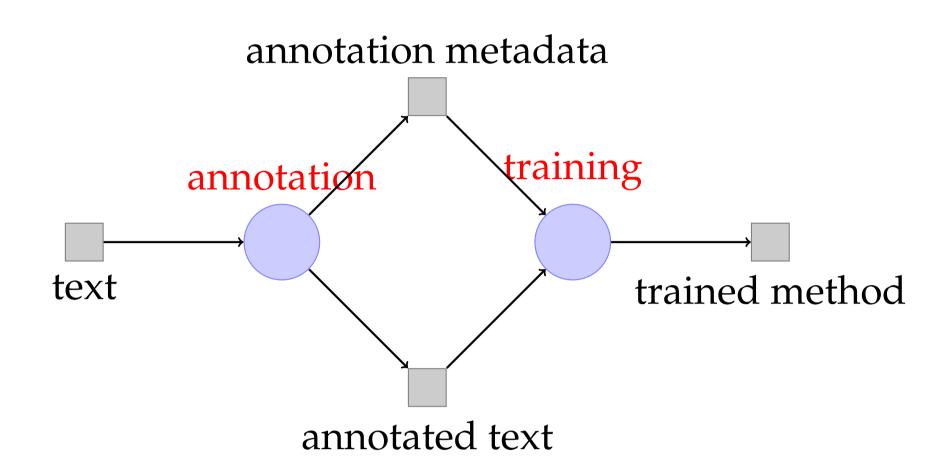


Figure 3. Including annotation meta-data in training supervised learning methods.

Tools We have developed **CorpCor**, a tool designed for correcting annotation errors in a text corpus. This webbased 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.

Szukaj w korowsie - Przeglądaj korows - Przeglądaj zmiany - Wyloguj się							
[base=Koziegłów]	zukaj	Edytuj zaznaczone	Zaznacz wszystkie	Kasuj zaznaczenie	Wyświetlaj modyfikowane	s	kładnia zapy

incrementally and modify its behavior on-line.

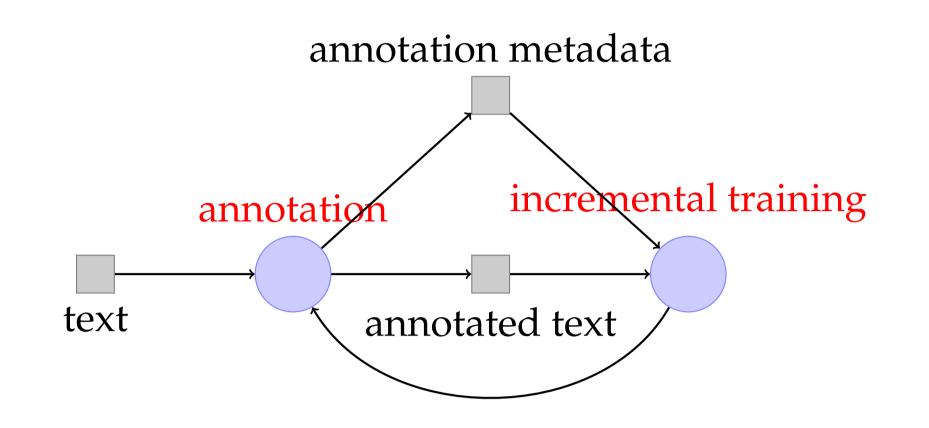


Figure 6. Providing suggestions for human annotators using predictions based on previously annotated text fragments.

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

Szukaj w korpusie Przeglądaj korpus Przeglądaj zmiany Wyloguj się

<<	•	>>	

Tekst ciągły	а							
ld	Forma ortograficzna	Forma bazowa	MSD	CTAG				
segm_1.1-seg	Po	ро	acc	prep				
segm_1.2-seg	raz	raz	sg:acc:m3	subst				
segm_1.3-seg	pierwszy	pierwszy	sg:acc:m3:pos	adj				
segm_1.4-seg	archiwum	archiwum	sg:nom:n	subst				
segm_1.5-seg	byłego	były	sg:gen:n:pos	adj				
segm_1.6-seg	KGB	KGB	sg:gen:n	subst				
segm_1.7-seg	przekazało	przekazać	sg:n:perf	praet				
Forma bazowa: archiwum								
MSD:								
sg:nom: m1								
m2 m3 k f tarz:								
K ' tarz: U n ty: 🗆								
Zapisz Historia ed	dycji Anuluj							

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).

Akcje przedsiębiorstwa to dokumenty stwierdzające udział ich właścicieli w jego kapital akcja zwykła; akcje przedsiębiorstwa • akcja zwykła; akcje przedsiębiorstwa • akcja zwykła; akcje przedsiębiorstwa Iw Welkiej Brytanii termin okraślający akcje • akcja użytkowa - akcja wydawana w zamian za akcję umorzoną przedsiębiorstwa. W Stanach Zjednoczorych analogicznym terminem jest common stock. Posiadaczem a z. przysługuje udział w podziale dywidendy, proporcjonalny do liczby akcji. • akcja użytkowa - akcja wystawiona na nazwisko akcjonariusza Akcje zwykłe zazwyczaj umozliwiają właścicielowi udział w podziale dywideni w syborach do władz spółki; • akcja protestacyjna - zorganizowane działanie zmierzające do osiągnięcia jakiego • akcja protestacyjna - zorganizowane działanie zmierzające do osiągnięcia jakiego • akcja protestacyjna - zorganizowane działanie zmierzające do osiągnięcia jakiego • przedsiębiorstwe owarunki. Posiadacze akcji preferencyjnych nie mają prawa głosu, lecz przysługuje im dywidenda, która musi zostać wybacona zanim nastąpi dalszy podział dywidendy między Maścicieli a z. W przypadku likwidacji przedsiębiorstwe a z. są obligacji mają prawa głosu, lecz przysługuje im dywidenda, która musi zostać wybacia działaniości gospodarczej - z osobowością prawną prowadzenia działalności gospodarczej - z osobowością prawną przypadku likwidacji przedsiębiorstwe a z. są obligacji przedsiębiorstwe a stywa, jakie pozystkie aktywa, jakie pozystkie aktywa, jakie pozystkie aktywa, jakie przedsiębiorstwe a z. są obligaci prawną prowadzenia działalności gospodarczej - bez osobowości prawną prowadzenia działałanie zmierzające do zosotawo przesiębiorstwe a stywoka, akcje kopajin zwystwie, jakcje pozediności w posiadacze a	1			Wyjaśnienie
• akcja.2 • akcja protestacyjna - zorganizowane działanie zmierzające do osiągnięcia jakiego pod warunkiem, że wszystkie odsetki zostały im zapłacone i spełnione są inne uzgodnione warunki. Posiadacze akcji preferencyjnych nie mają prawa głosu, lecz przysługuje im dywidenda, która musi zostać wypłacona zanim nastąpi dalszy podział dywidendy między właścicieli a.z. W przypadku likwidacji przedsiębiorstwa posiadacze a.z. są ostatni w kolejce do pozostałych po przedsiębiorstwie ostatni w kolejce do pozostałych po przedsiębiorstwie aktywa, jakie pozostaną po zrealizowaniu roszczeń innych podmiotów. Wszystkie oszczędności ulokowała w akcjach znanego browaru. Oprócz sprzedaży starych akcji bank emituje także nowe. Akcje kopalni zwyżkują.] • spółka osobowa partnerstwo, wspólne interesy forma prawna prowadzenia działalności gospodarczej - bez osobowości prawnej pod warunkiem, że wszystkie odsetki zostały im zapłacone i spełnione są inne uzgodnione warunki. Posiadacze akcji preferencyjnych nie mają prawa głosu, lecz przysługuje im dywidenda, która musi zostać wypłacona zanim nastąpi dalszy podział dywidendy między właścicieli a.z. W przypadku likwidacji przedsiębiorstwa posiadacze a.z. są ostatni w kolejce do pozostałych po przedsiębiorstwie aktywa, jakie pozostaną po zrealizowaniu roszczeń innych podmiotów. Wszystkie oszczędności ulokowała w akcjach znanego browaru. Oprócz sprzedaży starych akcji bank emituje także nowe. Akcje kopalni zwyżkują.] • spółka osobowa [akcje najczęściej sprzedawane w systemie SEAQ (Stock Exchange Automated Quotations System). Kiedy system	Con .	• na		Pokaż
Spółkach Definicja dywidenda, która musi zostać wypłacona zanim nastąpi Image: Spółka akcyjna forma prawna prowadzenia działalności gospodarczej - z osobowością prawną dalszy podział dywidendy między właścicieli a.z. W Image: Spółka akcyjna spółka akcyjna ostatni w kolejce do pozostałych po przedsiębiorstwia posiadacze a.z. są Image: Spółka z ograniczoną odpowiedzialnością - spółka, w której udziałowcy ponoszą c spółka skarbu państwa sktywów, ale przysługują im wszystkie aktywa, jakie Image: Spółka skarbu państwa pozostaną po zrealizowaniu roszczeń innych podmiotów. Wszystkie oszczędności ulokowała w akcjach znanego Image: Spółka osobowa forma prawna prowadzenia działalności gospodarczej - bez osobowości prawnej mowe. Akcje kopalni zwyżkują.] Image: Spółka osobowa spółka osobowa partnerstwo, wspólne interesy akcje (klasy) alfaa Image: Spółka osobowa gakcje najczęściej sprzedawane w systemie SEAQ (Stock Exchange Automated Quotations System). Kiedy system	ш			Pokaż
 spółka akcyjna spółka akcyjna spółka z ograniczoną odpowiedzialnością - spółka, w której udziałowcy ponoszą c spółka skarbu państwa forma prawna prowadzenia działalności gospodarczej - bez osobowości prawnej spółka osobowa spółka osobowa partnerstwo, wspólne interesy partnerstwo, wspólne interesy kiedy system 				Wyjaśnienie
• spółka osobowa nowe. Akcje kopalni zwyżkują.] • partnerstwo, wspólne interesy akcje (klasy) alfaa [akcje najczęściej sprzedawane w systemie SEAQ (Stock Exchange Automated Quotations System). Kiedy system				Pokaż
akcje (klasy) alfaa partnerstwo, wspólne interesy [akcje najczęściej sprzedawane w systemie SEAQ (Stock Exchange Automated Quotations System). Kiedy system				Pokaż
Exchange Automated Quotations System). Kiedy system				
 spółka cywilna - forma spółki o charakterze osobowym nieposiadająca osobowość którego osiągnięcia wspólnicy zobowiązują się dążyć; w oparciu o wkłady wspólni wspólnik nie może rozporządzać SEAQ był oficjalnie używany do obsługi Londyńskiej Giełdy Papierów Wartościowych, do grupy a.z. zaliczano 100 akcji 	•	z	zy, do	Pokaż

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).

Lewy kontekst				owanie		Prawy kontekst	
, Barczewa,	arczewa, Czarnego i		Kozieg	łów[base=Kozieg	łów, msd=pl:gen:n, ctag=subst]		
					1-1 of 1 🕩 🖲		
notacje Kont	ekst dopasowania 🛛 Informa	cje			-		
	Forma ortograficzna	Forma bazowa	MSD	CTAG	Forma ortograficzna:		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Koziegłów		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Koziegłów Forma bazowa:		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst			
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa:		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD:		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD: pl:gen:n		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD: pl:gen:n IICTAG:		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD: pl:gen:n ICTAG: subst		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD: pl:gen:n ICTAG: subst		
gm_1.56-seg	Koziegłów	Koziegłów	pl:gen:n	subst	Forma bazowa: Koziegłów MSD: pl:gen:n ICTAG: subst		

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.

Szukaj w korpusie Przeglądaj korpus Przeglądaj zmiany Wyloguj się	
Tekst ciągły Lista	
15 redaktorów 13 więziennych gazetek ukazujących się w zakładach karnych całej Polski rozpoczyna dziś	" Ogólnopolskie
Gazet Więziennych Pen-Press Białystok '99 " . Jest to pierwsze spotkanie tego typu w kraju . [Do Białegostoku
więziennej m.in . z Warszawy , Krakowa , Barczewa , Czarnego i Koziegłów .	
Historia anotacji	
Użytkownik Data Forma Forma MSD CTAG Komentarz	Zapytanie

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 problemsolving on the basis of document collections in the Internet), a national Ministry of Science and Higher Education Innovative Economy Operational Programme

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			

Forma ortograficzna:	gwt	2012 Sep 19 18:01:29	Koziegłów	Koziegłowy	pl:gen:n	subst	[base=Koziegłów]
Koziegłów	XML		Koziegłów	Koziegłów	pl:gen:n	subst	
Forma bazowa:							
Koziegłowy							
MSD:							
pl:gen:n							
CTAG:							
subst							
Komentarz:							
Usunięty:							
Zapisz Historia edycji Anuluj	Zamknij						

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.

(PO IG) grant,

• Automatic detection and correction of annotation errors in Polish language corpora, a National Science Centre research grant.