TaCo (tagset converter)

Bartosz Zaborowski

b.zaborowski@ipipan.waw.pl Institute of Computer Science, PAS

Motivation

problem:

- Annotated corpora use various types of tags to encode information along words, like POS-tags or more complex morphosyntactic information.
- The set of tags usually differs between corpora, often even for corpora of the same language.
- Various NLP tools often are tied to a specific tagset.
- Manual (re-)annotation of corpus in different tagset is expensive.
- Automatic (re-)tagging using regular tagger doesn't produce high quality results, regardless of the quality of previous annotation.

solution: TaCo

An automatic method of re-annotation similar to some kinds of taggers, but making use of existing annotation as a source of valuable information.

Method overview

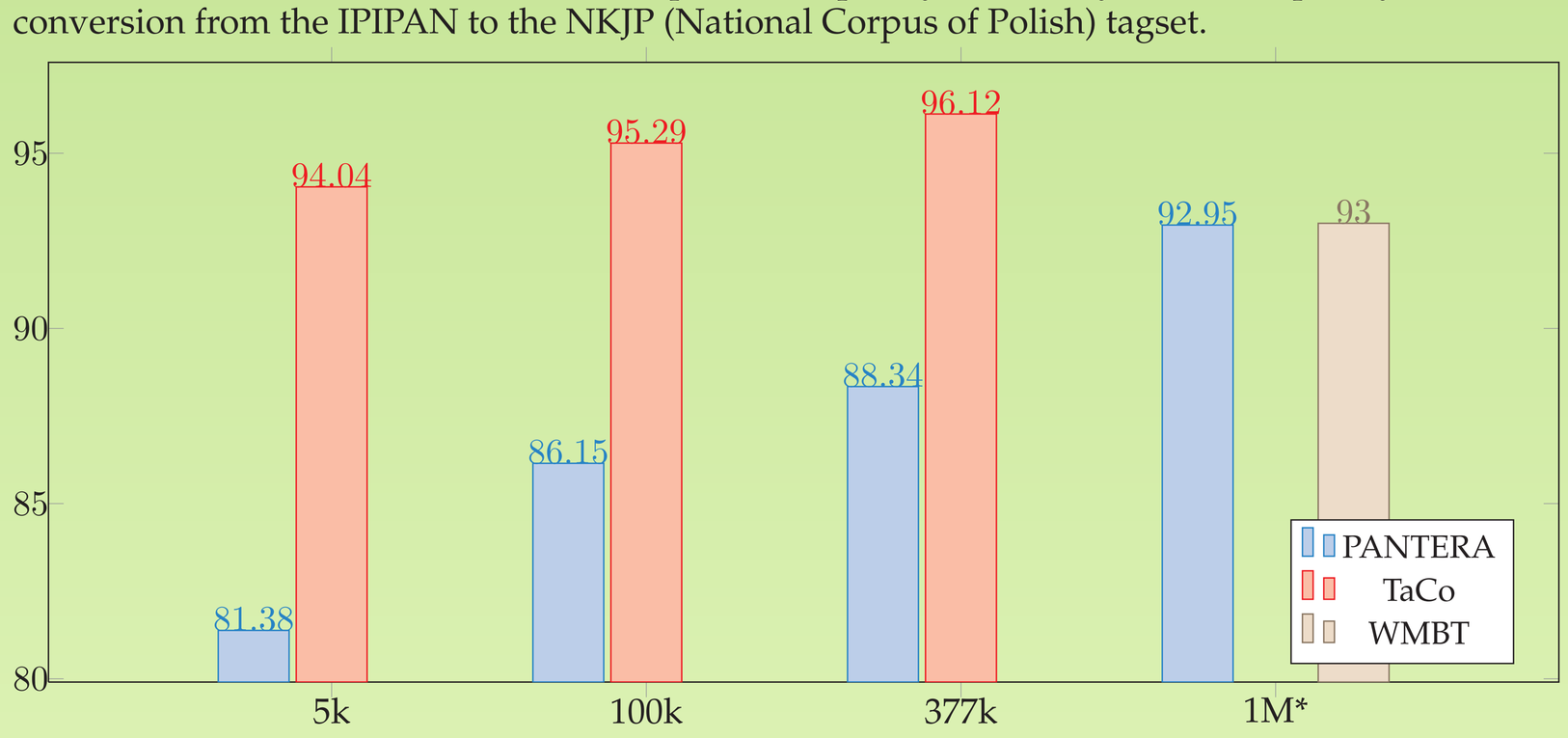
- statistical, supervised learning method
- for training, a corpus annotated by means of both, source and target tagsets is required
- TaCo extracts from the existing annotation as much information related to the target tagset as possible
- underlying classifier - decision tree - produced using well-known C5.0 algorithm
- morphological analysis using Morfeusz for further improvement of results
- not tied to specific tagsets or language (tagsets are parameter, morphological analyzer can be disabled or replaced with small effort)
- only few thousand words in training corpus is needed to achieve performance better than this of regular taggers trained on million of tokens (assuming the source corpus for conversion is high grade)
- high guessing capability for words unknown to morphological analyzer
- the algorithm adapts itself to specific corpora type during training thanks to error-driven learning possibilities
- fine-tuning of the training parameters may be performed for better results on specific types of corpora

Tool overview

- performs an automatic re-tagging of already annotated corpora using a different tagset
- high quality of results, assuming that the input data is high quality
- fast and fully automatic conversion after the model is trained
- language and tagset independent (requirement: tagset must be convertible to a positional format)
- various configurable parameters of training allow to fine-tune converter for use with specific language, tagset and corpora type
- implementation under GPL license (**homepage: http://zil.ipipan.waw.pl/TaCo**)

Evaluation results

TaCo was evaluated on the Enhanced Corpus of Frequency Dictionary of Contemporary Polish, on conversion from the IPIPAN to the NKJP (National Corpus of Polish) tagset.



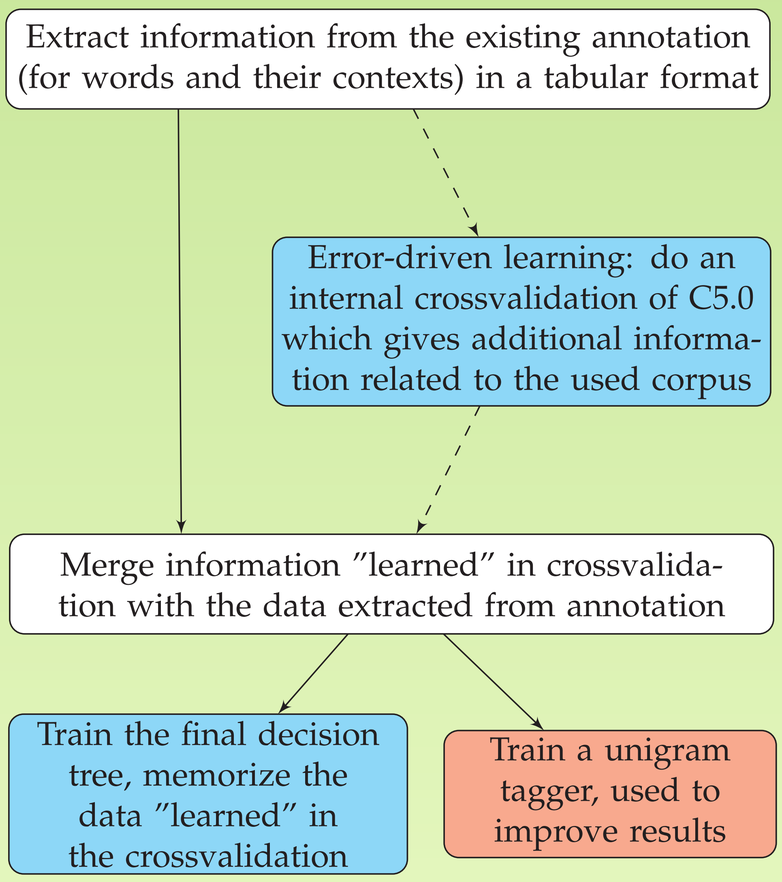
Correctness of TaCo in comparison with state-of-the-art polish taggers on corpora of different sizes.
* rough comparison: same tagset but different corpus used, values taken from literature.

corpus size	correctness			resources used
	all tokens	ambiguous	unknown (guessing)	
377k	96.13%	94.76%	83.98%	20h, 10.5GB of RAM
100k	95.29%	93.15%	86.32%	4h, 2.0GB of RAM
5k	94.04%	91.51%	68.75%	4min, 180MB of RAM

Detailed results of evaluation (resources are for the whole crossvalidation).

Training algorithm

The training algorithm in a nutshell:



Tagging algorithm

The conversion (re-tagging) algorithm: