Identification of Multiword Expressions in Parallel Latvian and Lithuanian Corpus (WG3)

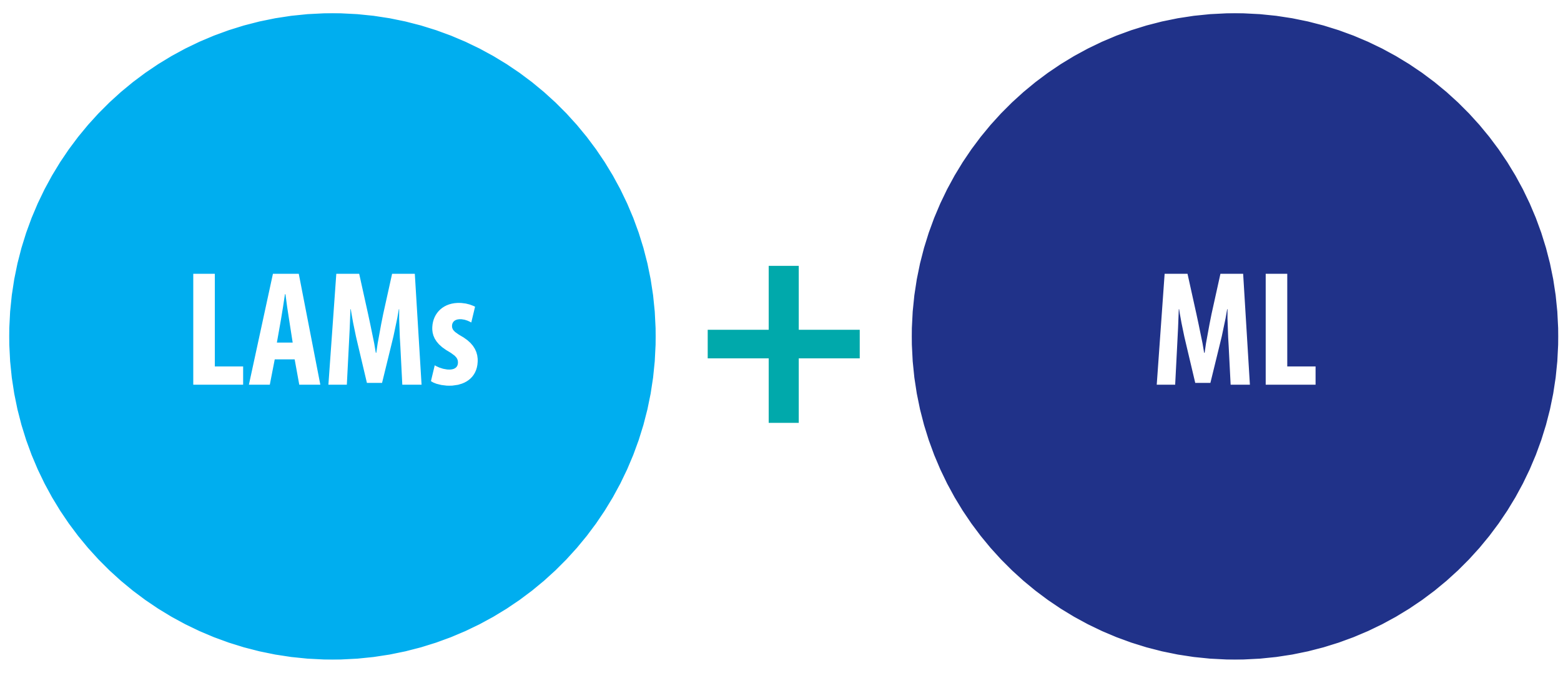
PURPOSE

Automatic identification of bi-gram multiword expressions (MWEs) in parallel Latvian and Lithuanian corpora. Our approach uses raw corpora and combination of lexical association measures (LAMs) and supervised machine learning (ML).

Justina Mandravickaitė
Baltic Institute of Advanced Technology
Vilnius University
justina@bpti.lt

Tomas Krilavičius
Baltic Institute of Advanced Technology
Vytautas Magnus University
t.krilavicius@bpti.lt

Inguna Skadiņa
University of Latvia
inguna.skadina@lumii.lv



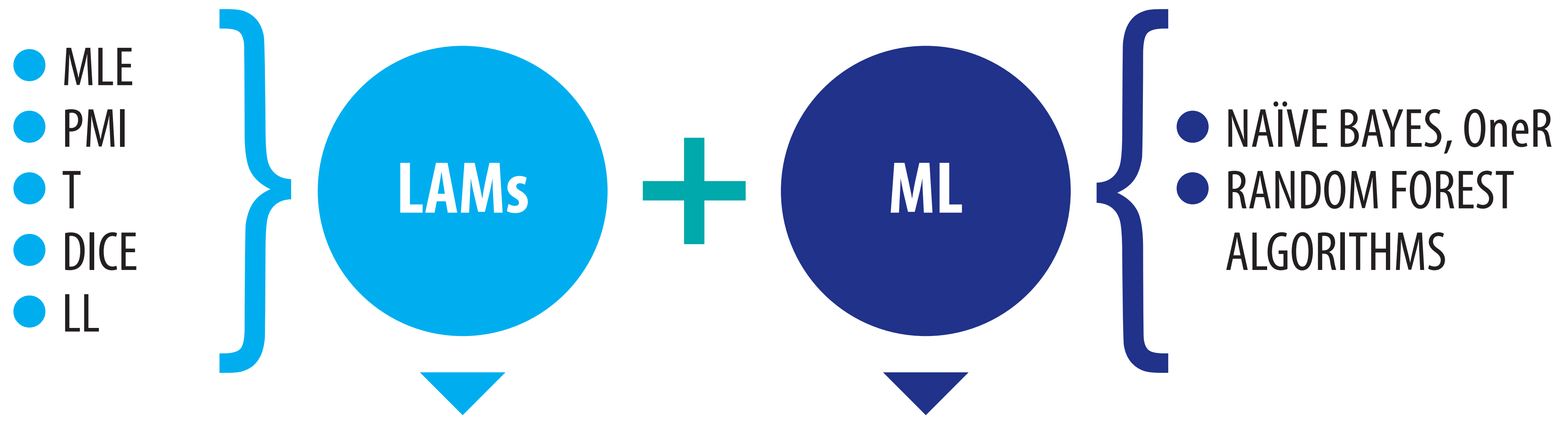
CORPORA AND LEXICAL RESOURCES FOR EVALUATION

1/3 Latvian and Lithuanian parts of JRC-Acquis Multilingual Parallel Corpus (~ *9 mln. words per language*)

EuroVoc, a Multilingual Thesaurus of the European Union, used as reference source for evaluation

- Bi-gram terms
- Separate MWE lists for Latvian (*3608 bi-grams*) and Lithuanian (*Lithuanian - 3783*)

METHOD



- MLE
- PMI
- T
- DICE
- LL

- NAÏVE BAYES, OneR
- RANDOM FOREST ALGORITHMS

Candidate list
Lexical association measures

- MLE (*Maximum Likelihood Estimation*)
- PMI (*Pointwise Mutual Information*)
- T (*Student's t score*)
- DICE (*Dice's coefficient*)
- LL (*Log-likelihood score*)

Reference list
Evaluation against the reference list

MWETOOLKIT

Supervised machine learning algorithms
Filters

- SMOTE (*Synthetic Minority Oversampling TEchnique*)
- Resample

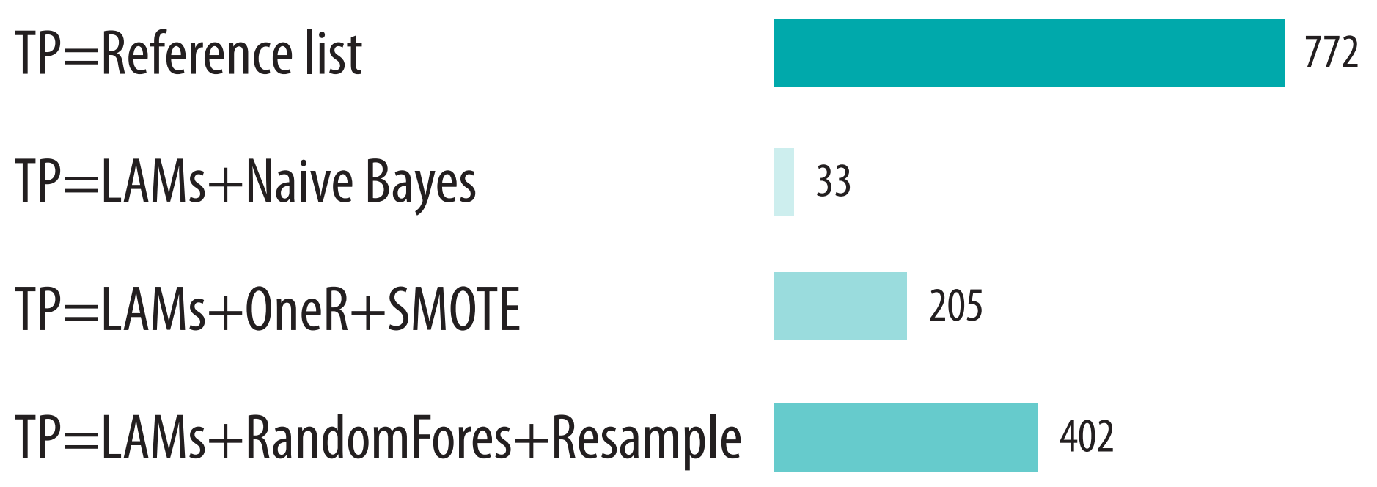
Evaluation
- Precision, Recall, F-measure
- 10-fold cross validation

WEKA

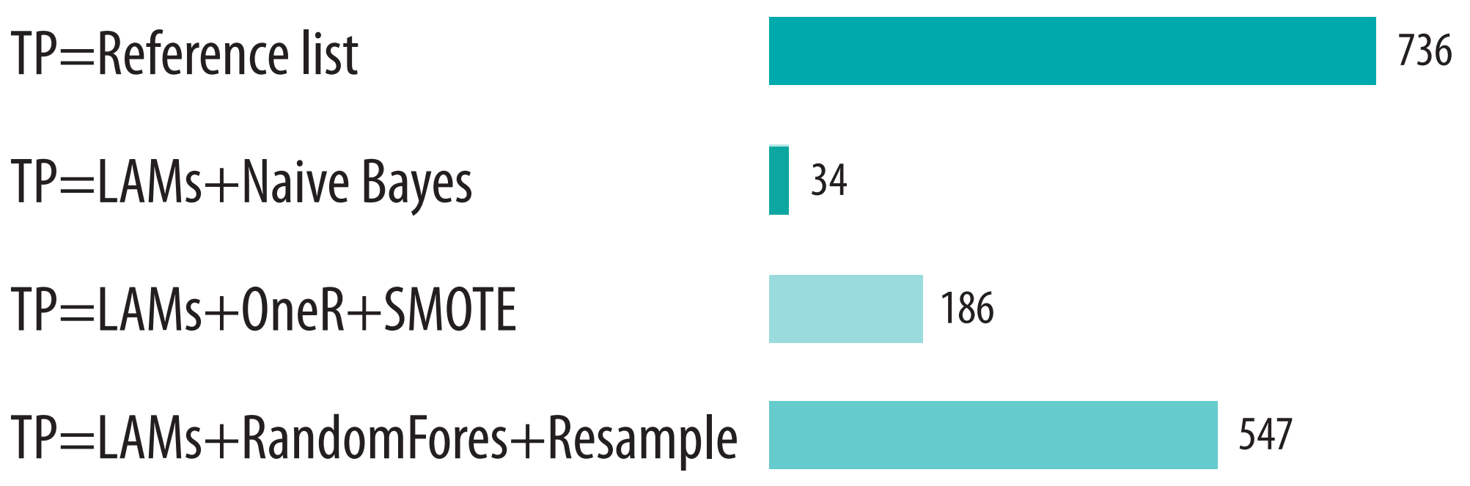
RESULTS

	SCENARIO	PRECISION	RECALL	F-MEAS.
LV	LAMs	0.1%	21.4%	0.3%
	LAMs+NayveBayes	0.6%	4.3%	1.1%
	LAMs+OneR+SMOTE	**100%**	13.3%	23.4%
	LAMs+Random Forest+Resample	92.4%	**52.2%**	**66.7%**
LT	LAMs	0.2%	19.4%	0.2%
	LAMs+NayveBayes	0.6%	4.6%	1.1%
	LAMs+OneR+SMOTE	**100%**	12.6%	22.4%
	LAMs+Random Forest+Resample	95.1%	**77.8%**	**85.6%**

LV = TP IN VARIOUS SCENARIOS



LT = TP IN VARIOUS SCENARIOS



CONCLUSION AND FUTURE PLANS

Extraction of bigram MWEs for Latvian and Lithuanian languages by combining LAMs and supervised ML improved results.

Future plans:

1. Automatic extractions of LT and LV MWEs
2. Experiments with wider set of features and tools, e.g. GIZA++ probability scores