Frederik Vaassen
Jurgita Kapočiūtė-Dzikienė
Walter Daelemans
Guy De Pauw

A simple method for topic classification for morphologically complex languages

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER
CLiPS
Universiteit Antwerpen
Lietuvos mokslo taryba

Overview

- Task Definition
- Experiments
  - Lithuanian
  - Russian
  - Swahili (WIP)
- Conclusion

Task Definition

- **Text classification (categorization)** is a problem of assigning an electronic text document to one or more categories, based on its contents.

- **Categories = topics**

- **Supervised classification** – the type of classification, having external mechanisms (human feedback) providing information on the correct decision.

➔ **Supervised topic classification**

Topic Classification for Morphologically Complex Languages

- Existing topic classification methods are effective for English (and other "popular" languages) having:
  - A wide range of annotated corpora
  - Grammatical tools (stemmers, lemmatizers…)
  - (Ontologies, databases…)

➔ Do these methods work for languages that are substantially different, or for resource-scarce languages?

Topic Classification for Morphologically Complex Languages

Inspiration: **Lithuanian**

- One of the most archaic and conservative living Indo-European languages
- Highly inflective (e.g. adjectives have 285 different word forms, expressed by different endings)
- Has rich word derivation system (e.g. 14 prefixes for phrasal verbs; 78 suffixes for diminutives and hypocoristic words, etc.)
- Has rich vocabulary (0.6 million headwords)

- Very little research on topic classification for Lithuanian (Kapočiūtė-Dzikienė et al. 2012)

Topic Classification for Morphologically Complex Languages

- The proposed topic classification method has to be able to cope with the complexity of Lithuanian
- The external information sources should be kept to a minimum
- Validate method on other, similarly morphologically complex languages:
  - Russian
  - Swahili

Experiments: Datasets

Language	Dataset	# of classes	# of documents	# of tokens/document
Lithuanian	Lietuvos rytas	11	8,936	37
	Supermamos	14	11,353	62
	Rinkimu programos ‘04	8	2,388	13
Russian	Forumishka	5	28,556	87
	Privet	11	17,909	47
Swahili	Wikipedia	15	1,671	346

- Varying levels of formality (political programs, forums)
- Varying distance between topics
- Varying number of topics, data set sizes and document lengths

Experiments: Lithuanian

Feature types:

- Unigrams based on word tokens (bag-of-words)
- Unigrams based on lemmatized words
- Character n-grams (sliding window)

Classifier:

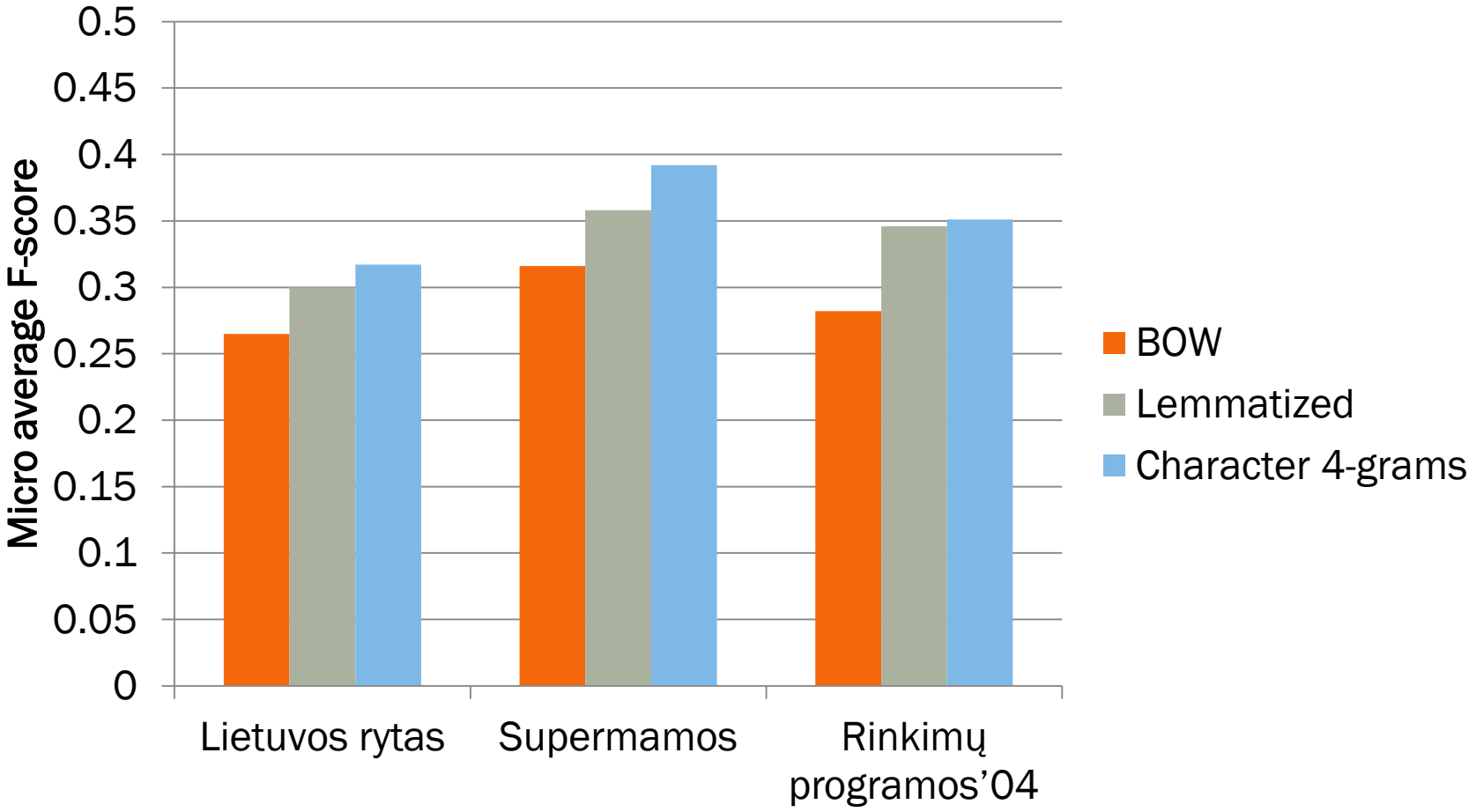
- SVM (libSVM)
- 10-fold CV

Experiments: Lithuanian

- **Hypothesis 1**: Bag-of-words approach should not be the best technique for Lithuanian, but lemmatization should improve classification results.

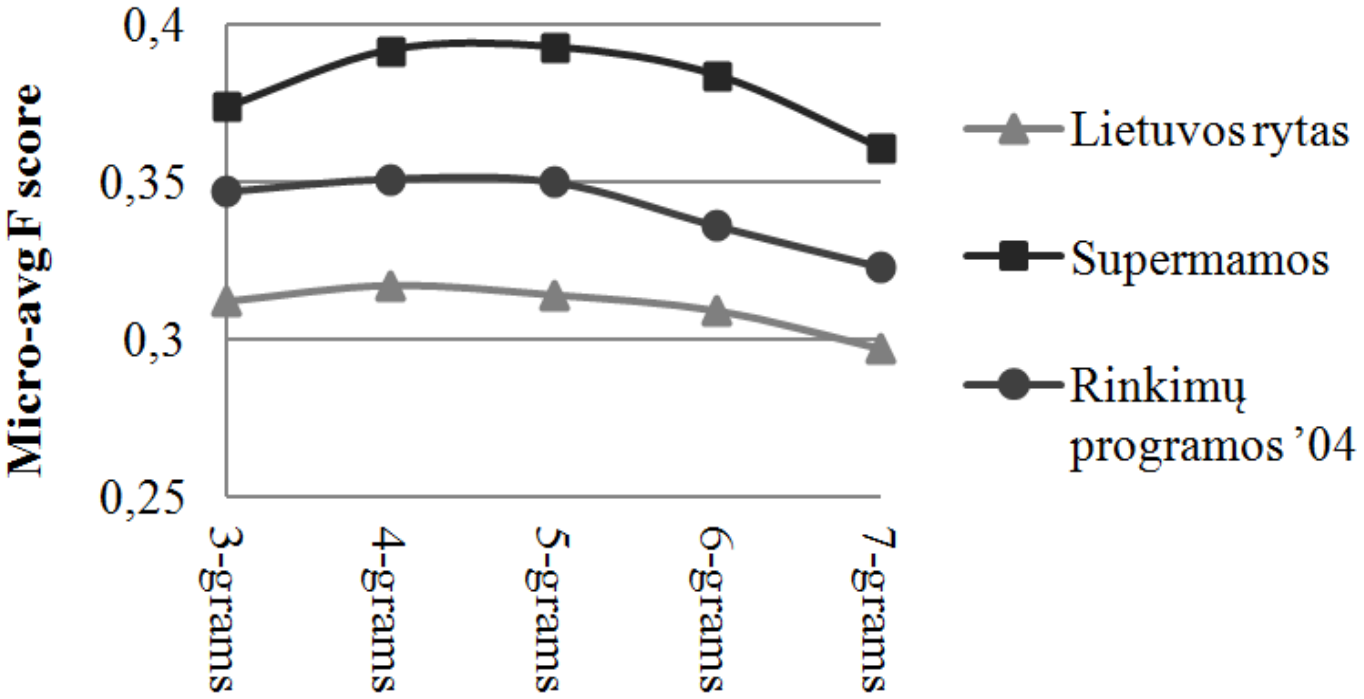
- **Hypothesis 2**: Character n-grams implicitly capture the relevant patterns within morphologically complex words (without having to resort to external grammatical tools).

Experiments: Lithuanian



Experiments: Lithuanian

Size of character n-grams:



Experiments: Lithuanian

- Examples of strong features:
  - "*vald*":
    - "valdymas"(management)
    - "valdžia" (authority)
    - "pavaldumas" (subordination)
    - "valdyti" (to govern);
    - "įvaldyti" (to master)
    - "suvaldyti" (to manage)
    - "savivaldybė" (municipality) ("savas", own + "valdyti", to govern)
    - "žemėvalda" (land-ownership) ("žemė", land + "valdyti", to govern)
    - …

Experiments: Lithuanian

- Examples of strong features:
  - “kari”:
    - “karininkas” (officer)
    - “kariuomenė” (army)
    - “karinis” (military)
    - “kariai” (soldiers)
    - …

Experiments: Russian and Swahili

- Can we reproduce these results on different languages with a similarly complex morphology?

Language	Dataset	# of classes	# of documents	# of tokens/document
Russian	Forumishka	5	28,556	87
	Privet	11	17,909	47
Swahili	Wikipedia	15	1,671	346

Experiments: Russian

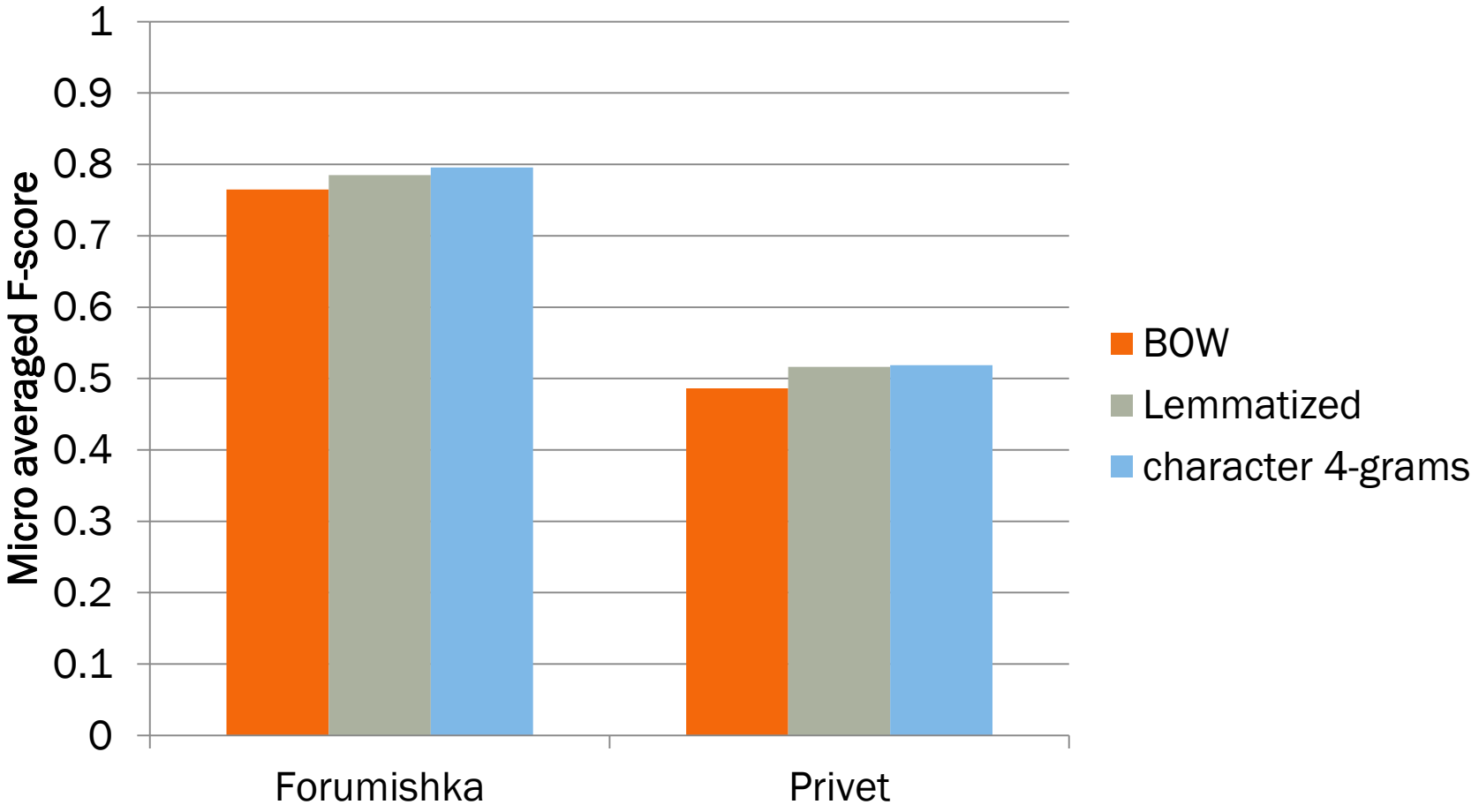
Feature types:

- Unigrams based on word tokens (bag-of-words)
- Unigrams based on lemmatized words
- Character n-grams (sliding window)

Classifier:

- SVM (libSVM)
- 10-fold CV

Experiments: Russian



Experiments: Russian

- Examples of strong features:
  - “хоро”
    - “хорош” (good), masc.
    - “хороша” (good), fem.
    - “нехороша” (not good), fem.
    - “хорошая” (good), fem. pron.
    - “хорошенькая” (pretty), coll. fem.
    - …

Experiments: Swahili

Feature types:

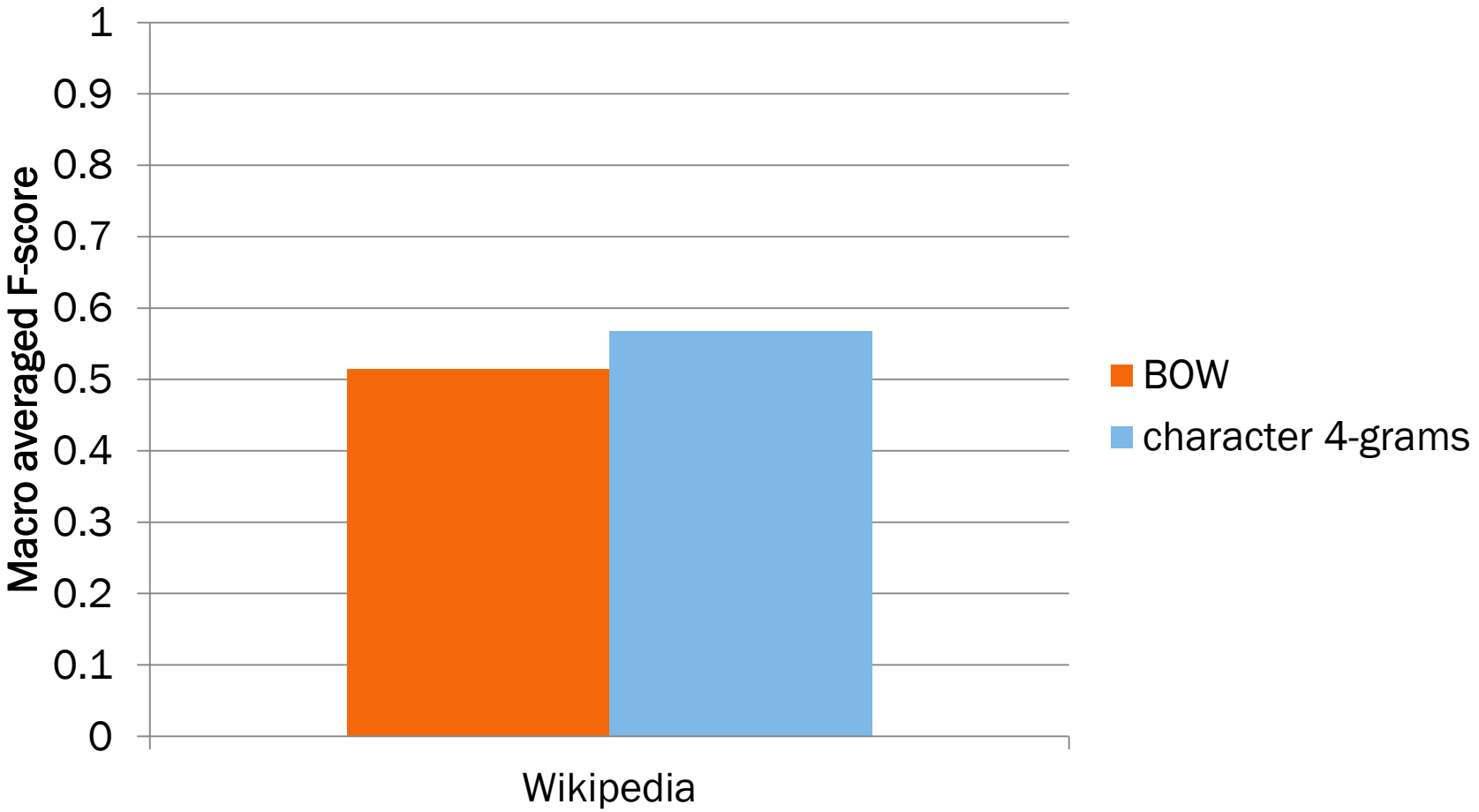
- Unigrams based on word tokens (bag-of-words)
- Character n-grams (sliding window)

Classifier:

- SVM (libSVM)
- 10-fold CV

Language	Dataset	# of classes	# of documents	# of tokens/document
Swahili	Wikipedia	15	1,671	346

Experiments: Swahili (WIP)



Conclusion

- We formulated and confirmed two hypotheses:
  - The common bag-of-words approach is not the best for morphologically complex languages; stemming or lemmatization may significantly improve topic classification performance.
  - Character n-grams implicitly capture relevant patterns and can even outperform classifiers trained on stemmed or lemmatized data (without resorting to external grammatical tools).

➔ Using character n-grams is a resource-independent and effective method for topic classification for morphologically complex languages

Contact

- frederik.vaassen@ua.ac.be
- j.kapociute-dzikiene@if.vdu.lt
- walter.daelemans@ua.ac.be
- guy.depauw@ua.ac.be

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER CLiPS Universiteit Antwerpen Lietuvos mokslo taryba