DKE

A Comparative Study on Language Identification Methods

Lena Grothe, Ernesto William De Luca, Andreas Nürnberger

DKE Group, Faculty of Computer Science, University of Magdeburg, Germany, www.findke.ovgu.de

Overview

- Motivation
- Language Identification
- Evaluation
- Conclusions & Future Work

Motivation

- Huge amount of available documents in WWW,
- BUT, language of documents is unknown

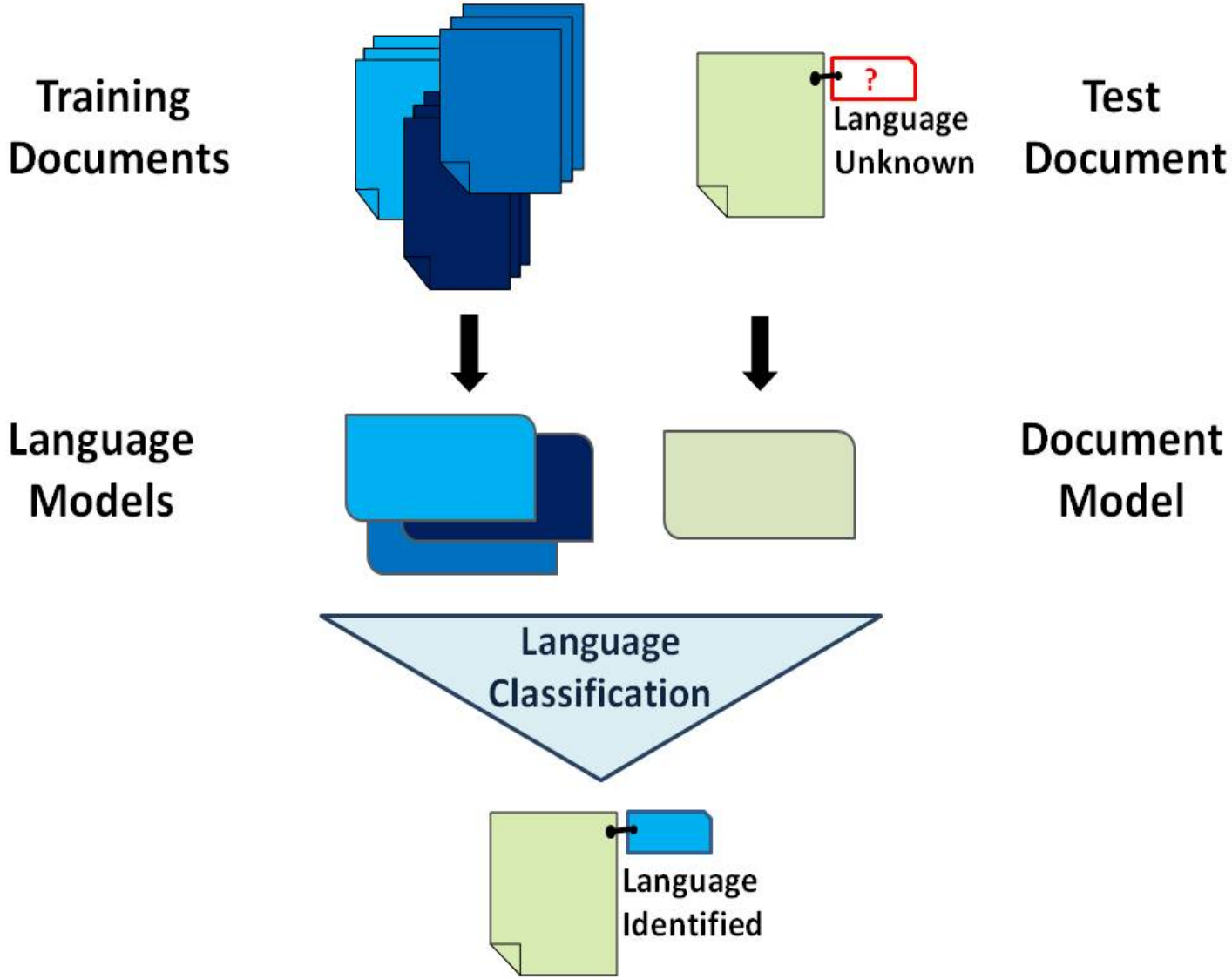
- Language can be used for:
  - Stemming
  - Machine Translation
  - Document Filtering
  - …

- Our Approach:
  - Language as **additional document annotation**
  - Language-specific **document filtering**

Overview

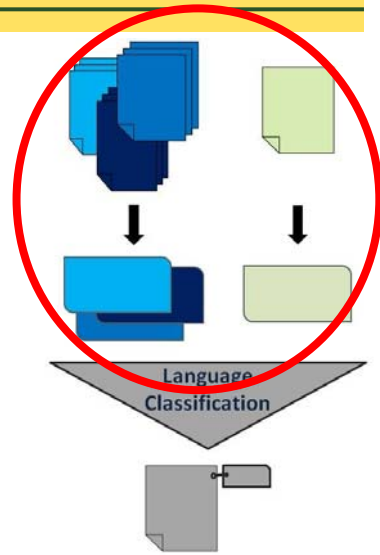
- Motivation
- Language Identification
  - Language Identification Process
  - Model Induction
  - Ad-Hoc Ranking
- Evaluation
- Conclusion & Future Work

Language Identification Process



Model Induction

- Approaches:
  - Frequent Words
  - Short Words
  - N-Grams

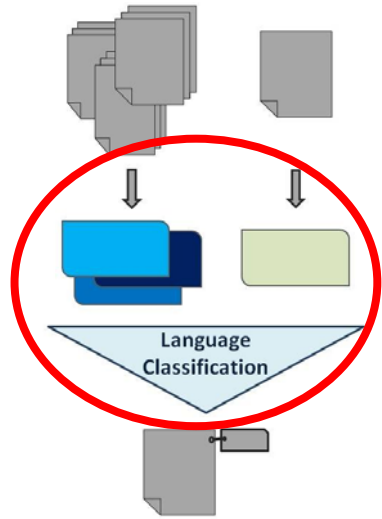


- Models:
  - Contain ranked entities
  - First rank entity: most frequent
  - Last rank entity: least frequent

  →Input for classification

Ad-hoc Ranking

- Main idea
  - model distance computation
- Model distance
  - sum of entity distances
- Entity distance
  - rank distance = Out-Of-Place Measure (OOPM)

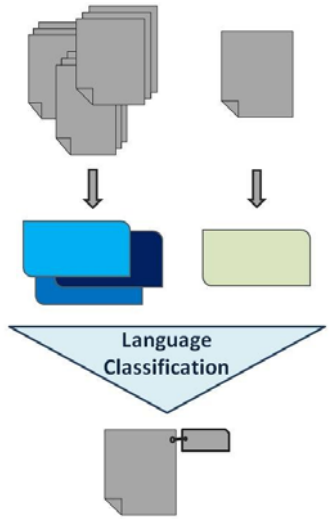


Out-Of-Place Measure (1/2)

- 2 Cases:
  - Entity exists in both models

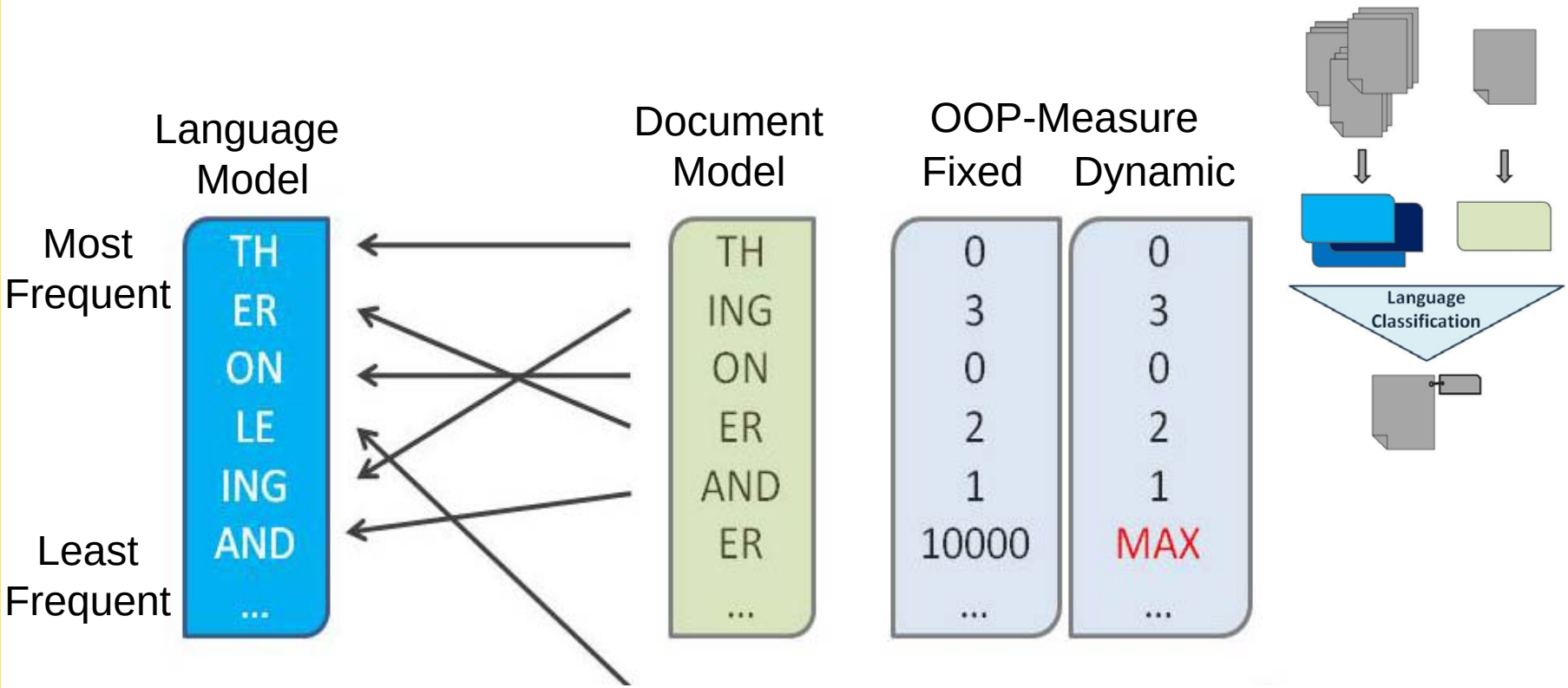
    → $oopm(e_i) = Rank_{DM}(e_i) - Rank_{LM}(e_i)$
  - Entity exists only in document model

    → $oopm(e_i) = MaxOOPM$



- MaxOOPM determination
  - Fixed value (model independent)
  - Dynamic value (model dependent)

Out-Of-Place Measure (2/2)



Overview

- Motivation
- Language Identification
- Evaluation
  - Evaluation Settings
  - Evaluation with Wikipedia
  - Evaluation with LCC
- Conclusion & Future Work

Evaluation Settings (1/2)

Wikipedia:

- Languages:
  Catalan, Danish, Dutch, English, French, German, Italian, Norwegian, Swedish

- Model Parameters:
  - Frequent Words (10%, 25%, 50%)
  - Short Words (3, 4, 5)
  - N-Grams (3, 4, 5)

Leipzig Corpora Collection:

- Languages:
  Catalan, Danish, Dutch, English, French, German, Italian, Norwegian, Swedish
  - Additional: Estionian, Finnish, Sorbian, Turkish

- Model Parameters:
  - Frequent Words (25%)
  - Short Words (4)
  - N-Grams (3)

Evaluation Settings (2/2)

Wikipedia:

- Training Data:
  - Closed subset of LCC
- Test Data:
  - Closed subset of Wikipedia (15 docs per language)
- **Fixed** MaxValue: 10000

Leipzig Corpora Collection:

- Training Data:
  - Closed subset of LCC
- Test Data:
  - Closed subset of LCC (250 docs/lang)
- **Dynamic** MaxValue

Fixed MaxValue for OOPM

Model (Parameter)	Correct Identified
Frequent Words (10%)	98,5%
Frequent Words (25%)	99,2%
Frequent Words (50%)	98,5%
Short Words (3)	93,3%
Short Words (4)	94,1%
Short Words (5)	91,8%
N-Grams (3)	79,2%
N-Grams (4)	30,3%
N-Grams (5)	1,5%
Frequent Words (25%) + Short Words (4)	94,1%
Frequent Words (25%) + N-Grams (3)	85,9%

Language Model Ex. (Short Words, 4)

Language	Uncl.	CAT	DE	DK	EN	FR	IT	NL	NO	SE	**Total**
Unclassif.	-										
Catalan		14			1						15
German			15								15
Danish				15							15
English					13	1			1		15
French						15					15
Italian							15				15
Dutch	1							14			15
Norwegian	1	1							12		15
Swedish	1									14	15
Total											135

Dynamic MaxValue for OOPM

Model (Parameter)	Correct Identified
Frequent Words (25%)	100%
Short Words (4)	100%
N-Grams (3)	100%

- MaxValue for OOPM changed dynamically per run
- Strong influence on classification performance
- Choice of model → less important

Conclusion & Future Work

- Language Identification
- Models: Frequent and Short Words, N-Grams
- Classification Method: Ad Hoc Ranking
- 2 Evaluation Sets: Static vs. Dynamic
- Importance of dynamic OOPM

- Future Work
  - Language families
  - Dialects
  - Minority languages

Thank You
for Your Attention!

Questions?