A learning-based Approach for Romanian Syllabification and Stress Assignment

DIANA BALC, ANAMARIA BELEIU, RODICA POTOLEA AND CAMELIA LEMNARU

Presentation

- Introduction & Objectives
- State of the Art
- Conceptual Model
- Testing
- Results
- Conclusion

Introduction & Objectives

- Orthographic Syllabification & Stress Assignment for Romanian language
- Motivation
- Romanian language hiatus/diphthong ambiguity

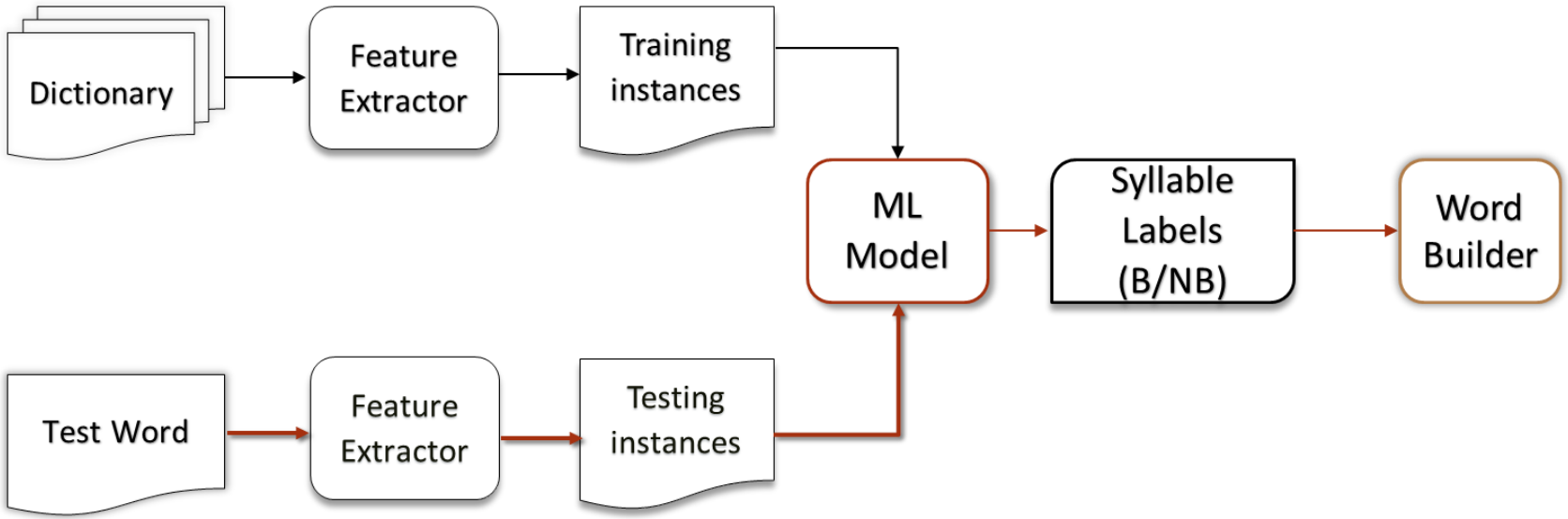
Sequence	Word	Diphthong	Word	Hiatus
ai	*hainǎ (noun)*	*h**ai**-nǎ*	*hainǎ (adj.)*	*h**a**-**i**-nǎ*
oa	*soare*	*s**oa**-re*	*croat*	*cr**o**-**a**t*
ou	*cadou*	*ca-d**ou***	*bour*	*b**o**-**u**r*

State of the Art

- Two approaches for solving this tasks:
  - Rule-based methods
  - Data driven techniques
- Note that the rule-based methods are language dependent and follow the exact set of grammar rules, but there are a large number of *exceptions*
- Data driven methods are based on the inference of *patterns* from a large and diverse number of examples in the training process
- Syllabification can be viewed as a structured learning problem
- For Romanian language, *MIRA** is known as a solution for Syllabification

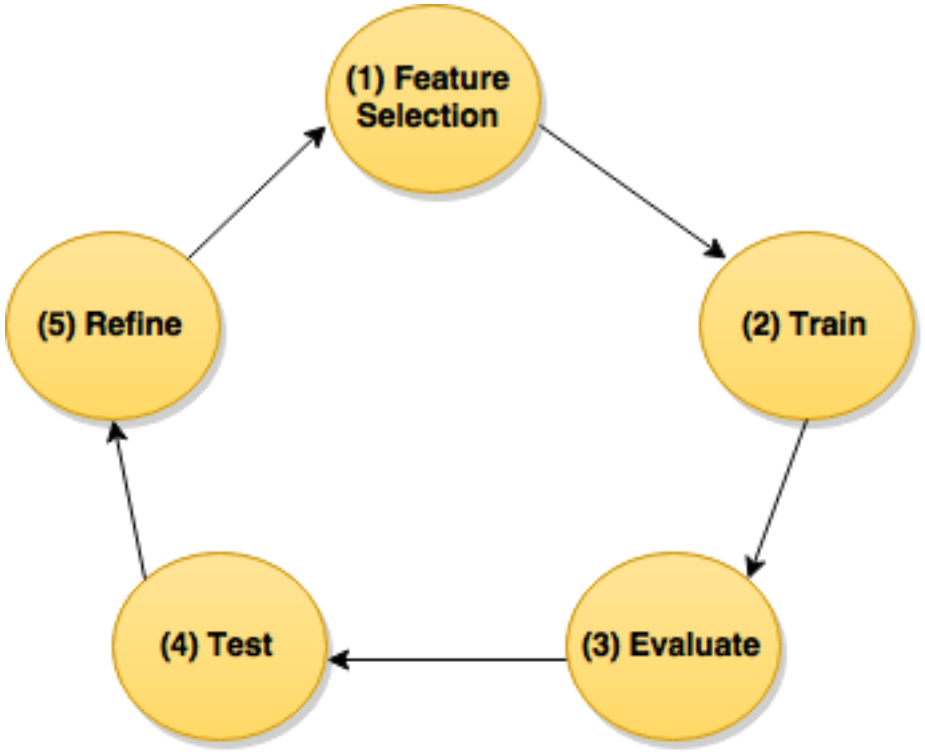
**T. Boroș, A Unified Lexical Processing Framework Based on the Margin Infused Relaxed Algorithm, 2013*

Conceptual Model



*Ana-Maria Barbu, *Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries*

Modelling Syllabification and Stress Assignment



Feature vector

Letter	Features	Class
î	*1 * * * * * î n ţ e l e*	**no**
n	*0 * * * * î n ţ e l e g*	**yes**
ţ	*0 * * * î n ţ e l e g i*	**no**
e	*1 * * î n ţ e l e g i * *	**yes**
l	*0 * î n ţ e l e g i * * *	**no**
e	*1 î n ţ e l e g i * * * *	**no**
g	*0 n ţ e l e g i * * * * *	**no**
i	*1 ţ e l e g i * * * * * *	**yes**

Table I. Example of features extracted for syllabification of the word "înţelegi"

Letter	Features	Class
î	** * * * * î n ţ e l e*	**yes**
e	** * î n ţ e l e g i **	**no**
e	*î n ţ e l e g i * * * *	**no**
i	*ţ e l e g i * * * * * *	**no**

Table II. Example of features extracted for stress assignment of the word "înţelegi"

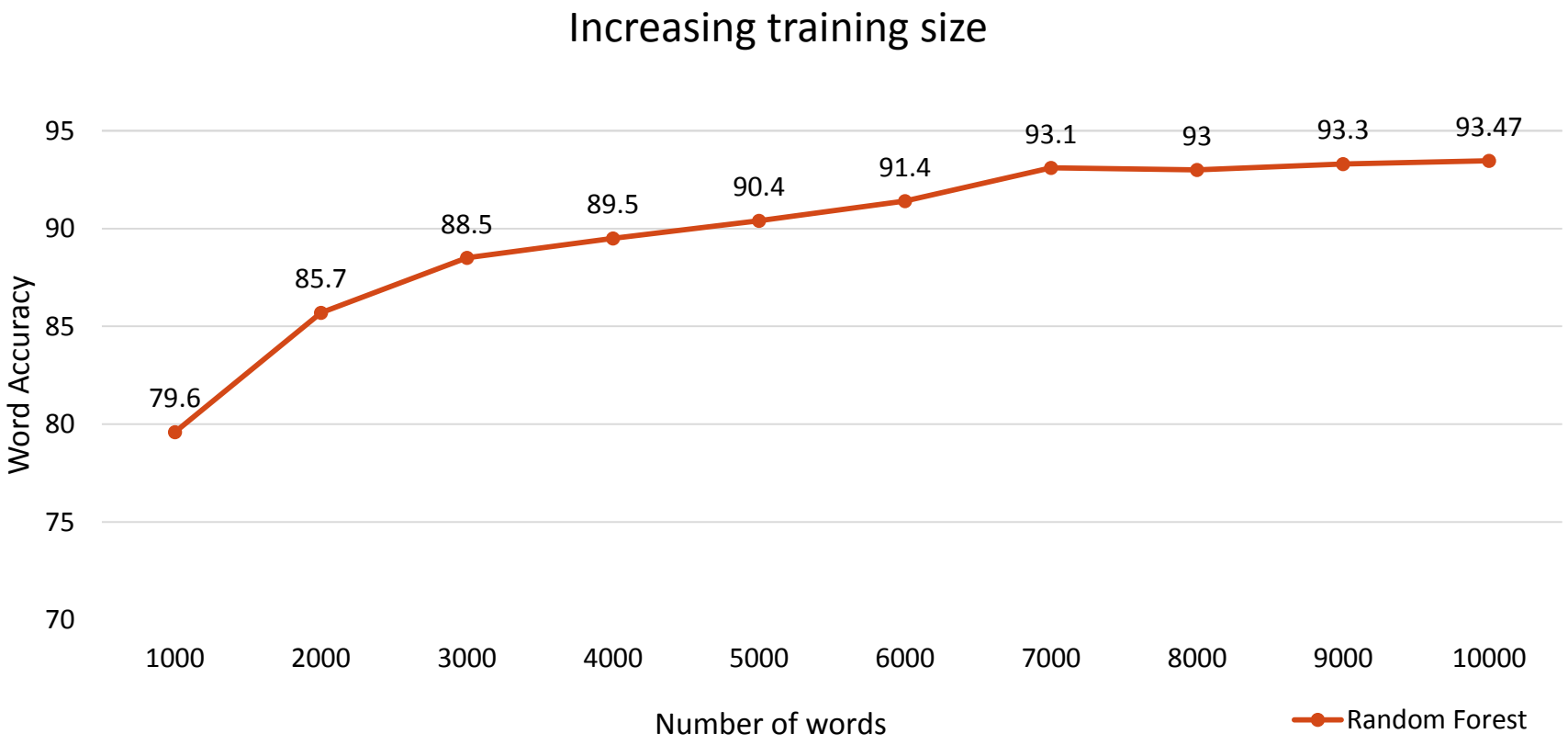
Stress assignment

Train set	Size	Accuracy (instance level)	Accuracy (word level)
Set 1	58.434	96,37 %	87,79 %
Set 2	58.626	96,37 %	86,74 %
Set 3	58.207	96,26 %	86,04 %
Set 4	58.293	96,58 %	86,74 %
Set 5	58.288	93,53 %	77,56 %

Table I. Random Forest models accuracy using the same test set for stress assignment – instance and word level

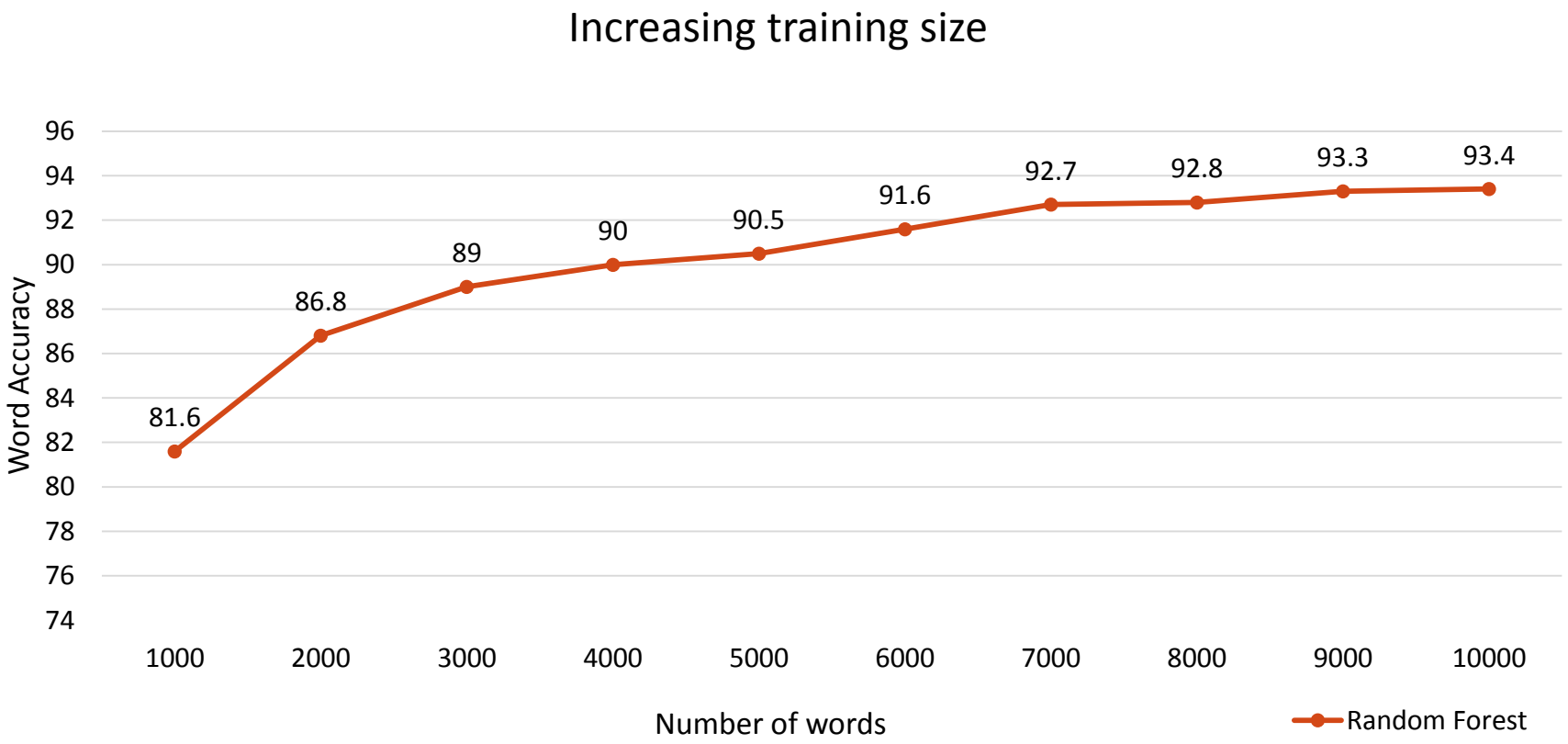
Test set: 860 Words (containing diacritics)

Training size - Syllabification



Test set: 35.000 Words (containing diacritics)

Training size – words without diacritics



Test set: 35.000 Words (no diacritics)

Conclusion

- Automatic syllabification and stress assignment can be modelled as a traditional classification problem by using supervised machine learning techniques
- The accuracy obtained at word level for syllabification is ~92% and for stress assignment ~85%
- Global component at phrase level - context dependency