

# 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
<i>ai</i>	<i>haină (noun)</i>	<i>h<u>ai</u>-nă</i>	<i>haină (adj.)</i>	<i>h<u>a</u>-<u>i</u>-nă</i>
<i>oa</i>	<i>soare</i>	<i>s<u>oa</u>-re</i>	<i>croat</i>	<i>cro<u>o</u>-<u>a</u>t</i>
<i>ou</i>	<i>cadou</i>	<i>ca-d<u>ou</u></i>	<i>bour</i>	<i>bo<u>o</u>-<u>u</u>r</i>

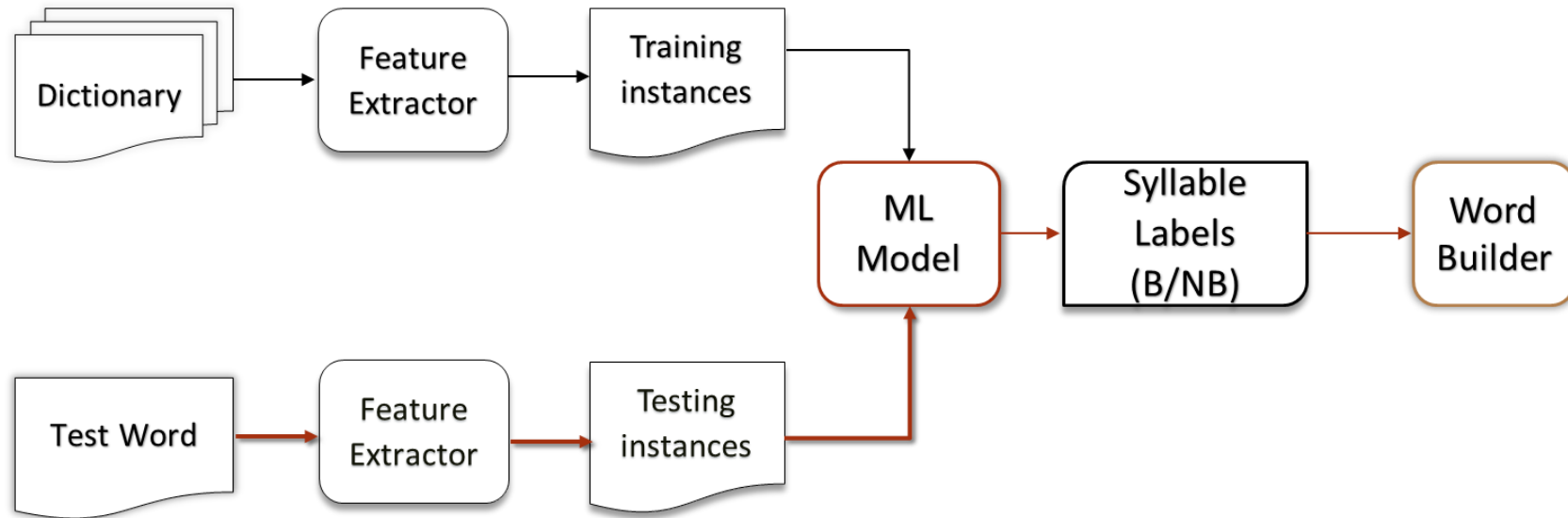
# State of the Art

---

- 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

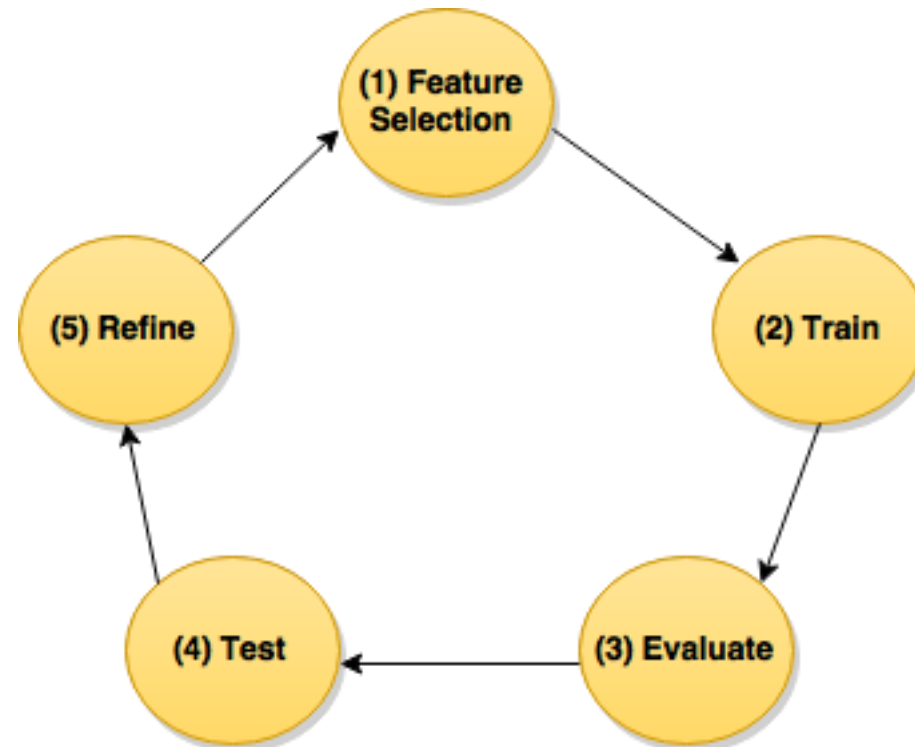
# Conceptual Model

---



# Modelling Syllabification and Stress Assignment

---



# Feature vector

---

Letter	Features	Class
<i>î</i>	<i>1 * * * * * î n ț e l e</i>	<b>no</b>
<i>n</i>	<i>0 * * * * * î n ț e l e g</i>	<b>yes</b>
<i>ț</i>	<i>0 * * * * * î n ț e l e g i</i>	<b>no</b>
<i>e</i>	<i>1 * * * * * î n ț e l e g i *</i>	<b>yes</b>
<i>l</i>	<i>0 * * * * * î n ț e l e g i * *</i>	<b>no</b>
<i>e</i>	<i>1 * * * * * î n ț e l e g i * * *</i>	<b>no</b>
<i>g</i>	<i>0 * * * * * n ț e l e g i * * * *</i>	<b>no</b>
<i>i</i>	<i>1 * * * * * ț e l e g i * * * * *</i>	<b>yes</b>

Table I. Example of features extracted for syllabification of the word “încelegi”

Letter	Features	Class
<i>î</i>	<i>* * * * * î n ț e l e</i>	<b>yes</b>
<i>e</i>	<i>* * * * * î n ț e l e g i *</i>	<b>no</b>
<i>e</i>	<i>* * * * * î n ț e l e g i * * *</i>	<b>no</b>
<i>i</i>	<i>* * * * * ț e l e g i * * * * *</i>	<b>no</b>

Table II. Example of features extracted for stress assignment of the word “încelegi”

# Stress assignment

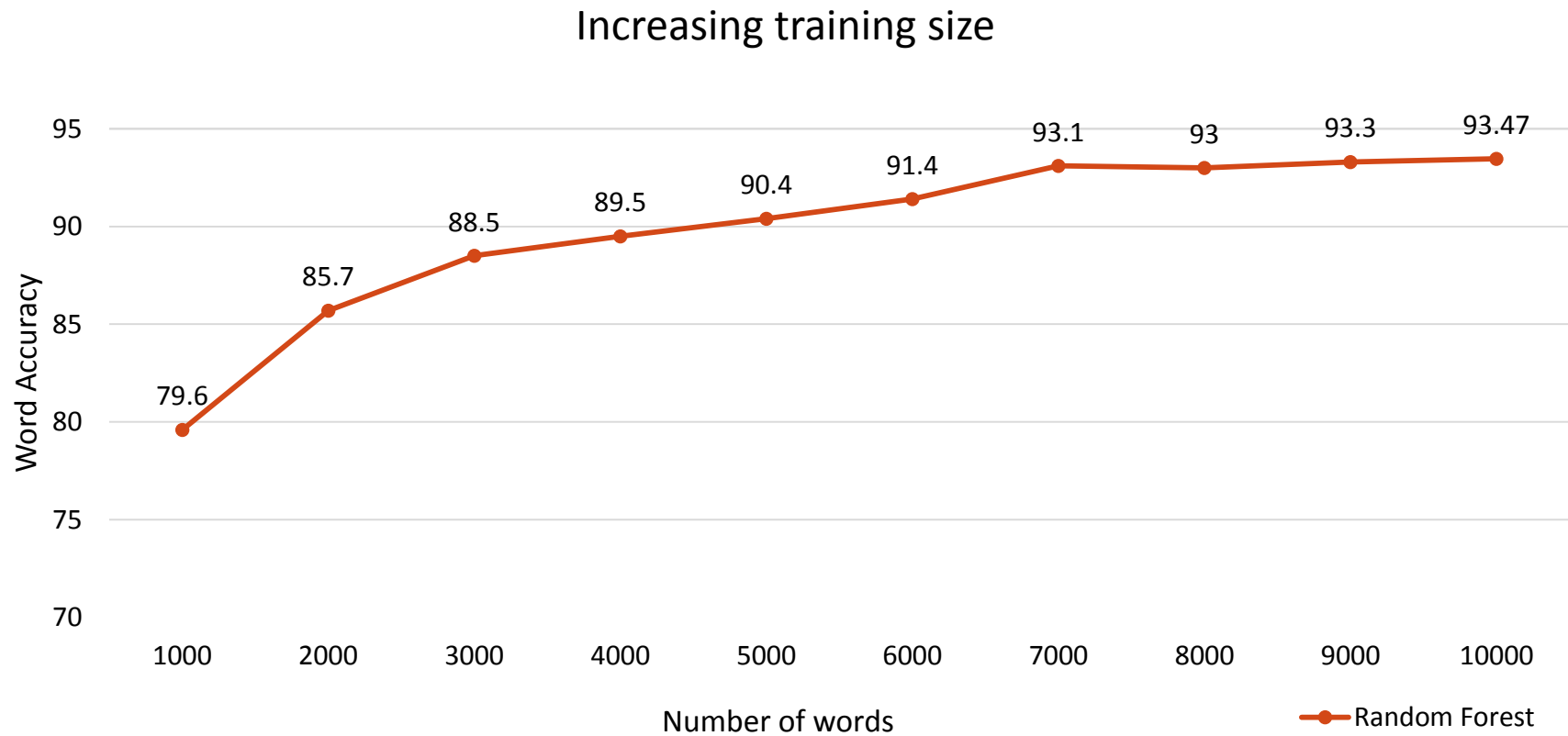
---

<b>Train set</b>	<b>Size</b>	<b>Accuracy (instance level)</b>	<b>Accuracy (word level)</b>
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

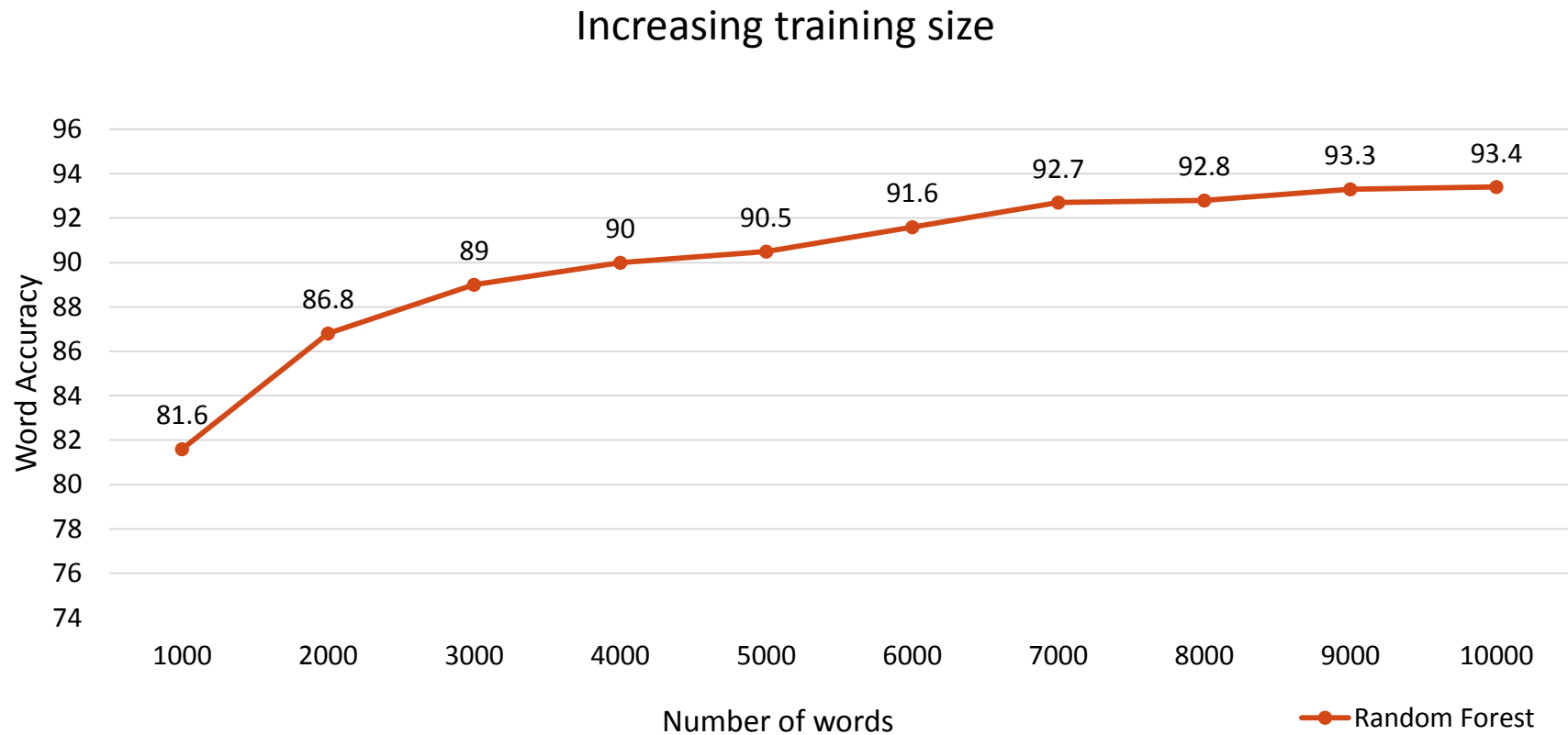


# Training size - Syllabification



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