Predicting Romanian Stress Assignment

Alina Maria Ciobanu[1,2], Anca Dinu[1,3], Liviu P. Dinu[1,2]
[1] Center for Computational Linguistics, University of Bucharest
[2] Faculty of Mathematics and Computer Science, University of Bucharest
[3] Faculty of Foreign Languages and Literatures, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, anca_d_dinu@yahoo.com, ldinu@fmi.unibuc.ro

Abstract

We train and evaluate two models for Romanian stress prediction: a baseline model which employs the consonant-vowel structure of the words and a cascaded model with averaged perceptron training consisting of two sequential models – one for predicting syllable boundaries and another one for predicting stress placement. We show in this paper that Romanian stress is predictable, though not deterministic, by using data-driven machine learning techniques.

1 Introduction

Romanian is a highly inflected language with a rich morphology. As dictionaries usually fail to cover the pronunciation aspects for all word forms in languages with such a rich and irregular morphology (Sef et al., 2002), we believe that a data-driven approach is very suitable for syllabication and stress prediction for Romanian words. Moreover, such a system proves extremely useful for inferring syllabication and stress placement for out-of-vocabulary words, for instance neologisms or words which recently entered the language.

Even if they are closely related, Romanian stress and syllabication were unevenly studied in the computational linguistic literature, i.e., the Romanian syllable received much more attention than the Romanian stress (Dinu and Dinu, 2005; Dinu, 2003; Dinu et al., 2013; Toma et al., 2009). One possible explanation for the fact that Romanian syllabication was more intensively studied than Romanian stress is the immediate application of syllabication to text editors which need reliable hyphenation. Another explanation could be that most linguists (most recently Dindelegan (2013)) insisted that Romanian stress is not predictable, thus discouraging attempts to investigate any systematic patterns.

Romanian is indeed a challenging case study, because of the obvious complexities of the data with respect to stress assignment. At first sight, no obvious patterns emerge for learning stress placement (Dindelegan, 2013), other than as part of individual lexical items. The first author who challenges this view is Chitoran (2002), who argues in favor of the predictability of the Romanian stress system. She states that stress placement strongly depends on the morphology of the language, more precisely on the distribution of the lexical items based on their part of speech (Chitoran, 1996). Thus, considering this type of information, lexical items can be clustered in a limited number of regular subpatterns and the unpredictability of stress placement is significantly reduced. A rule-based method for lexical stress prediction on Romanian was introduced by Oancea and Badulescu (2002).

Dou et al. (2009) address lexical stress prediction as a sequence tagging problem, which proves to be an accurate approach for this task. The effectiveness of using conditional random fields for orthographic syllabication is investigated by Trogkanis and Elkan (2010), who employ them for determining syllable boundaries and show that they outperform previous methods. Bartlett et al. (2008) use a discriminative tagger for automatic orthographic syllabication and present several approaches for assigning labels, including the language-independent *Numbered NB* tag scheme, which labels each letter with a value equal to the distance between the letter and the last syllable boundary. According to Damper et al. (1999), syllable structure and stress pattern are very useful in text-to-speech synthesis, as they provide valuable knowledge regarding the pronunciation modeling. Besides converting the letters to the corresponding phonemes, information about syllable boundaries and stress placement is also needed for the correct synthesizing of a word in grapheme-to-phoneme conversion (Demberg et al., 2007).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 64–68,
Gothenburg, Sweden, April 26-30 2014.

In this paper, we rely on the assumption that the stress system of Romanian is predictable. We propose a system for automatic prediction of stress placement and we investigate its performance by accounting for several fine-grained characteristics of Romanian words: part of speech, number of syllables and consecutive vowels. We investigate the consonant-vowel structure of the words (C/V structure) and we detect a high number of stress patterns. This calls for the need of machine learning techniques, in order to automatically learn such a wide range of variational patterns.

2 Approach

We address the task of stress prediction for Romanian words (out-of-context) as a sequence tagging problem. In this paper, we account only for the primary stress, but this approach allows further development in order to account for secondary stress as well. We propose a cascaded model consisting of two sequential models trained separately, the output of the first being used as input for the second. We use averaged perceptron for parameter estimation and three types of features which are described in detail further in this section: n-grams of characters, n-grams marking the C/V structure of the word and binary positional indicators of the current character with respect to the syllable structure of the word. We use one sequential model to predict syllable boundaries and another one to predict stress placement. Previous work on orthographic syllabication for Romanian (Dinu et al., 2013) shows that, although a rule-based algorithm models complex interactions between features, its practicality is limited. The authors report experiments on a Romanian dataset, where the rule-based algorithm is outperformed by an SVM classifier and a CRF system with character n-gram features.

We use a simple tagging structure for marking primary stress. The stressed vowel receives the positive tag 1, while all previous characters are tagged 0 and all subsequent ones 2. This structure helps enforce the uniqueness of the positive tag. The main features used are character n-grams up to $n = W$ in a window of radius W around the current position. For example, if $W = 2$, the feature template consists of `c[-2]`, `c[-1]`, `c[0]`, `c[1]`, `c[2]`, `c[-2:-1]`, `c[-1:0]`, `c[0:1]`, `c[1:2]`. If the current letter is the fourth of the word *dinosaur*, *o*, the feature values would be *i, n, o, s, a, in, no, os, sa*. We use two additional types of features:

- features regarding the C/V structure of the word: n-grams using, instead of characters, markers for consonants (C) and vowels (V);
- binary indicators of the following positional statements about the current character, related to the statistics reported in Table 1:
  - exactly before/after a split;
  - in the first/second/third/fourth syllable of the word, counting from left to right;
  - in the first/second/third/fourth syllable of the word, counting from right to left

The syllabication prediction is performed with another sequential model of length $n-1$, where each node corresponds to a position between two characters. Based on experimenting and previous work, we adopted the *Numbered NB* labeling. Each position is labeled with an integer denoting the distance from the previous boundary. For example, for the word *diamond*, the syllable (above) and stress annotations (below) are as follows:

```
d   i   a   m   o   n   d
  1   0   0   1   2   3
0   1   2   2   2   2   2
```

The features used for syllabication are based on the same principle, but because the positions are in-between characters, the window of radius W has length $2W$ instead of $2W+1$. For this model we used only character n-grams as features.

3 Data

We run our experiments for Romanian using the *RoSyllabiDict* (Barbu, 2008) dictionary, which is a dataset of annotated words comprising 525,528 inflected forms for approximately 65,000 lemmas. This is, to our best knowledge, the largest experiment conducted and reported for Romanian so far. For each entry, the syllabication and the stressed vowel (and, in case of ambiguities, also grammatical information or type of syllabication) are provided. For example, the word *copii* (*children*) has the following representation:

```
<form w="copii" obs="s."> co-píi</form>
```

We investigate stress placement with regard to the syllable structure and we provide in Table 1 the percentages of words having the stress placed on different positions, counting syllables from the beginning and from the end of the words as well.

For our experiments, we discard words which do not have the stressed vowel marked, compound

Syllable	%words
1st	5.59
2nd	18.91
3rd	39.23
4th	23.68
5th	8.52

(a) counting syllables from the beginning of the word

Syllable	%words
1st	28.16
2nd	43.93
3rd	24.14
4th	3.08
5th	0.24

(b) counting syllables from the end of the word

Table 1: Stress placement for *RoSyllabiDict*

words having more than one stressed vowel and ambiguous words (either regarding their part of speech or type of syllabication).

We investigate the C/V structure of the words in *RoSyllabiDict* using raw data, i.e., *a, ă, â, e, i, î, o, u* are always considered vowels and the rest of the letters in the Romanian alphabet are considered consonants. Thus, we identify a very large number of C/V structures, most of which are not deterministic with regard to stress assignment, having more then one choice for placing the stress[1].

4 Experiments and Results

In this section we present the main results drawn from our research on Romanian stress assignment.

4.1 Experiments

We train and evaluate a cascaded model consisting of two sequential models trained separately, the output of the first being used as input to the second. We split the dataset in two subsets: train set (on which we perform cross-validation to select optimal parameters for our model) and test set (with unseen words, on which we evaluate the performance of our system). We use the same train/test sets for the two sequential models, but they are trained independently. The output of the first model (used for predicting syllabication) is used for determining feature values for the second one (used for predicting stress placement) for the test set. The second model is trained using gold syllabication (provided in the dataset) and we report results on the test set in both versions: using gold syllabication to determine feature values and using predicted syllabication to determine feature values. The results with gold syllabication are reported only for providing an upper bound for learning and for comparison.

We use averaged perceptron training (Collins, 2002) from *CRFsuite* (Okazaki, 2007). For the stress prediction model we optimize hyperparameters using grid search to maximize the 3-fold cross-validation F_1 score of class 1, which marks the stressed vowels. We searched over $\{2,3,4\}$ for W and over $\{1,5,10,25,50\}$ for the maximum number of iterations. The values which optimize the system are 4 for W and 50 for the maximum number of iterations. We investigate, during grid search, whether employing C/V markers and binary positional indicators improve our system's performance. It turns out that in most cases they do. For the syllabication model, the optimal hyperparameters are 4 for the window radius and 50 for the maximum number of iterations. We evaluate the cross-validation F_1 score of class 0, which marks the position of a hyphen. The system obtains 0.995 instance accuracy for predicting syllable boundaries.

We use a "majority class" type of baseline which employs the C/V structures described in Section 3 and assigns, for a word in the test set, the stress pattern which is most common in the training set for the C/V structure of the word, or places the stress randomly on a vowel if the C/V structure is not found in the training set[2]. The performance of both models on *RoSyllabiDict* dataset is reported in Table 2. We report word-level accuracy, that is, we account for words for which the stress pattern was correctly assigned. As expected, the cascaded model performs significantly better than the baseline.

Model	Accuracy
Baseline	0.637
Cascaded model (gold)	0.975
Cascaded model (predicted)	0.973

Table 2: Accuracy for stress prediction

Further, we perform an in-depth analysis of the sequential model's performance by accounting for

[1] For example, for CCV-CVC structure (1,390 occurrences in our dataset) there are 2 associated stress patterns: CCV-CV̲C (1,017 occurrences) and CC̲V-CVC (373 occurrences). Words with 6 syllables cover the highest number of distinct C/V structures (5,749). There are 31 C/V structures (ranging from 4 to 7 syllables) reaching the maximum number of distinct associated stress patterns (6).

[2] For example, the word *copii* (meaning *children*) has the following C/V structure: CV-CVV. In our training set, there are 659 words with this structure and the three stress patterns which occur in the training set are as follows: CV-CV̲V (309 occurrences), C̲V-CVV (283 occurrences) and CV-CV̲V̲ (67 occurrences). Therefore, the most common stress pattern CV-CV̲V is correctly assigned, in this case, for the word *copii*.

several fine-grained characteristics of the words in *RoSyllabiDict*. We divide words in categories based on the following criteria:

- part of speech: verbs, nouns, adjectives
- number of syllables: 2-8, 9+
- number of consecutive vowels: with at least 2 consecutive vowels, without consecutive vowels

Category	Subcategory	♯ words	Accuracy G	Accuracy P
POS	Verbs	167,193	0.995	0.991
	Nouns	266,987	0.979	0.979
	Adjectives	97,169	0.992	0.992
Syllables	2 syllables	34,810	0.921	0.920
	3 syllables	111,330	0.944	0.941
	4 syllables	154,341	0.966	0.964
	5 syllables	120,288	0.981	0.969
	6 syllables	54,918	0.985	0.985
	7 syllables	17,852	0.981	0.989
	8 syllables	5,278	0.992	0.984
	9+ syllables	1,468	0.979	0.980
Vowels	With VV	134,895	0.972	0.972
	Without VV	365,412	0.976	0.974

Table 3: Accuracy for cascaded model with gold (G) and predicted (P) syllabication

We train and test the cascaded model independently for each subcategory in the same manner as we did for the entire dataset. We decided to use cross-validation for parameter selection instead of splitting the data in train/dev/test subsets in order to have consistency across all models, because some of these word categories do not comprise enough words for splitting in three subsets (words with more than 8 syllables, for example, have only 1,468 instances). The evaluation of the system's performance and the number of words in each category are presented in Table 3.

4.2 Results Analysis

The overall accuracy is 0.975 for the cascaded model with gold syllabication and 0.973 for the cascaded model with predicted syllabication. The former system outperforms the latter by only very little. With regard to the part of speech, the highest accuracy when gold syllabication is used was obtained for verbs (0.995), followed by adjectives (0.992) and by nouns (0.979). When dividing the dataset with respect to the words' part of speech, the cascaded model with predicted syllabication is outperformed only for verbs. With only a few exceptions, the accuracy steadily increases with the number of syllables. The peak is reached for words with 6 syllables when using the gold syllabication and for words with 7 syllables when using the predicted syllabication. Although, intuitively, the accuracy should be inversely proportional to the number of syllables, because the number of potential positions for stress placement increases, there are numerous stress patterns for words with 6, 7 or more syllables, which never occur in the dataset[3]. It is interesting to notice that stress prediction accuracy is almost equal for words containing two or more consecutive vowels and for words without consecutive vowels. As expected, when words are divided in categories based on their characteristics the system is able to predict stress placement with higher accuracy.

5 Conclusion and Future Work

In this paper we showed that Romanian stress is predictable, though not deterministic, by using data-driven machine learning techniques. Syllable structure is important and helps the task of stress prediction. The cascaded sequential model using gold syllabication outperforms systems with predicted syllabication by only very little.

In our future work we intend to experiment with other features as well, such as syllable n-grams instead of character n-grams, for the sequential model. We plan to conduct a thorough error analysis and to investigate the words for which the systems did not correctly predict the position of the stressed vowels. We intend to further investigate the C/V structures identified in this paper and to analyze the possibility to reduce the number of patterns by considering details of word structure (for example, instead of using raw data, to augment the model with annotations about which letters are actually vowels) and to adapt the learning model to finer-grained linguistic analysis.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. The contribution of the authors to this paper is equal. Research supported by a grant of ANRCS, CNCS UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

[3]For example, for the stress pattern CV-CV-CV-CV-CV-CVCV, which matches 777 words in our dataset, the stress is never placed on the first three syllables.

References

Ana-Maria Barbu. 2008. Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 1937–1941.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2008. Automatic Syllabification with Structured SVMs for Letter-to-Phoneme Conversion. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2008*, pages 568–576.

Ioana Chitoran. 1996. Prominence vs. rhythm: The predictability of stress in Romanian. In *Grammatical theory and Romance languages*, pages 47–58. Karen Zagona.

Ioana Chitoran. 2002. *The phonology of Romanian. A constraint-based approach*. Mouton de Gruyter.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP 2002*, pages 1–8.

Robert I. Damper, Yannick Marchand, M. J. Adamson, and K. Gustafson. 1999. Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech & Language*, 13(2):155–176.

Vera Demberg, Helmut Schmid, and Gregor Möhler. 2007. Phonological Constraints and Morphological Preprocessing for Grapheme-to-Phoneme Conversion. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL 2007*, pages 96–103.

Gabriela Pană Dindelegan. 2013. *The Grammar of Romanian*. Oxford University Press.

Liviu P. Dinu and Anca Dinu. 2005. A Parallel Approach to Syllabification. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2005*, pages 83–87.

Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Șulea. 2013. Romanian Syllabication Using Machine Learning. In *Proceedings of the 16th International Conference on Text, Speech and Dialogue, TSD 2013*, pages 450–456.

Liviu Petrisor Dinu. 2003. An Approach to Syllables via some Extensions of Marcus Contextual Grammars. *Grammars*, 6(1):1–12.

Qing Dou, Shane Bergsma, Sittichai Jiampojamarn, and Grzegorz Kondrak. 2009. A Ranking Approach to Stress Prediction for Letter-to-Phoneme Conversion. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, ACL 2009*, pages 118–126.

Eugeniu Oancea and Adriana Badulescu. 2002. Stressed Syllable Determination for Romanian Words within Speech Synthesis Applications. *International Journal of Speech Technology*, 5(3):237–246.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Tomaz Sef, Maja Skrjanc, and Matjaz Gams. 2002. Automatic Lexical Stress Assignment of Unknown Words for Highly Inflected Slovenian Language. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue, TSD 2002*, pages 165–172.

S.-A. Toma, E. Oancea, and D. Munteanu. 2009. Automatic rule-based syllabication for Romanian. In *Proceedings of the 5th Conference on Speech Technology and Human-Computer Dialogue, SPeD 2009*, pages 1–6.

Nikolaos Trogkanis and Charles Elkan. 2010. Conditional Random Fields for Word Hyphenation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 366–374.