SYLLABIFICATION WITH FREQUENT SEQUENCE PATTERNS

adrianulbona@gmail.com, {camelia.lemnaru and rodica.potolea}@cs.utcluj.ro

UNIVERSITATEA TEHNICĂ DIN CLUJ-NAPOCA

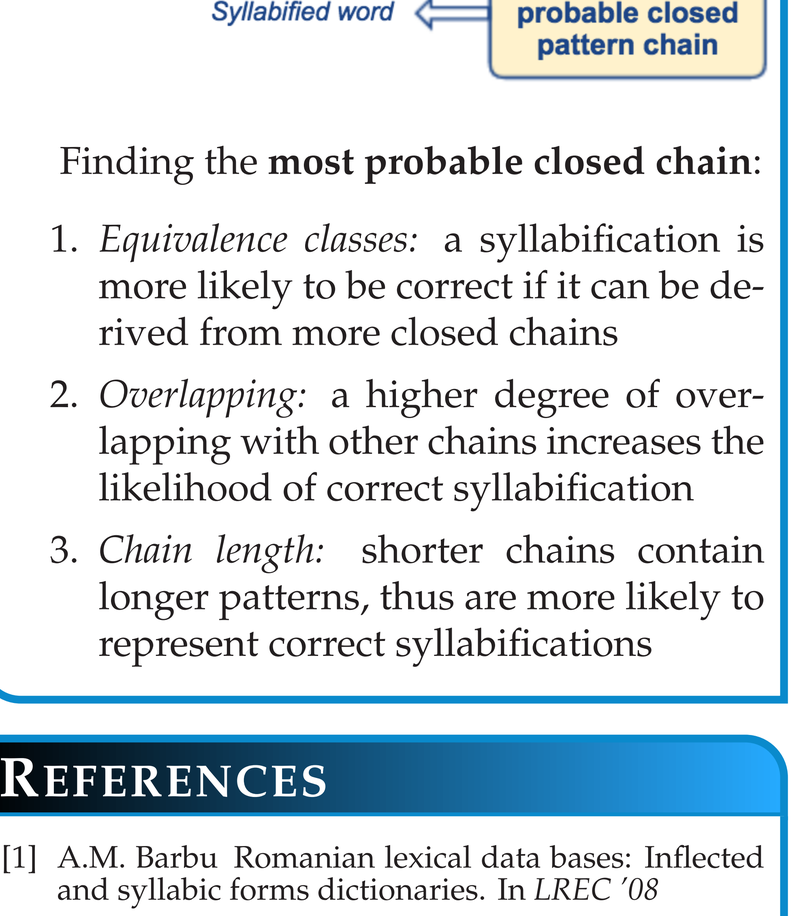
PROBLEM

Splitting words by various criteria has always presented high interest in a wide range of fields, such as linguistics and artificial intelligence. Text to speech systems and automatic end-of-line hyphenation in text editors could greatly benefit from an efficient, language independent method for syllabification.

CONTRIBUTIONS

We present a novel approach for word syllabification, based on frequent pattern mining, but also a more general framework for syllabification. Preliminary evaluations on Romanian and English words indicated a word level accuracy around 77% for Romanian words and around 70% for English words. However, we believe the method can be refined in order to improve performance.

METHOD



Finding the **most probable closed chain**:

1. *Equivalence classes:* a syllabification is more likely to be correct if it can be derived from more closed chains
2. *Overlapping:* a higher degree of overlapping with other chains increases the likelihood of correct syllabification
3. *Chain length:* shorter chains contain longer patterns, thus are more likely to represent correct syllabifications

REFERENCES

[1] A.M. Barbu Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *LREC '08*

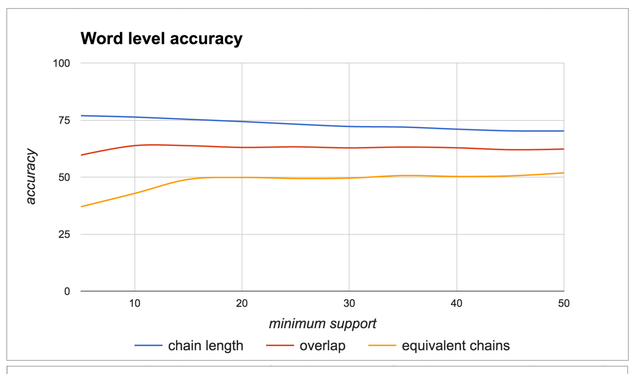
[2] C. Li and J. Wang. Efficiently mining closed subsequences with gap constraints. In *SDM '08*

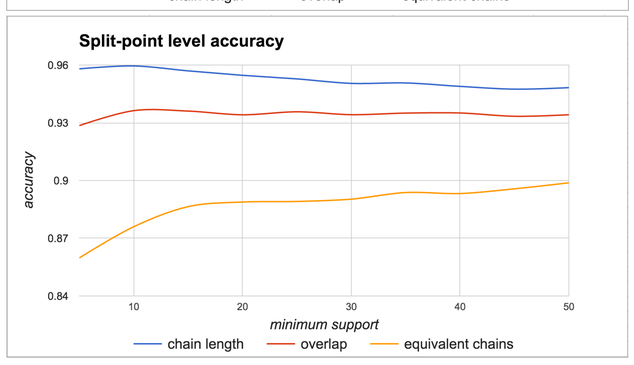
[3] Moby Hyphenator: http://icon.shef.ac.uk/Moby/mhyph.html

RESULTS

For Romanian, RoSyllabiDict [1] was employed. We sampled randomly:

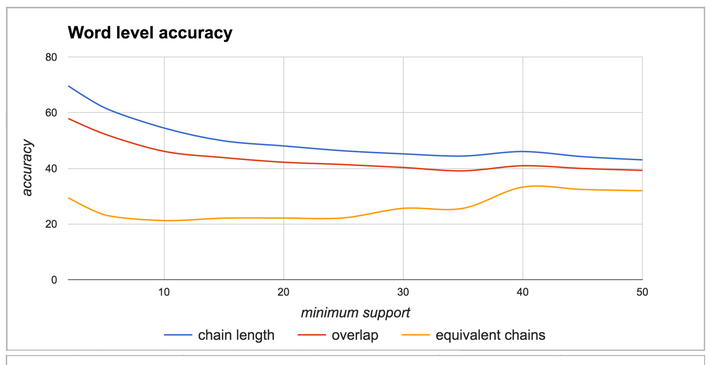
- 200.000 words for training
- 10.000 words for evaluation

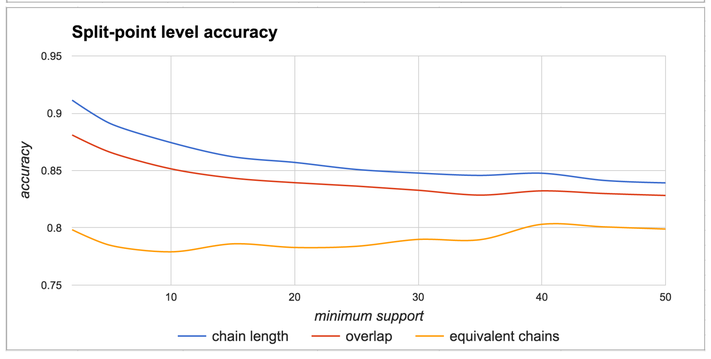




For English, we employed a dataset of ~190.000 words [3]. We sampled randomly:

- ~180.000 words for training
- 5.000 words for evaluation





SYLLABLE PATTERNS

The number of patterns found in RoSyllabiDict [1] (525486 Romanian syllabified words), by varying the minimum support:

Supp.	Len. 2	Len. 3	Len. 4	Len. 5
100	6754	2309	162	2
50	12671	7453	853	47
20	25731	28056	5384	674
10	41126	63388	17812	2940
5	63443	123429	52271	10553
2	104254	248227	186623	66915

For the pattern identification part we employed an implementation of the gapBIDE algorithm [2].

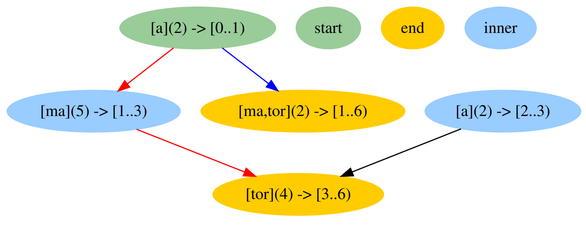
A WALKTHROUGH EXAMPLE

Syllabification of the word ***amator***:

1. Find matched patterns:

Pattern	Matching
$(< a >, 2)$	$[0, 1), [2, 3)$
$(< ma >, 5)$	$[2, 4)$
$(< ma, tor >, 2)$	$[2, 6)$
$(< tor >, 4)$	$[4, 6)$

2. Classify patterns and build pattern graph:



3. Choose closed pattern chain (red or blue):

- Equivalence classes $\Longrightarrow$ there is only one equivalence class $\Longrightarrow$ ***a-ma-tor*** ✓
- Overlapping $\Longrightarrow$ none of the chains have overlapping $\Longrightarrow$ ***a-ma-tor*** ✓
- Chain length $\Longrightarrow$ blue $\Longrightarrow$ ***a-ma-tor*** ✓

FUTURE DIRECTIONS

Preliminary evaluations performed on both Romanian and English words indicate that the method has potential. We believe that we can produce a significant boost in accuracy by providing a solution for syllabifying words for which there is no closed chain of patterns found. Also, none of the three syllable boundary prediction strategies employs support information in the prediction process. We believe considering such information will further boost the accuracy.

Furthermore, we intend to evaluate the solution on a broader range of languages (Hungarian is one potential candidate, as it is not an Indo-European language) and we are working on evaluating the impact of special characters (such as diacritics in Romanian) on the performance of our approach.

SOURCE CODE

https://github.com/adrianulbona/rosil

ACKNOWLEDGEMENT

The work presented in this paper is partially supported by the Romanian Ministry of Education, under grant agreement PN-II-PT-PCCA-2013-4-1660 (SWARA)