ADBIS 2018

September 2-5
Budapest Hungary

Integrating Approximate String Matching with Phonetic Similarity

(UFPR) Junior Ferri `junior.ferri@ufpr.br`

(UFPR) Dr Hegler Tissot **`hegler@gmail.com`**

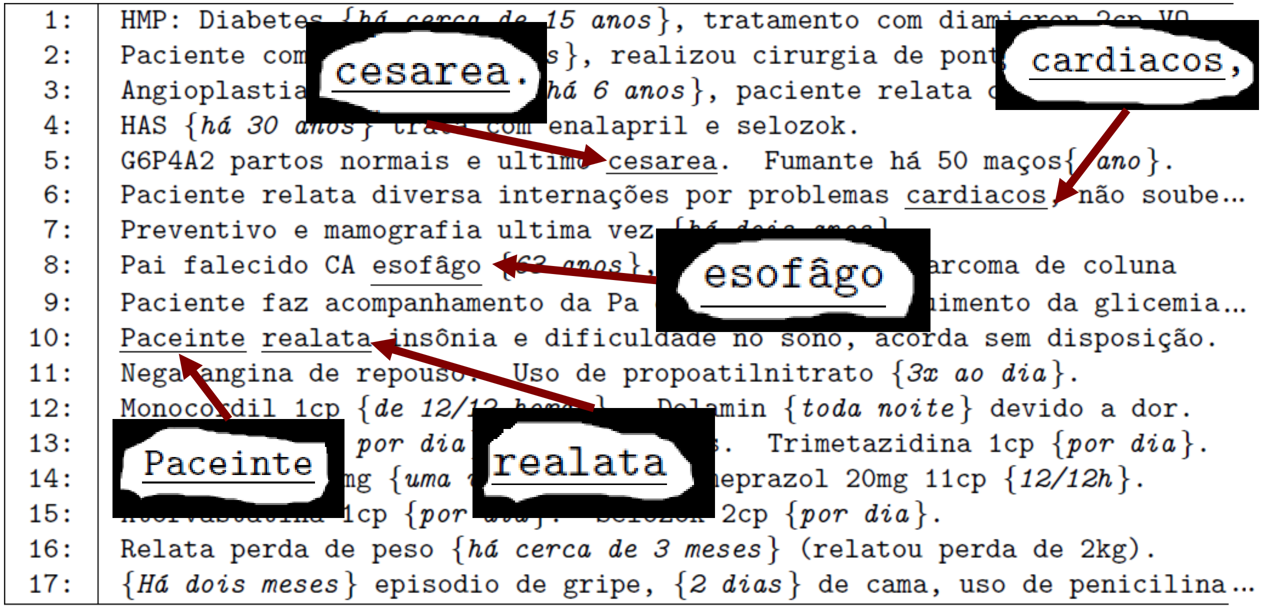
(UFPR) Dr Marcos D Del Fabro `marcos.ddf@inf.ufpr.br`

Context

- Natural Language Processing (NLP)
- Named Entity Recognition (NER)
  - Organizations
  - Cities and Countries
  - Names of Drugs
- A usual approach
  - Dictionary-based (gazetteer) – exact match

Problem

- Imperfections (spelling errors)



HMP: Diabetes {*há cerca de 15 anos*}, tratamento com diamicron 2cp VO
Paciente com s}, realizou cirurgia de pont
Angioplastia *há 6 anos*}, paciente relata c
HAS {*há 30 anos*} trata com enalapril e selozok.
G6P4A2 partos normais e ultimo cesarea. Fumante há 50 maços{ *ano*}.
Paciente relata diversa internações por problemas cardiacos, não soube...
Preventivo e mamografia ultima vez {*há dois anos*}
Pai falecido CA esofâgo {*63 anos*}, arcoma de coluna
Paciente faz acompanhamento da Pa imento da glicemia...
Paceinte realata insônia e dificuldade no sono, acorda sem disposição.
Nega angina de repouso. Uso de propoatilnitrato {*3x ao dia*}.
Monocordil 1cp {*de 12/12 horas*}. Dolamin {*toda noite*} devido a dor.
por dia s. Trimetazidina 1cp {*por dia*}.
mg {*uma* eprazol 20mg 11cp {*12/12h*}.
1cp {*por dia*}. Selozok 2cp {*por dia*}.
Relata perda de peso {*há cerca de 3 meses*} (relatou perda de 2kg).
{*Há dois meses*} episodio de gripe, {*2 dias*} de cama, uso de penicilina...

- Approximate String Matching (ASM)
  - Indexed (fast) search + Edit Distance

Problem

- ED does not capture all the aspects of text imperfections, such as phonetic dissimilarity.

- Phonetics is a language-dependent problem

- Hybrid string and phonetic similarity approaches can lead to more consistent results compared to traditional ED-based methods [Tissot et al., DEXA-2014]

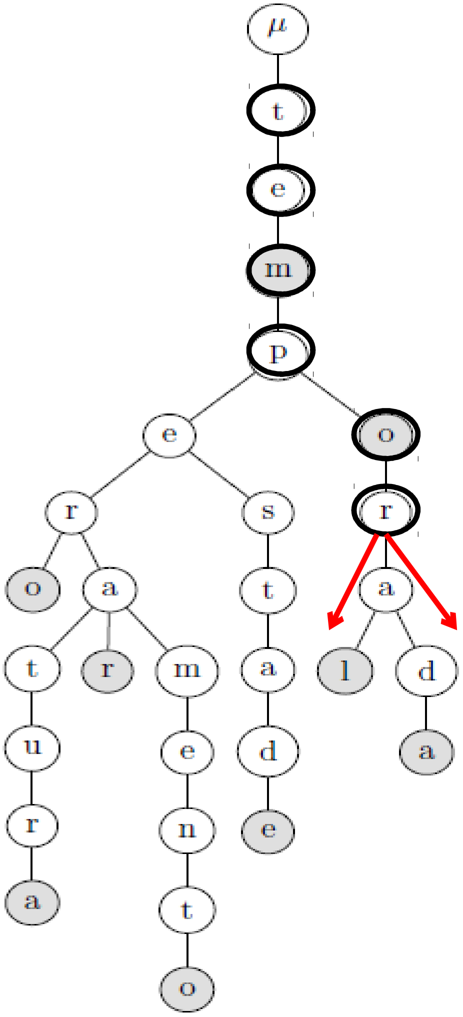
Method

- Task:
  - NER

    Matching phonetically similar words
- *TRIE: indexed search*
  - *ED: String similarity threshold*

    Used to auto-complete
- *Language-dependent components*
  - *Phonetic representation (Metaphone)*
    - *more compact (memory optimization)*
  - *String Similarity Function*

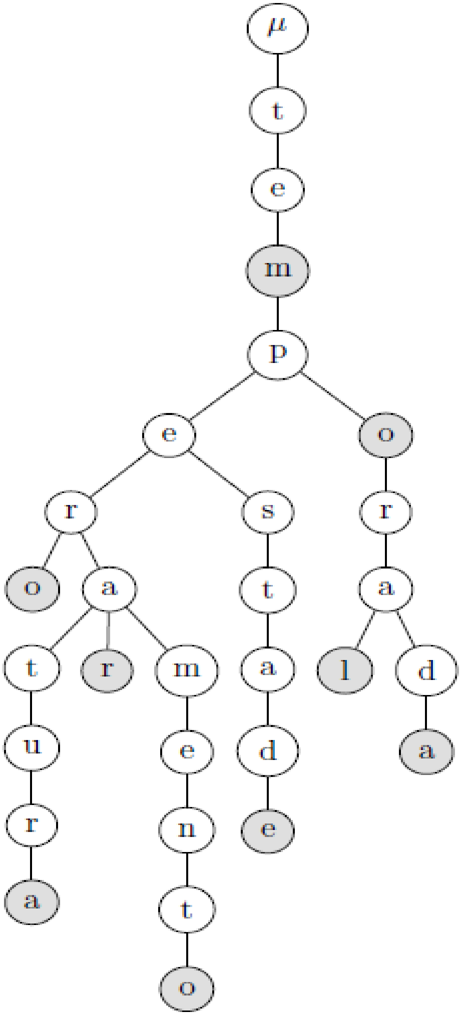
[Tissot et al., DEXA-2014]



Method

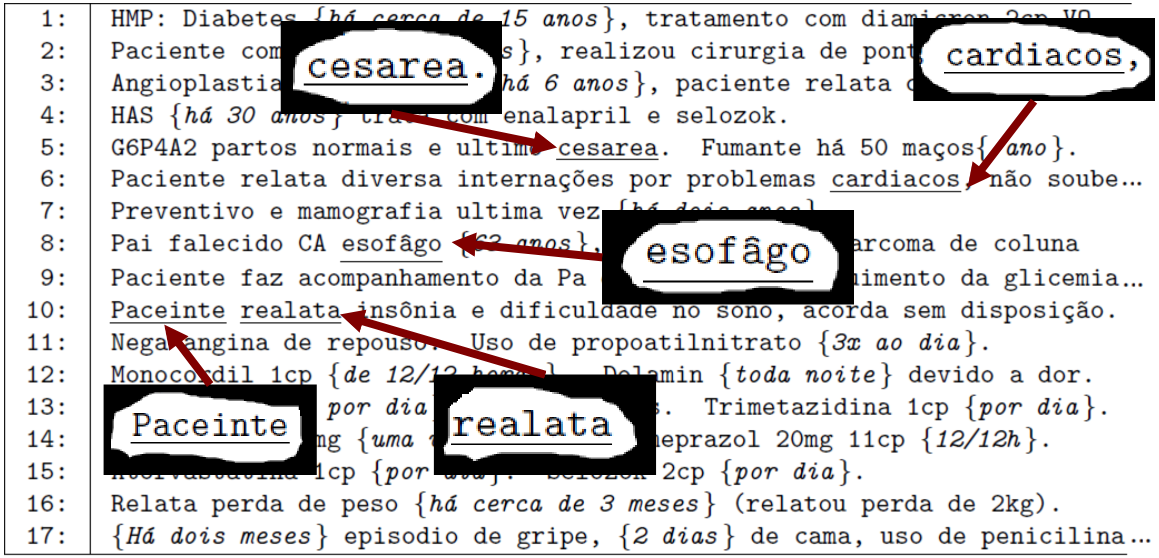
(A) Both gazetteer entries and text are converted to a phonetic representation (Metaphone)

Word	Phonetic representation
medroxalol	MTRKSLL
amoxicillin	AMKSSLN
bromfenac	BRMFNK
New York	NYRK
Avondale Estates	AFNTLSTTS
Washington	WXNKTN



Method

(B) TRIE search combines the phonetic representation with a Edit Distance threshold (ED <= 2)



1: HMP: Diabetes {*há cerca de 15 anos*}, tratamento com diamicron 2cp VO

2: Paciente com s}, realizou cirurgia de pont

3: Angioplastia *há 6 anos*}, paciente relata c

4: HAS {*há 30 anos*} trata com enalapril e selozok.

5: G6P4A2 partos normais e ultimo cesarea. Fumante há 50 maços{*ano*}.

6: Paciente relata diversa internações por problemas cardiacos, não soube...

7: Preventivo e mamografia ultima vez {*há dois anos*}.

8: Pai falecido CA esofâgo {*63 anos*}, arcoma de coluna

9: Paciente faz acompanhamento da Pa imento da glicemia...

10: Paceinte realata insônia e dificuldade no sono, acorda sem disposição.

11: Nega angina de repouso. Uso de propoatilnitrato {*3x ao dia*}.

12: Monocordil 1cp {*de 12/12 horas*}. Dolamin {*toda noite*} devido a dor.

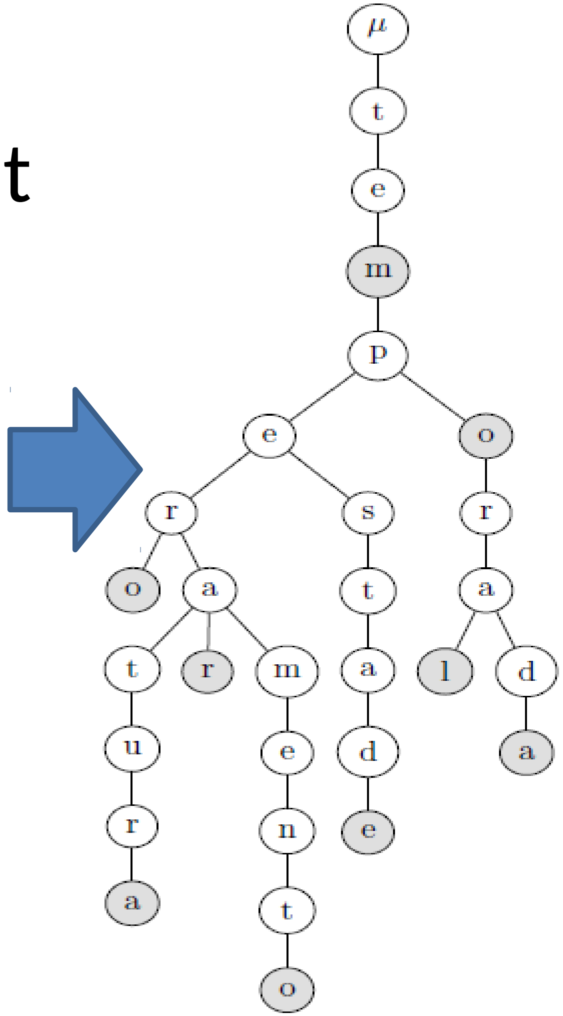
13: *por dia* s. Trimetazidina 1cp {*por dia*}.

14: mg {*uma* eprazol 20mg 11cp {*12/12h*}.

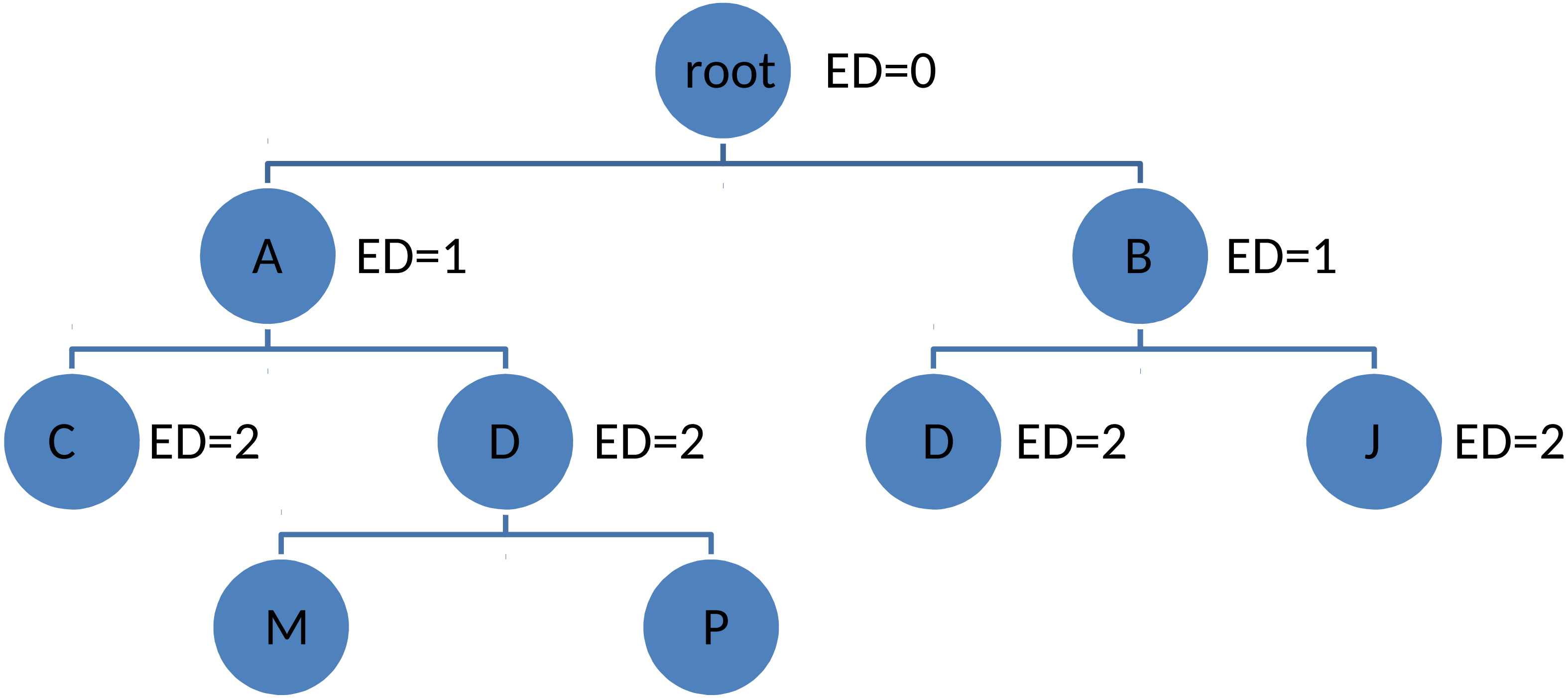
15: Atorvastatina 1cp {*por dia*}. Selozok 2cp {*por dia*}.

16: Relata perda de peso {*há cerca de 3 meses*} (relatou perda de 2kg).

17: {*Há dois meses*} episodio de gripe, {*2 dias*} de cama, uso de penicilina...

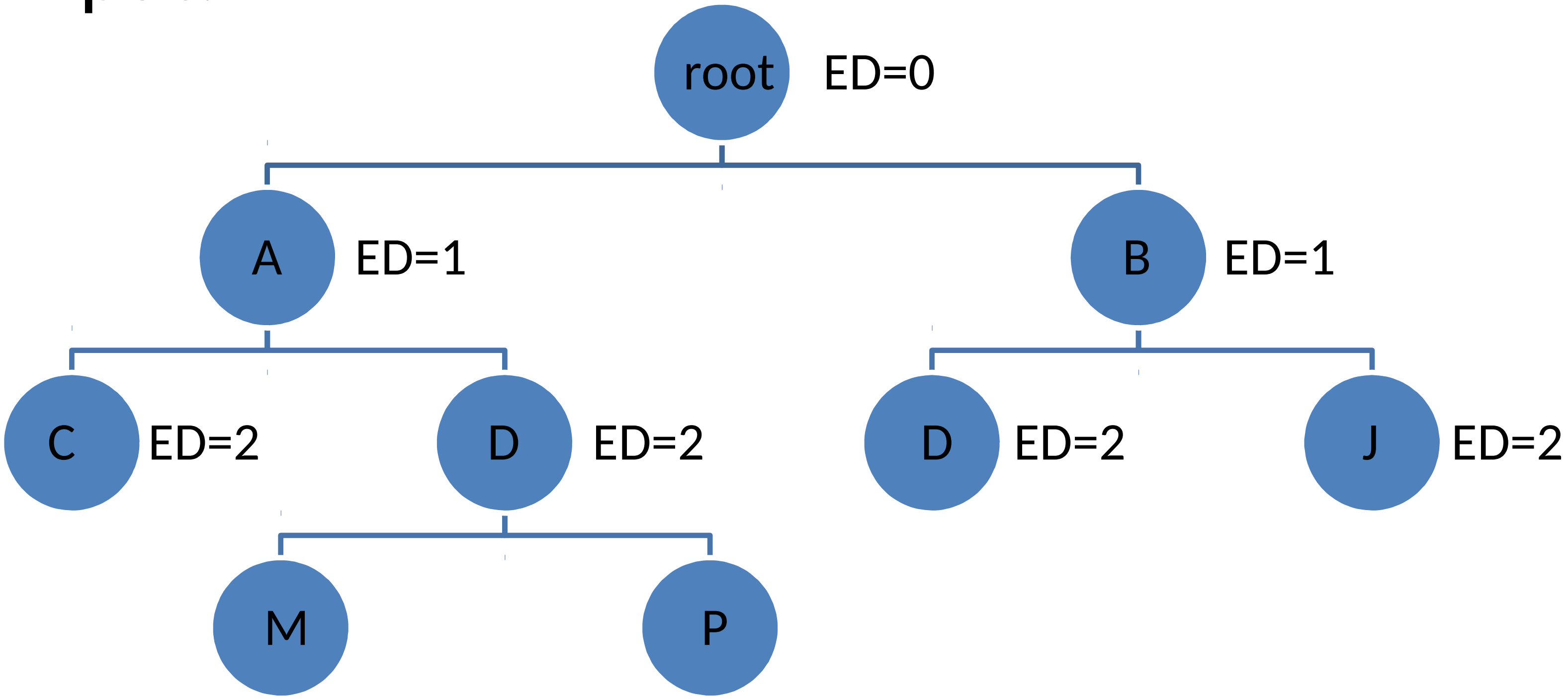


Method



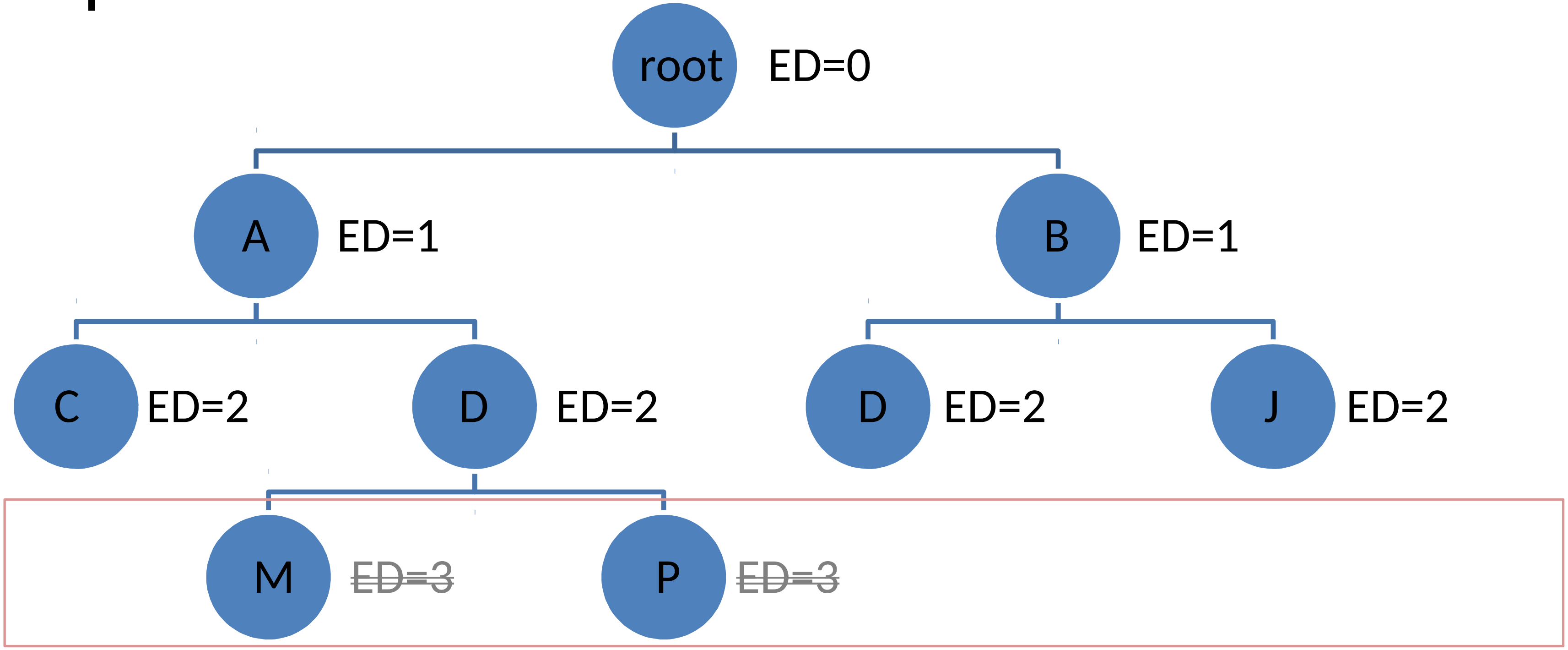
Method

Input: D



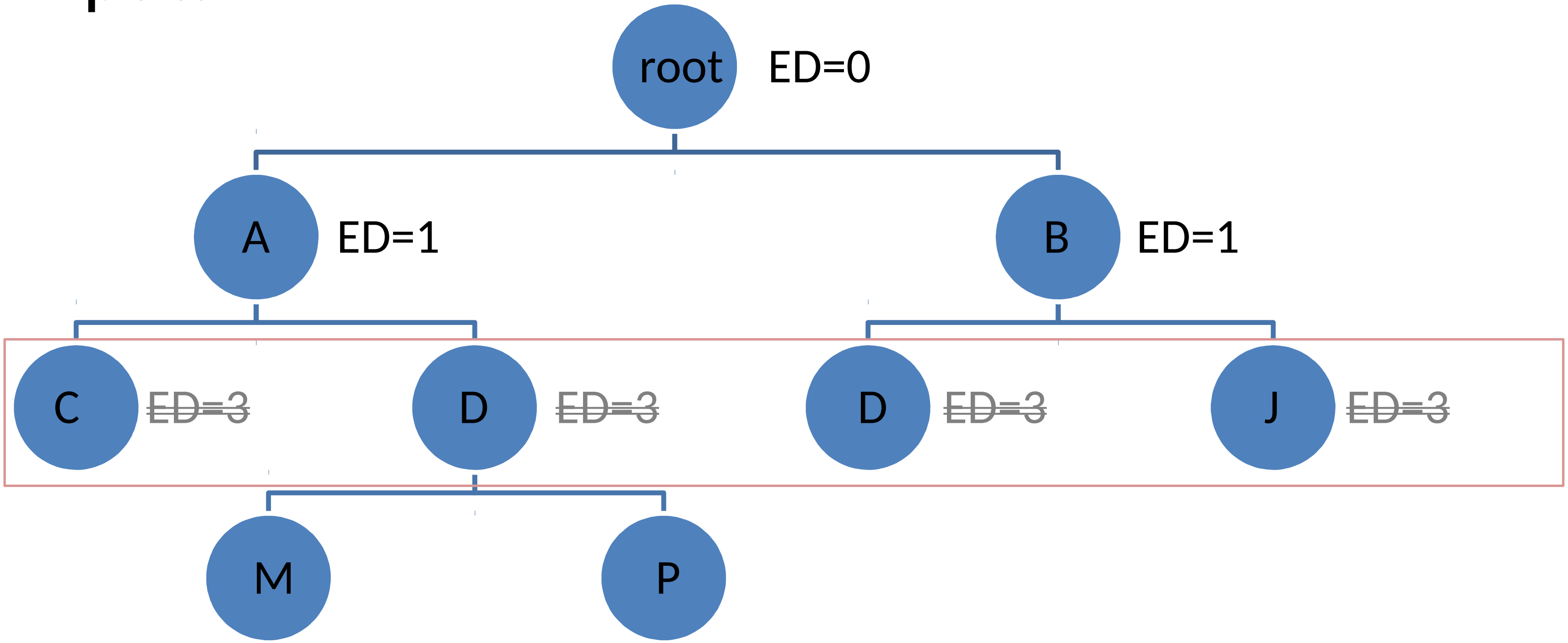
Method

Input: D



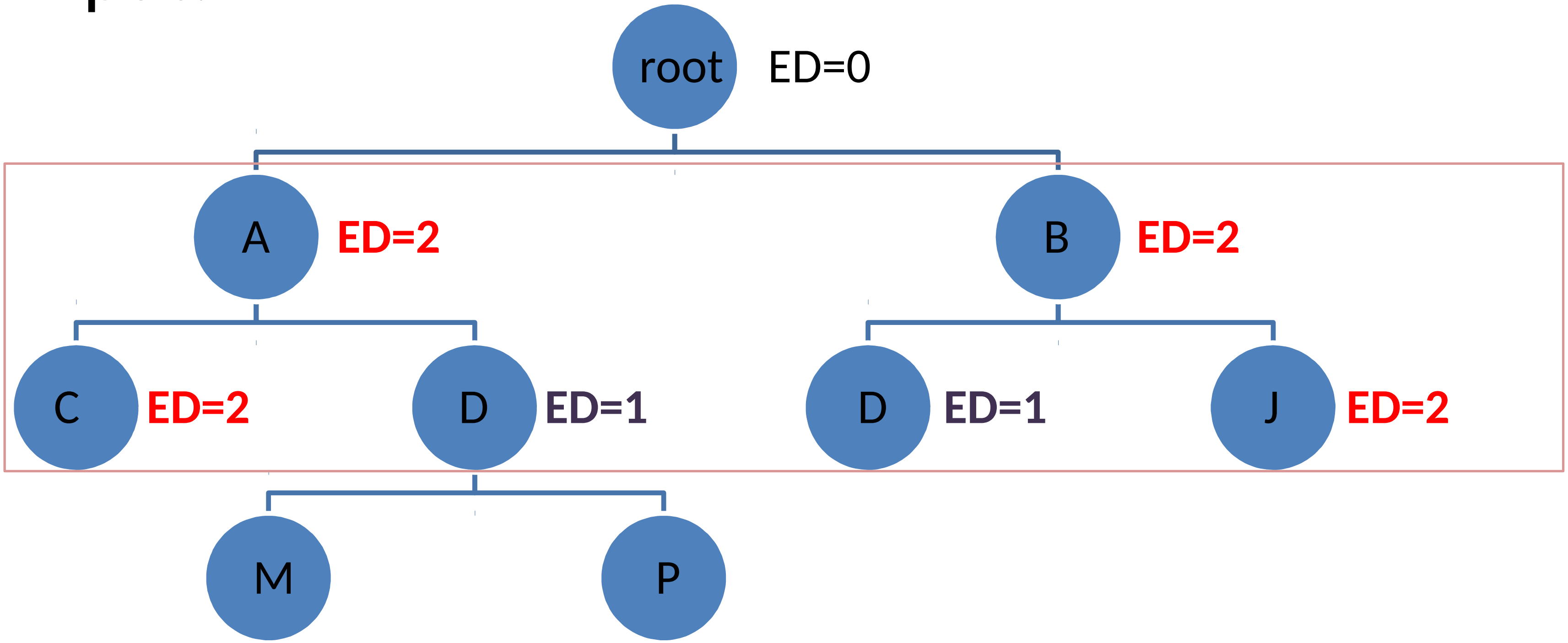
Method

Input: D



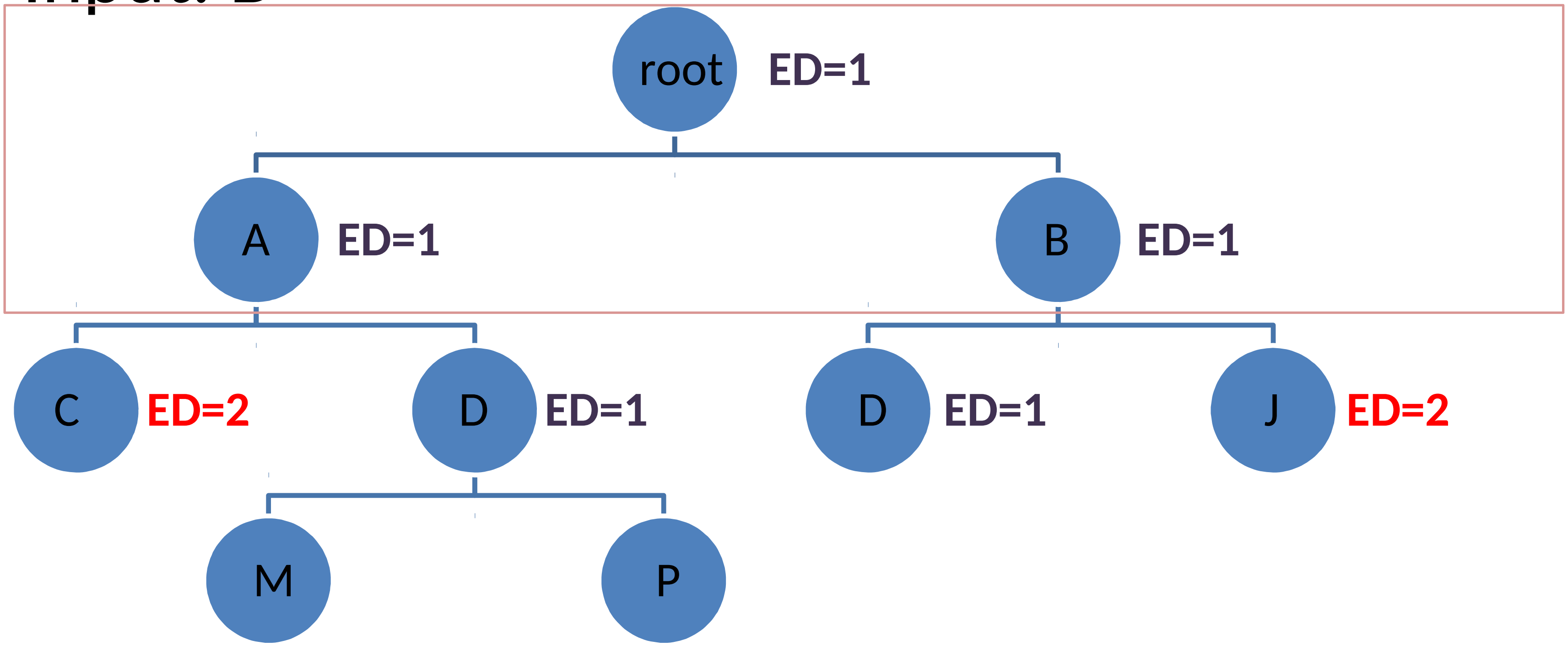
Method

Input: D



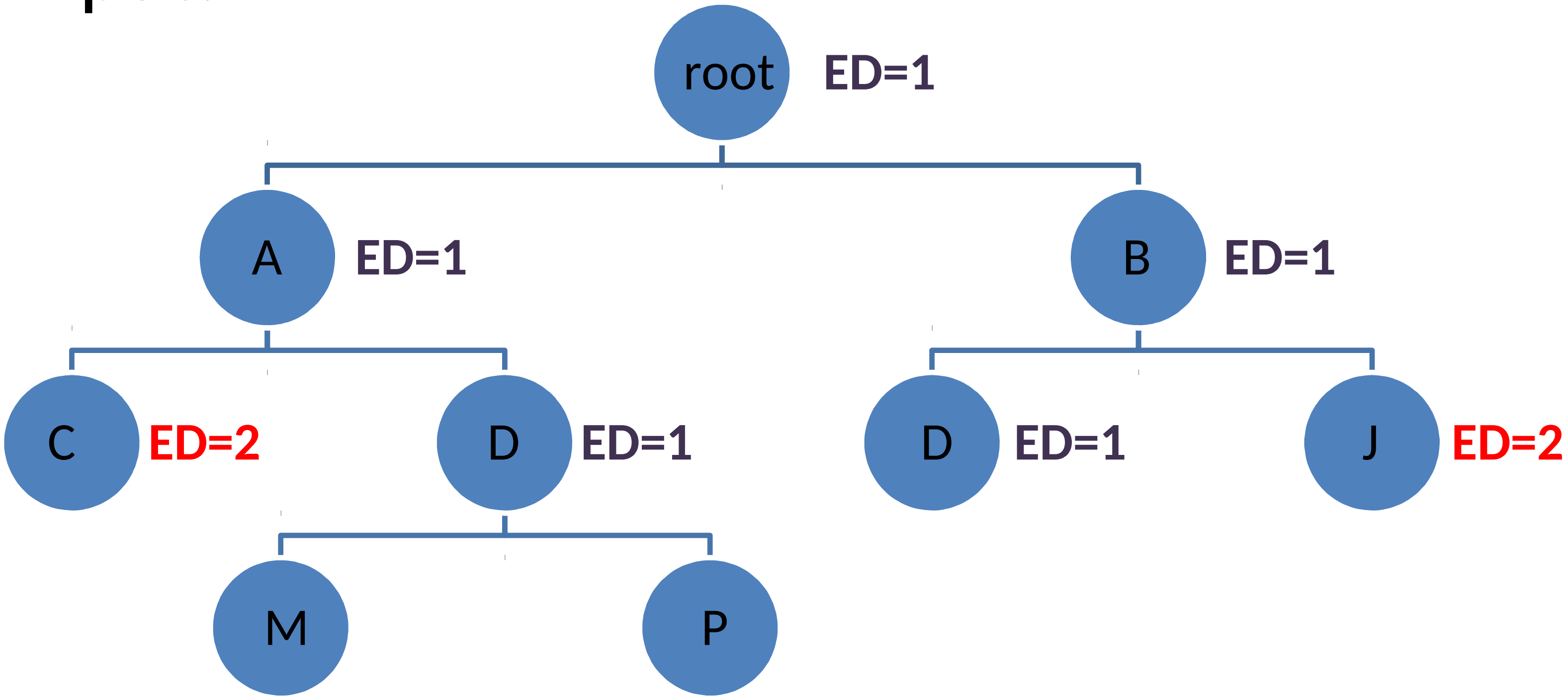
Method

Input: D



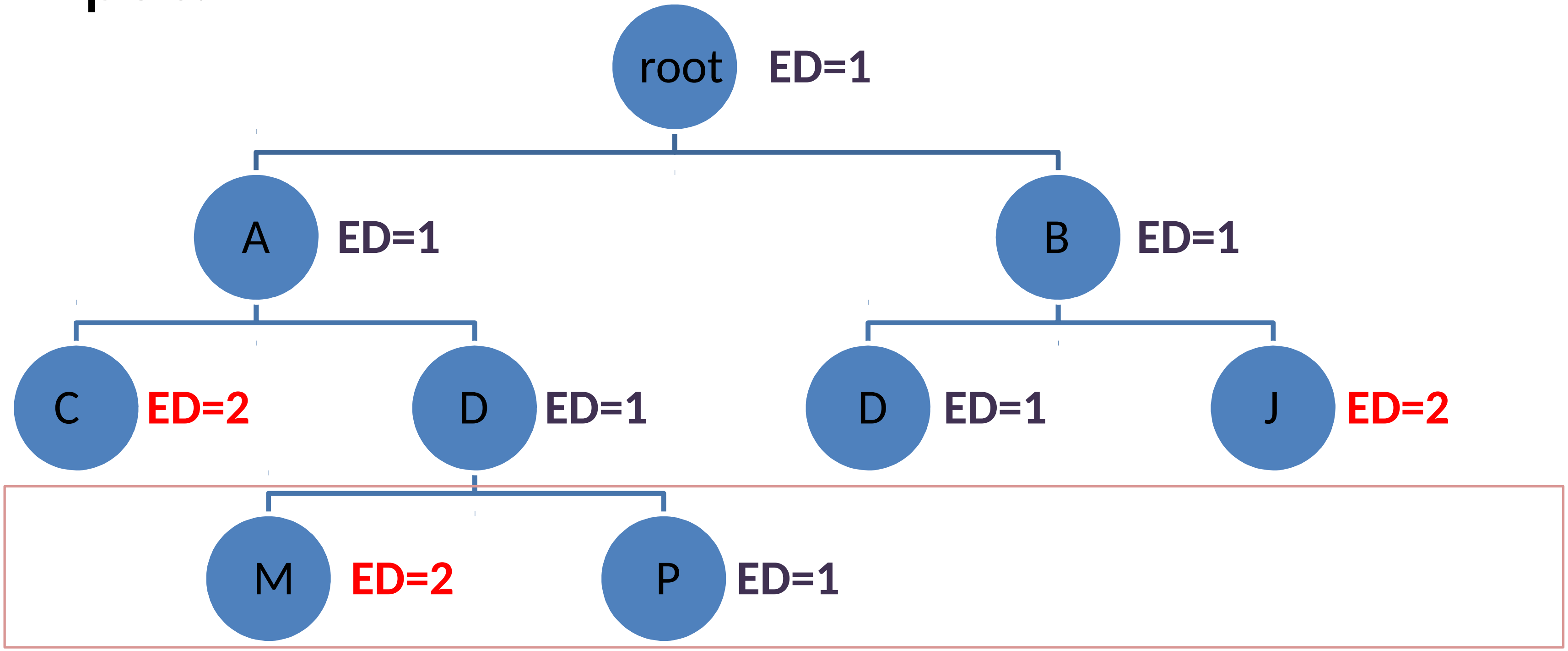
Method

Input: DP



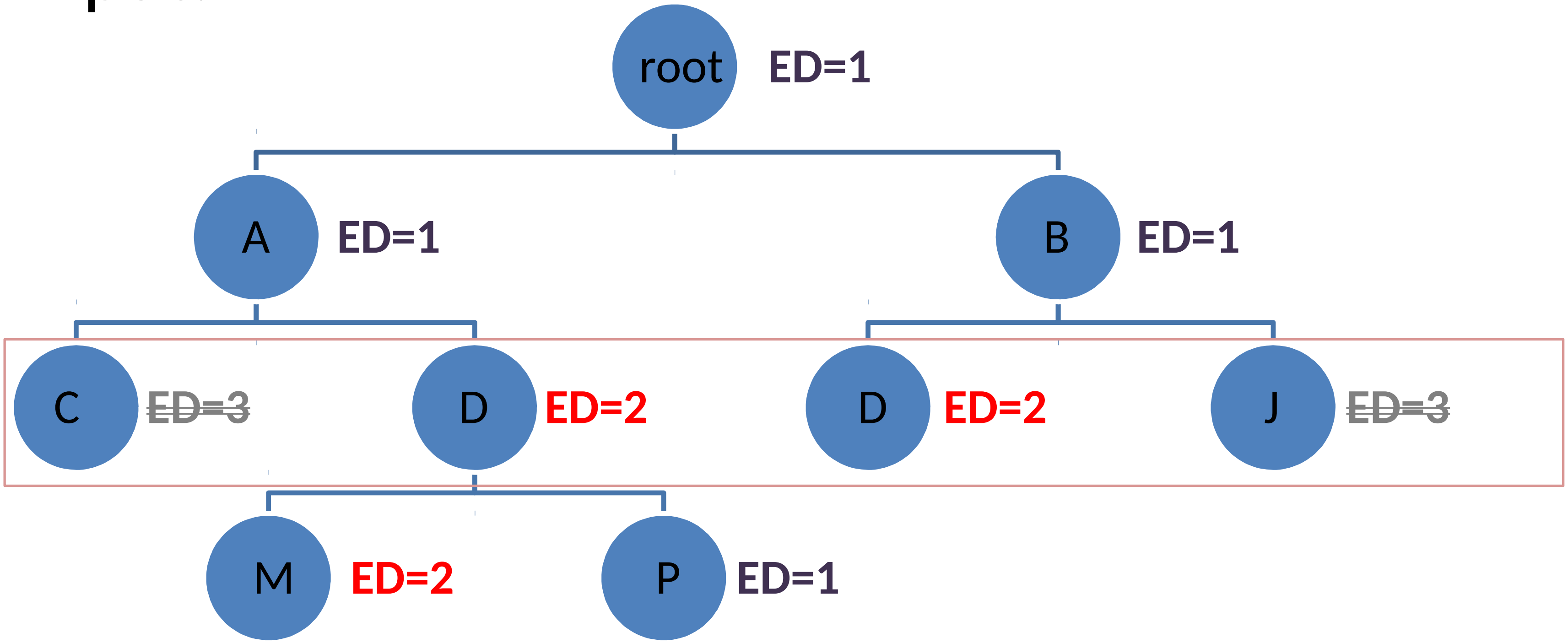
Method

Input: DP



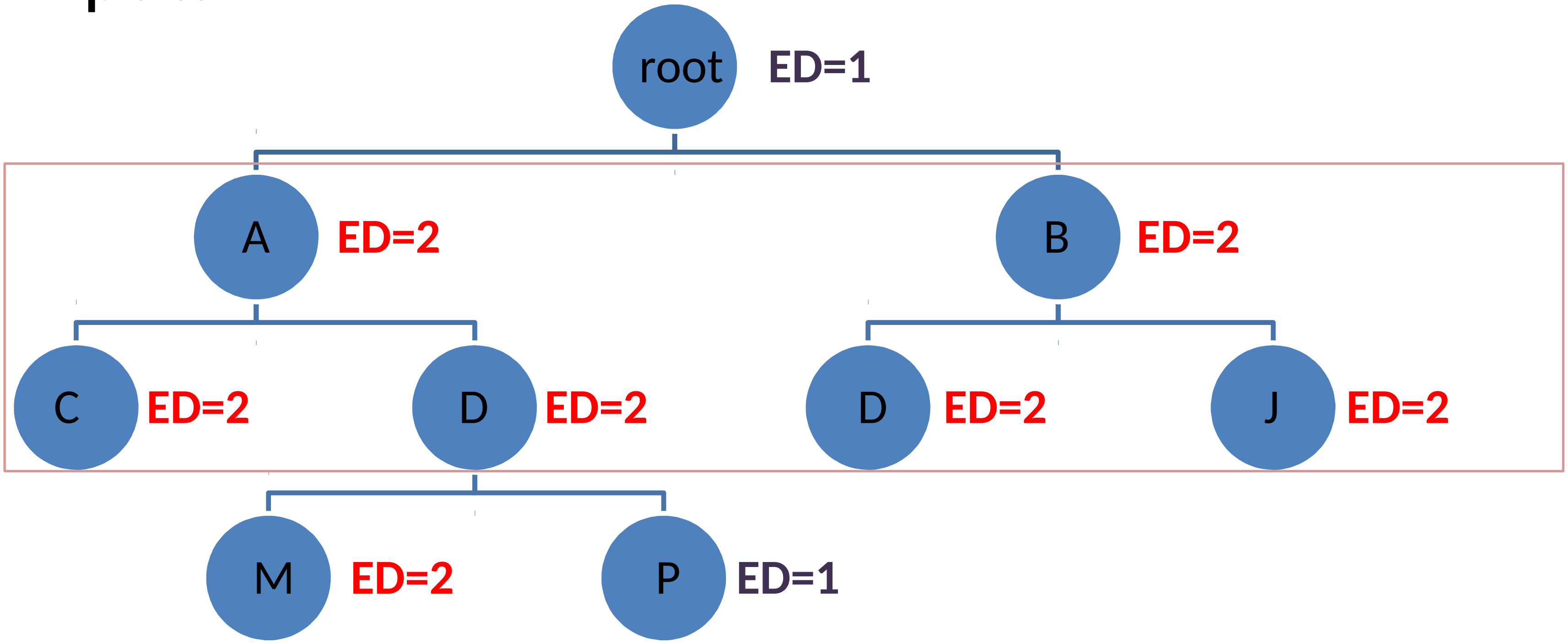
Method

Input: DP



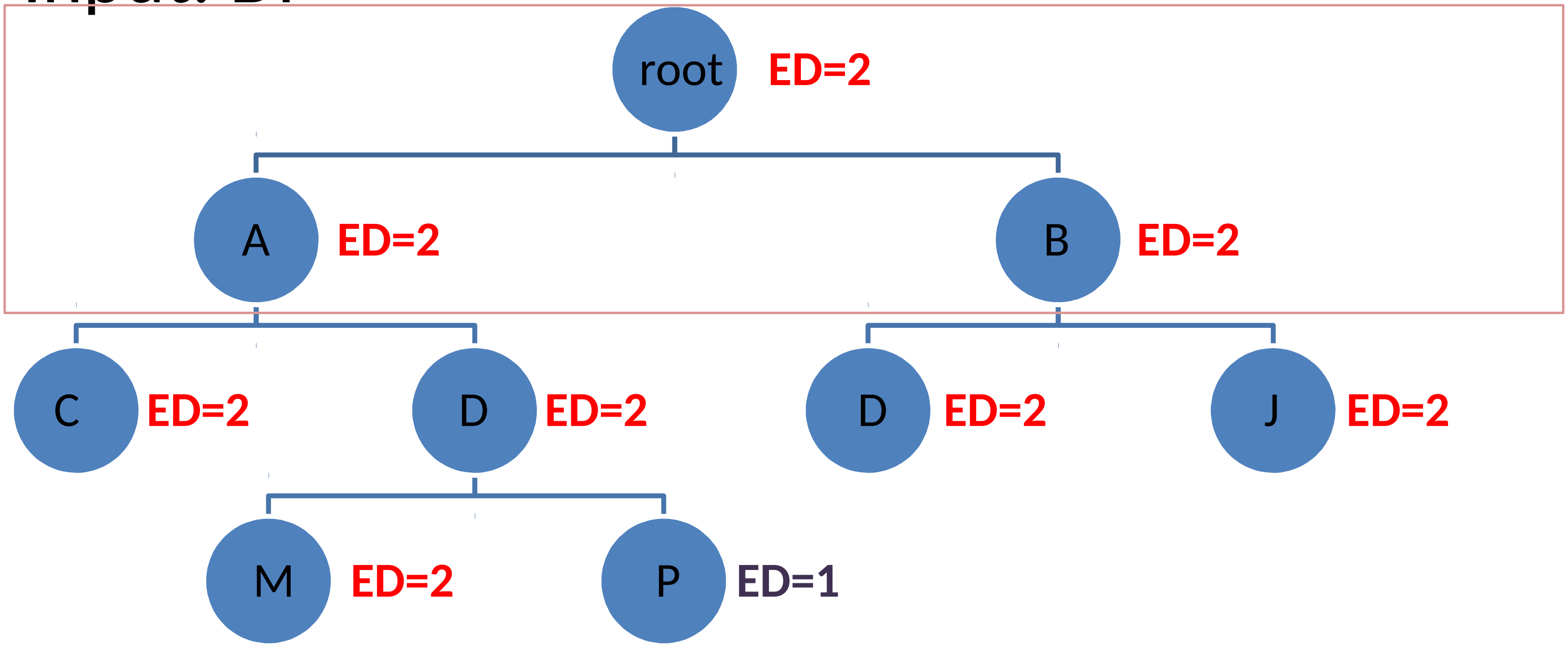
Method

Input: DP



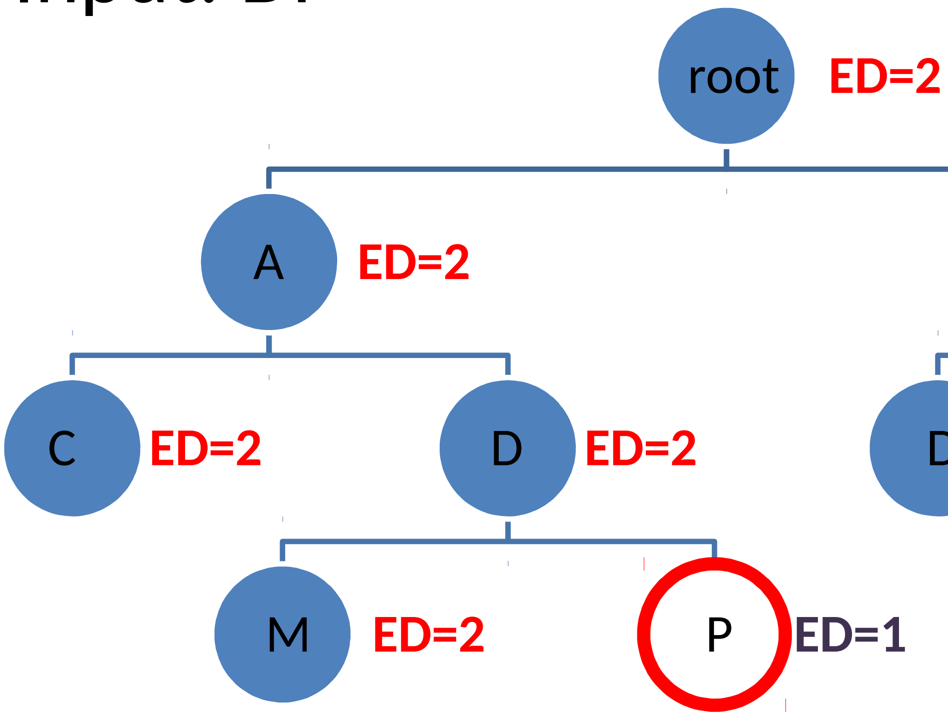
Method

Input: DP



Method

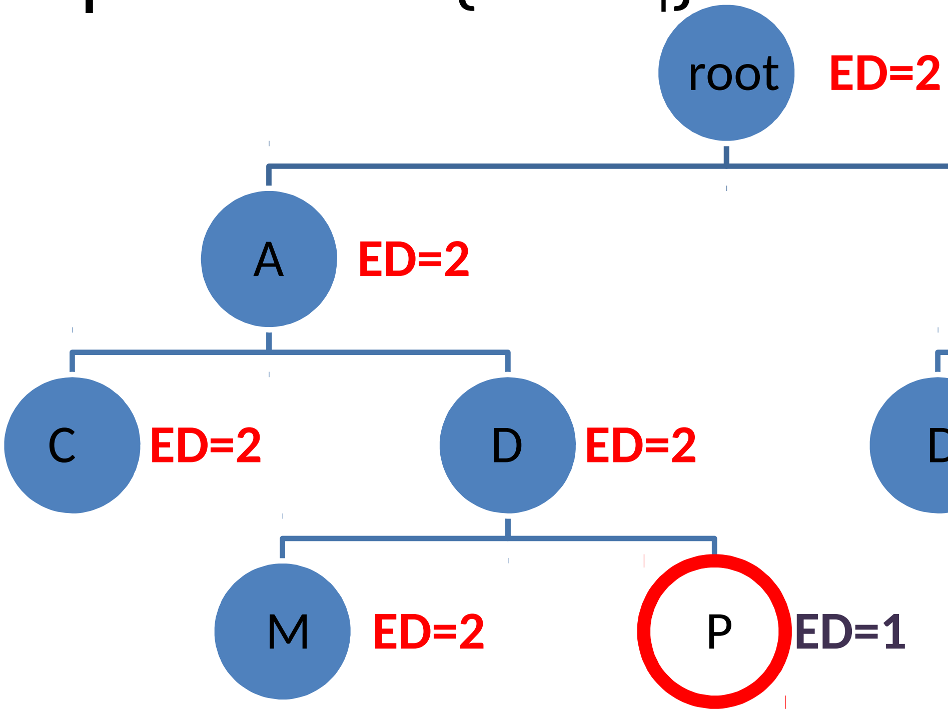
Input: DP



TRIE: ADP (ED=1)

Method

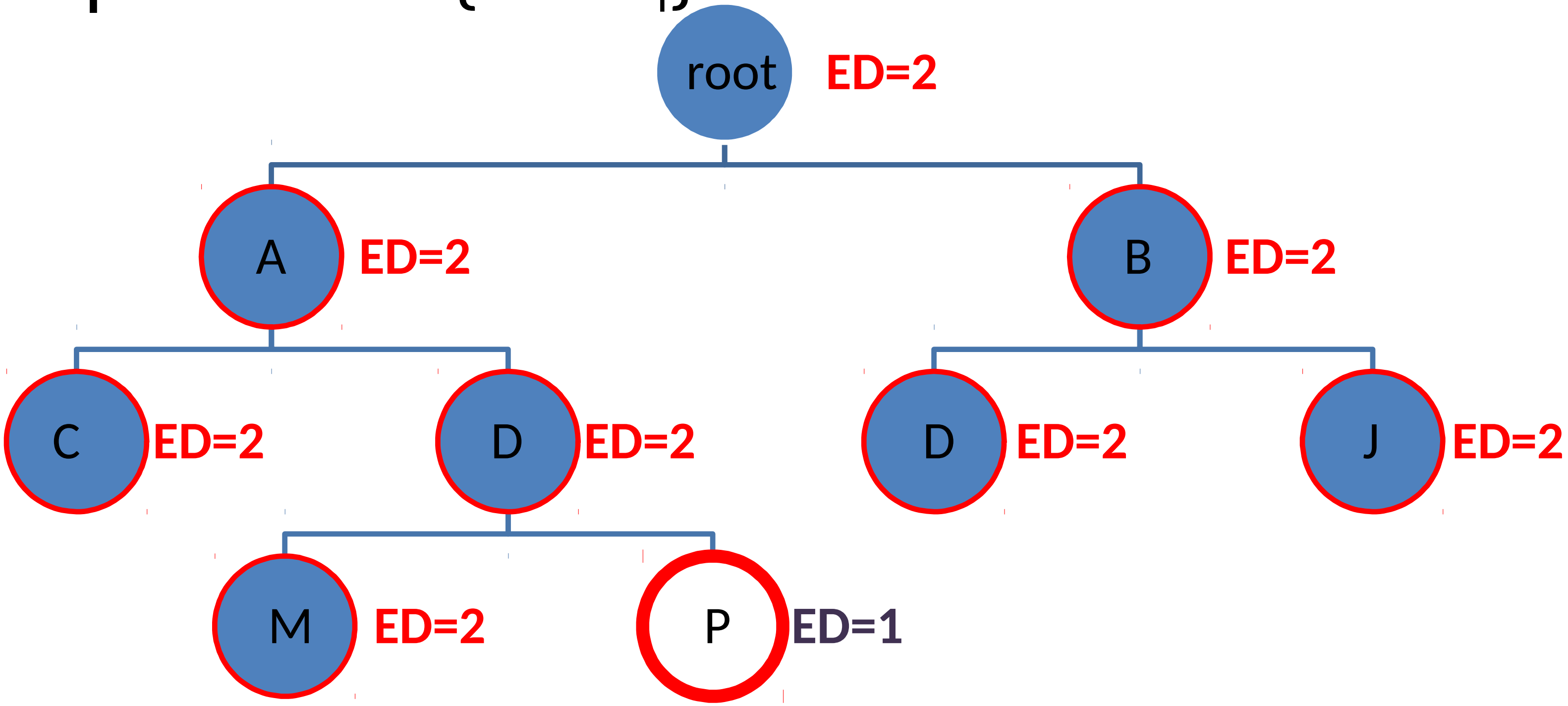
Input: DP → $\{word_i\}$



TRIE: ADP (ED=1) → $\{word_1, word_2,\ldots,word_n\}$

Method

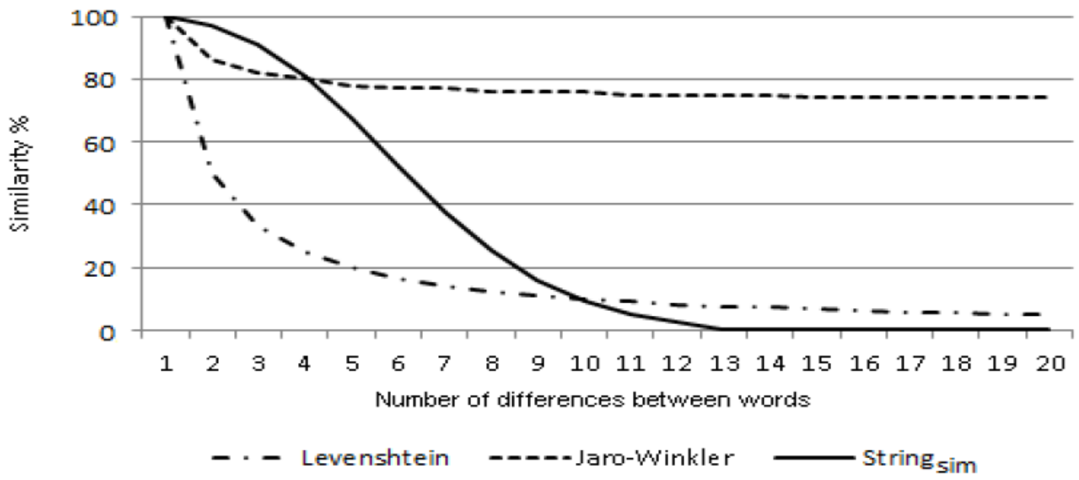
Input: DP → {word$_i$}



TRIE: ADP (ED=1) → {word$_1$, word$_2$,…,word$_n$}

Method

(C) A string similarity function is used to finally filter out possible false positive matches



ED vs Jaro-Winkler vs **String**$_{\mathbf{Sim}}$

Validating Experiment

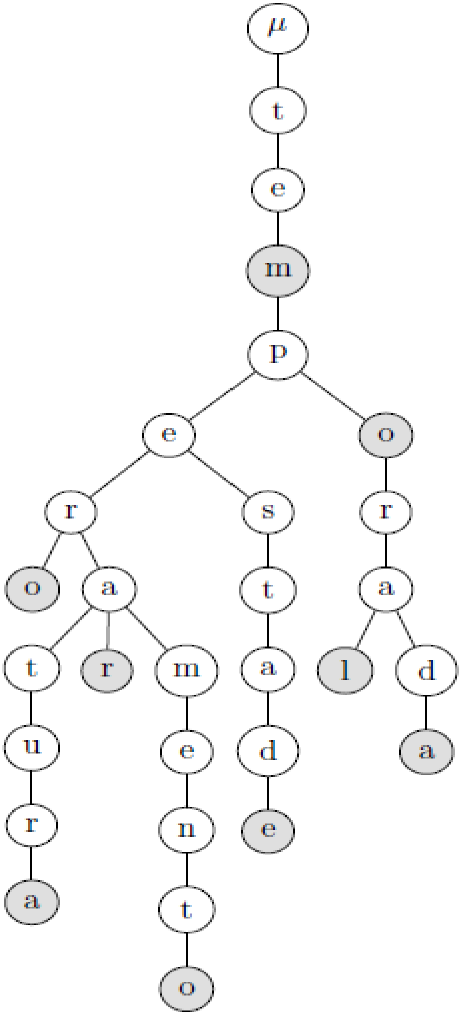
- 76,912 English words

  (WordNet)

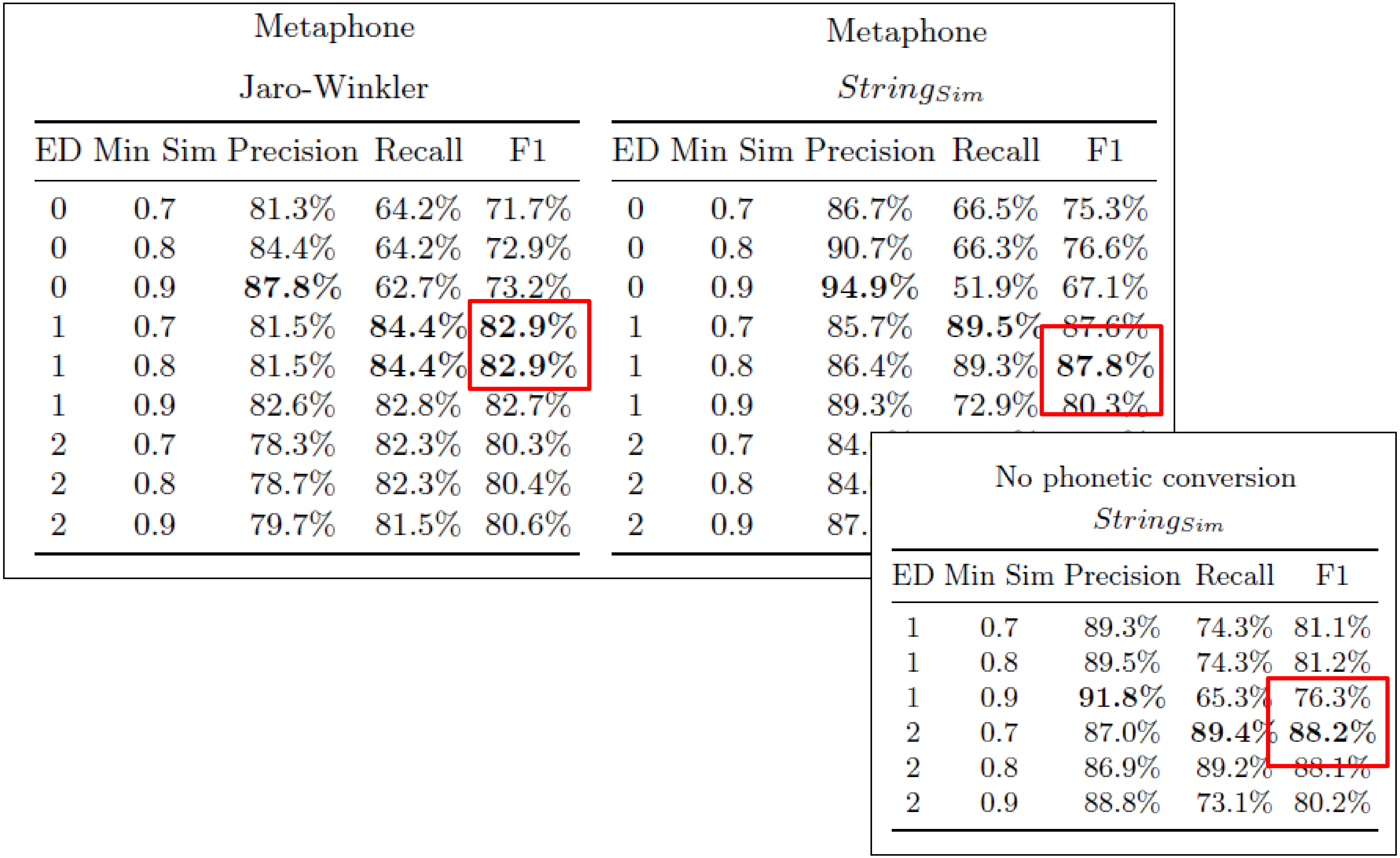
  vs

- Lists of 1,000 common misspellings

  (Wikipedia)



Validating Results



Metaphone

Jaro-Winkler

ED	Min Sim	Precision	Recall	F1
0	0.7	81.3%	64.2%	71.7%
0	0.8	84.4%	64.2%	72.9%
0	0.9	**87.8%**	62.7%	73.2%
1	0.7	81.5%	**84.4%**	**82.9%**
1	0.8	81.5%	**84.4%**	**82.9%**
1	0.9	82.6%	82.8%	82.7%
2	0.7	78.3%	82.3%	80.3%
2	0.8	78.7%	82.3%	80.4%
2	0.9	79.7%	81.5%	80.6%

Metaphone

$String_{Sim}$

ED	Min Sim	Precision	Recall	F1
0	0.7	86.7%	66.5%	75.3%
0	0.8	90.7%	66.3%	76.6%
0	0.9	**94.9%**	51.9%	67.1%
1	0.7	85.7%	**89.5%**	87.6%
1	0.8	86.4%	89.3%	**87.8%**
1	0.9	89.3%	72.9%	80.3%
2	0.7	84.		
2	0.8	84.		
2	0.9	87.		

No phonetic conversion

$String_{Sim}$

ED	Min Sim	Precision	Recall	F1
1	0.7	89.3%	74.3%	81.1%
1	0.8	89.5%	74.3%	81.2%
1	0.9	**91.8%**	65.3%	76.3%
2	0.7	87.0%	**89.4%**	**88.2%**
2	0.8	86.9%	89.2%	88.1%
2	0.9	88.8%	73.1%	80.2%

Conclusions

- Inexact Match
  - Information Extraction
  - Natural Language Processing

- Temporal Information Extraction
  - December: Dcember, Decmebr, Deceber, remember(?)
  - August: Augusto, Augustus (person or month?)

- Future work
  - Disambiguation

ADBIS 2018

September 2-5
Budapest Hungary

Integrating Approximate String Matching with Phonetic Similarity

(UFPR) Junior Ferri `junior.ferri@ufpr.br`

(UFPR) Dr Hegler Tissot `hegler@gmail.com`

(UFPR) Dr Marcos D Del Fabro `marcos.ddf@inf.ufpr.br`

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ