Statistically Augmented Preprocessing/Normalization Module for a Romanian Text-to-Speech System

Cătălin Ungurean, Dragoș Burileanu, Mihai Surmei

Speech and Dialogue (SpeeD) Laboratory

Faculty of Electronics, Telecommunications and Information Technology

University "Politehnica" of Bucharest, Romania

Importance for TTS

- The preprocessor is the first level in a TTS system
  - Responsible for text segmentation
  - Necessary for unitary text representation
- Helps all the following stages:
  - Diacritic restoration
  - Phonetic conversion
  - Prosody generation, etc.

Importance for TTS

- Commercial TTS can have both naturalness and intelligibility issues due to incomplete preprocessing
- The tendency is to use more acronyms/abbreviations, especially for mobile TTS applications
- Acronym/abbreviation (AC/AB) arrival is a productive process
- Lack of extensive acronyms/abbreviations dictionaries and improper end sentence detection can alter both prosody and message
- When unseen acronyms/abbreviations expanding cannot be done => the TTS solution is spelling: S.R.L. (se re le) but the detection is necessary

The problems of preprocessing

- Numeral representation:
  - 123 ⇒ *o sută douăzeci și trei (one hundred and thirty three)*
  - 07234567892 ⇒ *zero șapte doi trei patru ...*
    (zero seven two three four ...)
- Acronyms/abbreviations must be expanded:
  - dr ⇒ *doctor*
  - mg ⇒ *milligram*
- Date representation:
  - 20.08.2008 ⇒ *douăzeci august două mii opt*
    (eight of august two thousand and eight)
- The extraction of graphical signs for prosody (i.e., period marks, etc.)
- Sentence boundary detection (SBD)
- Word segmentation

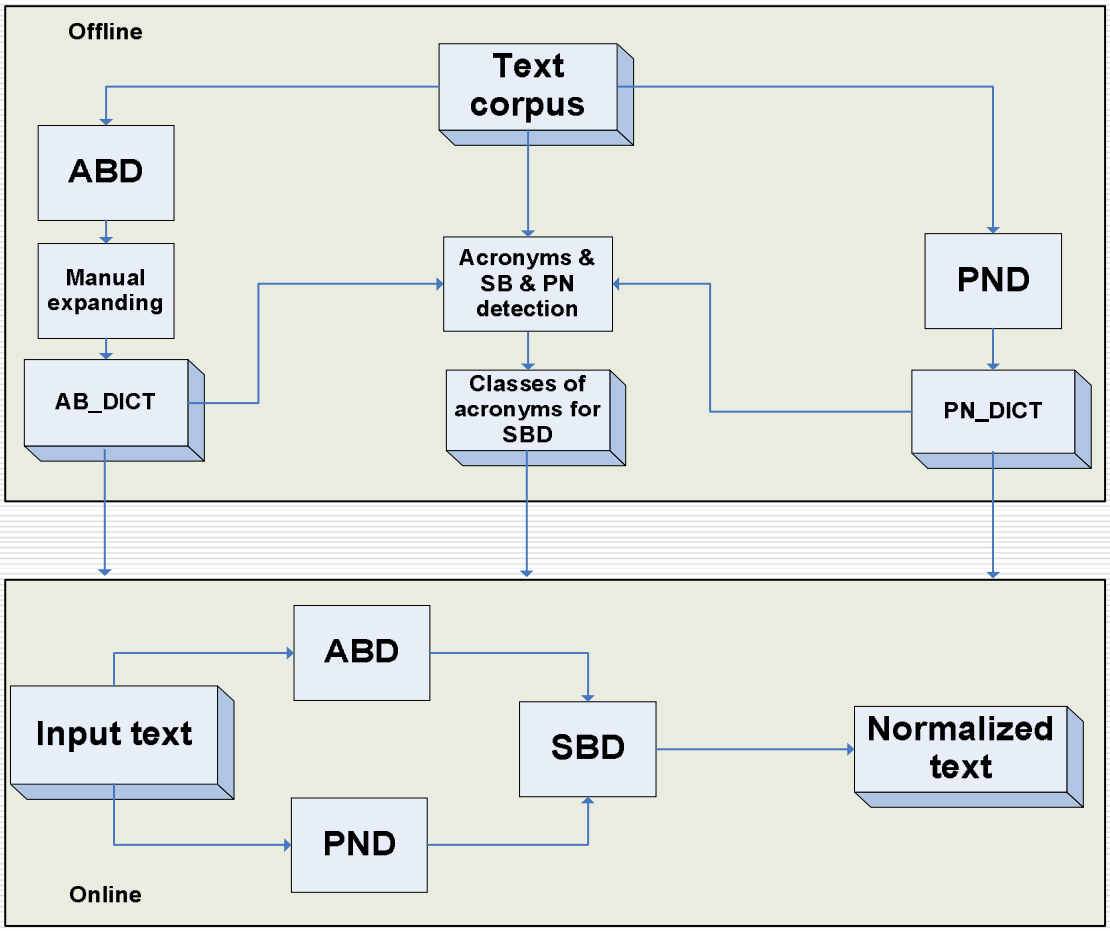
The problems of preprocessing

- Rule-based preprocessor is limited:
  - Solve date, numeral representation, multiple period acronyms, clear SBD (period mark, etc.)
- The most difficult problem is an abbreviation (AB) followed by period, followed by proper name (PN):
  ... AB. PN ...

  e.g., L-am sunat pe dr. Marcu să vină la mine.

  (I called dr. Marcu to come to me.)
  - The period can be a mark for abbreviation, sentence end, or both
  - Multiple period (MP) acronyms (MP+PN, 2MP+PN, 3MP+PN) can occur: e.g., I.L. Caragiale, I.H. Rădulescu, R.A.T.B., etc.

General algorithm description



- Interdependence between:
  - ABD - acronym/ abbreviation detection
  - PND - proper name detection
  - SBD - sentence boundary detection
- Two layers:
  - Offline: used to enlarge dictionaries
  - Online: applied on the input text

Linguistic resources

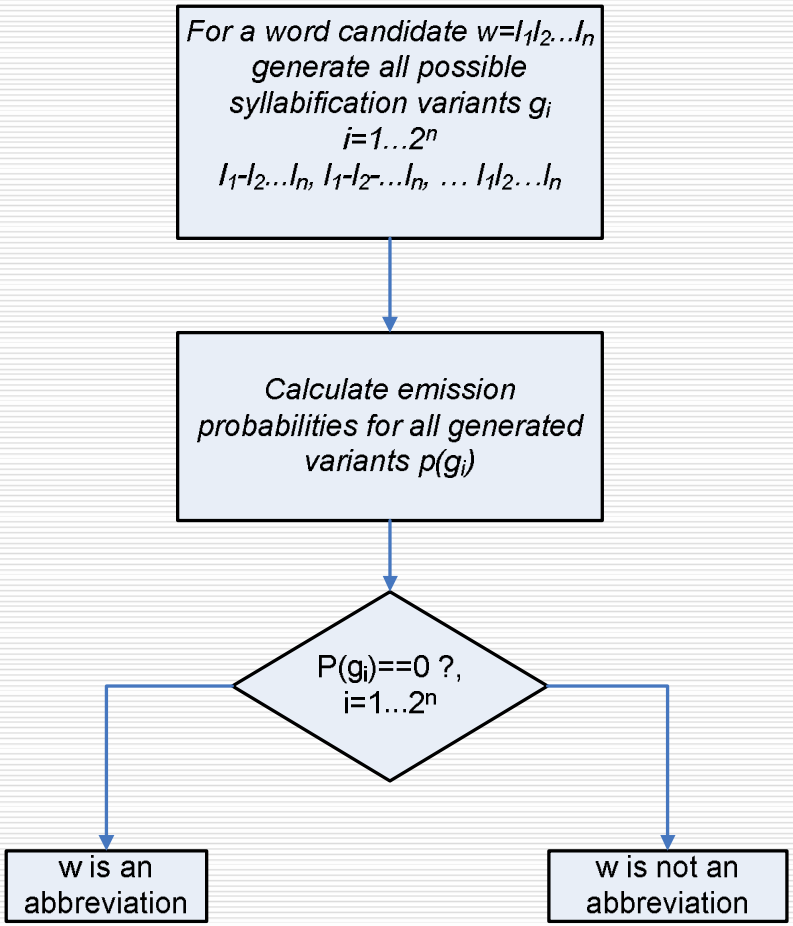
- **9am corpus** – 3.5×10^6 phrases, 63×10^6 tokens in Romanian
- **TS_1 ÷ TS_5** – test sets randomly extracted from **9am**, count in between 125.000-190.000 sentences
- **9am_res** – a remaining set, used to enlarge offline AB_DICT and PN_DICT
- **TS_ABD** – test set for abbreviations/acronyms detection (ABD), contains 200 AC/AB and 200 common words
- **SYL_DICT** – 55,000 correctly syllabified Romanian words for ABD training

- **AB_DICT** – acronyms/abbreviations dictionary, with expanded forms
- **PN_DICT** – proper names dictionary

ABD – acronym/abbreviation detection algorithm

- It is trained on SYL_DICT dictionary by calculating *n*-grams probabilities
- It relies on the assumption that abbreviations do not respect the linguistic syllable structure
- Relies on a modified *n*-grams based syllabification algorithm
- An abbreviation has the structure of an impossible syllable

ABD – acronym/abbreviation detection algorithm



Evaluation and results for ABD

- TS_ABD contains 400 tokens: 200 AC/AB, 200 common words

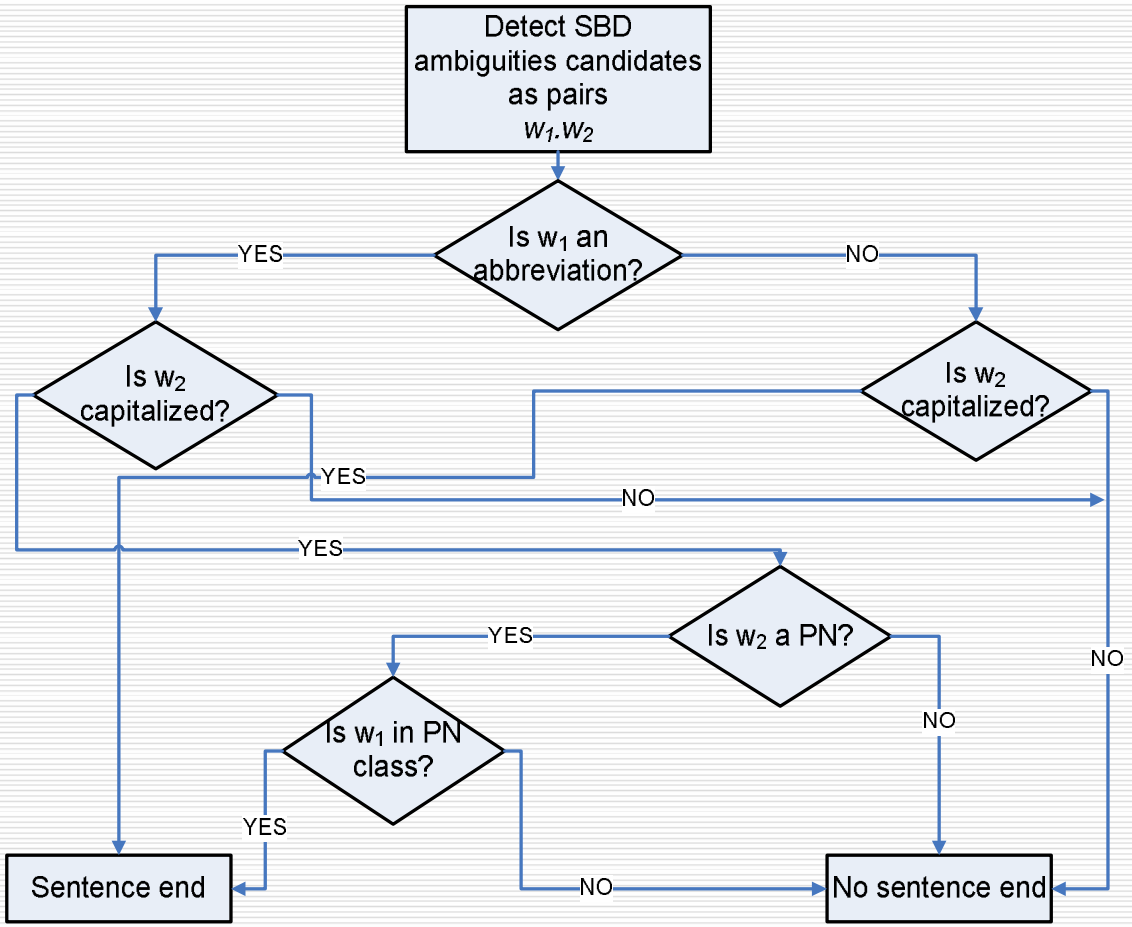
Precision (%)	Recall (%)	Accuracy (%)
96.9	78.5	88,7

- Enlarged Romanian AB_DICT: 2,480 tokens

SBD – sentence boundary detection

- SBD is difficult for ambiguities like:
  - AB. PN: *L-am sunat pe dr. Marcu să vină la mine.* (*I called dr. Marcu to come to me.)*
- Relies on ABD and PND
- For ABD, further AB classification into classes is made
  - Some AB require sentence end, other require continuation of the sentence, other are neutral before capital letters

SBD – implementation



Evaluation and results for SBD

Initial context

Test set	Sentences	AB+PN	AB+CN	2MP+PN	3MP+PN
TS1	149,050	520	3,092	176	35
TS2	125,905	387	2,595	127	33
TS3	152,646	483	3,199	208	48
TS4	190,756	563	4,185	228	48
TS5	180,804	563	3,924	158	22

Final results

Test set	AB+PN	AB+CN	2MP+PN	3MP+PN	SBD accuracy[%]
TS1	68	312	24	8	99.73
TS2	63	231	31	10	99.74
TS3	60	326	22	2	99.79
TS4	97	380	41	12	99.73
TS5	52	518	12	7	99.68

Conclusions

- The preprocessing/normalization stage in a TTS system has a fundamental contribution to the intelligibility and naturalness of the synthesized text
- Abbreviations/acronyms detection is used together with proper name detection for correct sentence boundary segmentation
- Enhanced resources were obtained in an offline manner: AB_DICT, PN_DICT $\Rightarrow$ useful for a high quality Romanian TTS system

Conclusions

- Acronym/abbreviation appearance is a productive process
- A statistical acronym/abbreviation detection method was obtained to solve ambiguous conditions
- Expanding acronyms/abbreviations requires a large AB_DICT obtained offline from a large corpus in a semi-supervised manner
- ABD is also responsible to provide a triggering signal to the speller of the speech engine

Conclusions

- ABD method lead to 88.7% accuracy
- Enlarged Romanian AB_DICT: 2,480 tokens
- Most ABD errors comes from words imported from other languages: lady, baby, kong, york
- PND method result is 99% accuracy
- The errors come from homographs: ion, popa, doru
- For SBD the final result is 99.74%
- SBD errors comes from both ABD or PND

Thank you !