Diacritization

Weirui Chen

What are diacritics

What are diacritics

P-side: Accent/tone markers to supplement the alphabet

S-side: Avoid ambiguity (when language is written down)

TABLE I. CHARACTER TRANSFORMATION

Characters without diacritics	Characters with diacritics
a	a, á, à, ả, ã, ạ, ă, ắ, ằ, ẳ, ẵ, ặ,â, ấ, ầ, ẩ, ẫ, ậ
d	d, đ
e	e, é, è, ẻ, ẽ, ẹ, ê, ế, ề, ể, ễ, ệ
i	i, í, ì, ỉ, ĩ, ị
o	o, ó, ò, ỏ, õ, ọ,ô, ố, ồ, ổ, ỗ, ộ, ơ, ớ, ờ, ở, ỡ, ợ
u	u, ú, ù, ủ, ũ, ụ, ư, ứ, ừ, ử, ữ, ự
y	y, ý, ỳ, ỷ, ỹ, ỵ

Nga et al. 2019

Diacritics in some European languages

TABLE I
DIACRITICS IN EUROPEAN LANGUAGES WITH LATIN BASED ALPHABETS

Language	Diacritics	Language	Diacritics
Albanian	ç ë	Italian	á è é ì í ï ò ó ù ú
Basque	ñ ü	Lower Sorbian	ċ č ě ł ń ŕ ś š ź ž
Breton	â ê ñ ú ö	Maltese	ċ ġ ż
Catalan	à ç è é í ï ò ó ú ü	Norwegian	å æ ø
Czech	á č é í ñ ó ř š ý ž	Polish	ą ę ć ł ń ó ’s ’z ż
Danish	å æ ø	Portuguese	â ă ç ê ó ô ŭ ü
Dutch	ë	Romanian	ă â î ş ţ
English	none	Sami	á ï ĉ d- ń ŋ š t- ž
Estonian	ä č ō ö ž	Serbo-Croatian	ć č d- š ž
Faroese	á æ d- ó øú ý	Slovak	á ä č d’ é Í ñ ó ô ŕ š t’ ú ý ž
Finnish	ä å ö š ž	Slovene	č š ž
French	á â æ ç é è ë ê î œ ù û ÿ	Spanish	á é í ó ú ü ñ
Gaelic	á é í ó ú	Swedish	ä å ö
German	ä ö ü ß	Turkish	ç ğ ö ş ü
Hungarian	á é í ó ö ő ú ü ű	Upper Sorbian	ć č ě ł ń ó ř š ž
Icelandic	á æ ð é í ó ö ú ý	Welsh	ă ĕ ĭ ŏ ŭ ŵ ŷ

Nutu et al. 2019

Why do we need diacritics

Disambiguation

Arabic Sentence	English Sentence	Voiced	Pronunciation	Translation
علم السعودية أخضر وأبيض اللون	The **flag** of Saudi Arabia is green and white	عَلَمُ	[ʕalamu]	flag
أحب الفلك علم	I love space **science**	عِلمَ	[ʕilma]	science
علم ناصر أحمد السباحة	Nasser **taught** Ahmad how to swim	عَلَّمَ	[ʕalːama]	taught

Diacritization

What is diacritization

Add diacritics to bare character if necessary

ni ki balfi ⟶ ní ki bālfî

Also named: Diacritics Restoration, Diacritics Generation, Diacritics Recovery, Accent Restoration (old-fashioned), Unicodification (old-fashioned), diacritics insertion (old-fashioned)

Why Diacritics are ignored

Historical reasons

1. Only English keyboards and ASCII encoding

2. Cross-OS or Cross-platform encoding-decoding issue

3. Even when Unicode is widely adopted, encoding issue can still arise
    - One system component is using different encoding scheme from the other components
        - utf-8, utf-16, utf-32
    - Different versions of unicode: 13.0 on March 2020, 14.0 on September 2021

Why Diacritics are ignored (Cont'd)

Convenience reasons

- People are lazy switching between different keyboard layout when Code-switching happens a lot
- People are lazy to type diacritics
  - In the interest of time and typing speed

Technical reasons

- Diacritics are sometimes ignored for capital letters

Why is diacritization important

ASR

Not specifying the diacritics may create ambiguity

Text-to-speech

People may be able to read through it because of contextual information but may find it weird when listening to a wrongly pronounced word

How to perform diacritization

3 Approaches

Rule-based (language-dependent)

- Corpus, dictionary: Scannell 2011, Náplava 2018 (as baseline)
- Requires linguistic expertise
- Not extendible to low-resource languages which has no dictionary and/or with small corpus

Machine learning (language-dependent)

- Extract linguistic features and formulate as a classification task: Mihalcea 2002, Shahrour 2015
- Lower accuracy but higher interpretability

Neural (language-dependent / language-independent)

- NN-based: Belinkov 2015, Náplava 2018, Fadel 2019
- Transformer: Mubarak 2019, Laki 2020, Ali 2021

Metrics

DER: The percentage of misclassified Arabic characters whether the character has 0, 1 or 2 diacritics (Fadel 2019)

WER: The percentage of Arabic words which have at least one misclassified Arabic character (Fadel 2019)

How difficult it is to perform diacritization

Lexical Diffusion: Average number of possible orthographic instantiations of the same Latin form. (De Pauw et al. 2007)

By comparing the statistics of linguistic characteristics and baseline performances. (Náplava 2018)

Language	Words with diacritics	Word error rate of dictionary baseline
Vietnamese	88.4%	40.53%
Romanian	31.0%	29.71%
Latvian	47.7%	8.45%
Czech	52.5%	4.09%
Slovak	41.4%	3.35%
Irish	29.5%	3.15%
French	16.7%	2.86%
Hungarian	50.7%	2.80%
Polish	36.9%	2.52%
Swedish	26.4%	1.88%
Portuguese	13.3%	1.83%
Galician	13.3%	1.62%
Estonian	19.7%	1.41%
Spanish	11.3%	1.28%
Norwegian-Nynorsk	12.8%	1.20%
Turkish	30.0%	1.16%
Catalan	11.1%	1.10%
Slovenian	14.0%	0.97%
Finnish	23.5%	0.89%
Norwegian-Bokmaal	11.7%	0.79%
Danish	10.2%	0.69%
German	8.3%	0.59%
Croatian	16.7%	0.34%

Rule-based approaches

Scannell 2011

For each ASCII word in the input text, this algorithm first finds all words in the first layer whose asciification equals the input word. If there is just one such word, this is taken as the output. If there is more than one, the most common one in the training data is taken.

Machine Learning approaches

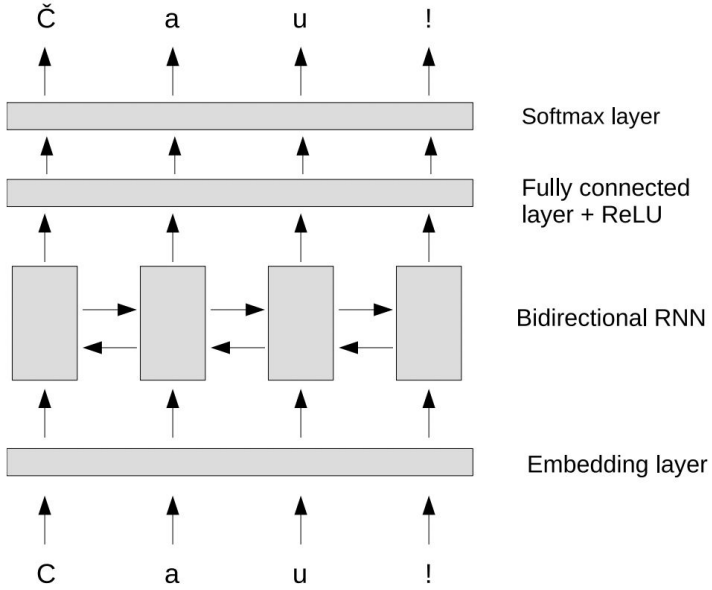
Mihalcea 2002

Feature extraction: Hence, we decided for very simple features, for the extraction of whom no particular processing is required. We are using surrounding letters, with a special notation assigned to white spaces, commas, dots and colons.

Model: decision tree

Neural Network approaches

Náplava 2018



Neural network approach variants

Multi-task learning: diacritization and translation (Ali 2021)

MT task may provide the model semantic and linguistic knowledge to resolve ambiguities in diacritization.

Arabic Sentence	English Sentence	Voiced	Pronunciation	Translation
علم السعودية أخضر وأبيض اللون	The **flag** of Saudi Arabia is green and white	عَلَمُ	[ʕalamu]	flag
أحب الفلك علم	I love space **science**	عِلمَ	[ʕilma]	science
علم ناصر أحمد السباحة	Nasser **taught** Ahmad how to swim	عَلَّمَ	[ʕalːama]	taught

Thank you!