A term identification system for a specific domain

{Corpus, rule-based identification system and tagger development}

Cécile MACAIRE, Ludivine ROBERT

M2 NLP
UE902 EC3 Terminology

November 25th, 2020

UNIVERSITÉ DE LORRAINE

Institut des sciences du Digital Management & Cognition

Table of Contents

Pipeline

Corpus

- NLP field: *Text-to-Speech* domain.
- Texts taken from scientific publications.
- **.pdf** converted into **.txt** format.

Corpus size:

Train	**Test**
22	2

Lexicon creation

- Automatic extraction with TermSuite [1].
- Parameters: pattern, lemma, frequency for each term.

#	type	pattern	pilot	lemma	freq
	T	N N	1D Convolution	1d convolution	
	T	N N	1D convolutional	1d convolutional	
9	T	A N	acoustic model	acoustic model	72
497	T	A N N	acoustic model training	acoustic model training	5
2346	T	A	acoustic-phonotactic	acoustic-phonotactic	1
478	T	N N	Adam optimization	adam optimization	5
225	T	N N	Adam optimizer	adam optimizer	9
676	T	A A N	adaptive cepstral analysis	adaptive cepstral analysis	4
279	T	N N	adversarial loss	adversarial loss	8
334	T	N N N	adversarial speaker classifier	adversarial speaker classifier	7
	T	A N N	artificial human speech	artificial human speech	
1345	T	N	ASR	asr	
438	T	N N	ASR performance	asr performance	5
82	T	N N	attention alignments	attention alignment	19
121	T	N N	attention block	attention block	14
	T	N N N	attention context vector	attention context vector	

Figure: Beginning of the lexicon file data.

Rule-based approach

- Based on dictionary lookup & heuristics.
- Main idea: identify the terms by looking into the list of terms and the rules.
- Important steps: lemmatization, define the rules.

Rule-based approach

Some rules:

Rules		Examples
<N-Prep-N> +	<N,N> <N> <Adj,N>	text-to-speech synthesis system end-to-end pipeline end-to-end neural speech
<T1 - T2> +	<N, N, N> <N, N> <N>	HMM-based speech synthesis system n-gram language modeling grapheme-based model
[data, voice, etc.] + <Adj | N> +	<N> [data, voice, etc.]	voice quality audio data
<Adj> +	T	autoregressive models
T +	<W>	phoneme representations

Rule-based approach

```
We conduct experiments on the [LJSpeech dataset] to test [FastSpeech].
The results show that in terms of [speech quality], [FastSpeech] nearly matches
the [autoregressive Transformer model].
Furthermore, [FastSpeech] achieves 270x speedup on [mel-spectrogram generation]
and 38x speedup on final [speech synthesis] compared with the [autoregressive
Transformer TTS model], almost eliminates the problem of word skipping and
repeating, and can adjust [voice speed] smoothly.
We attach some [audio files] generated by our method in the supplementary
materials.
```

Figure: Extract of an annotated file.

Sequence tagger

IOB tagger

```
We (O) conduct (O) experiments (O) on (O) the (O) [LJSpeech (B) dataset (I)]
to (O) test (O) [FastSpeech (B)].
The (O) results (O) show (O) that (O) in (O) terms (O) of (O) [speech (B)
quality (I)], [FastSpeech (B)] nearly (O) matches (O) the (O) [autoregressive
(B) Transformer (I) model (I)].
Furthermore, (O) [FastSpeech (B)] achieves (O) 270x (O) speedup (O) on (O)
[mel-spectrogram (B) generation (I)] and (O) 38x (O) speedup (O) on (O) final
(O) [speech (B) synthesis (I)] compared (O) with (O) the (O) [autoregressive
(B) Transformer (I) TTS (I) model (I)], almost (O) eliminates (O) the (O)
problem (O) of (O) word (O) skipping (O) and (O) repeating, (O) and (O) can
(O) adjust (O) voice (O) speed (O) smoothly. (O)
We (O) attach (O) some (O) [audio (B) files (I)] generated (O) by (O) our (O)
method (O) in (O) the (O) supplementary (O) materials. (O)
```

Figure: Extract of an annotated and IOB tagged file.

Sequence tagger

POS tagger with *spacy* library

```
 We PRON (O) conduct NOUN (O) experiments VERB (O) on ADP (O) the DET (O)
[LJSpeech NOUN (B) dataset NOUN (I)] to PART (O) test NOUN (O) [FastSpeech
PROPN (B)] .
The DET (O) results VERB (O) show NOUN (O) that SCONJ (O) in ADP (O) terms
NOUN (O) of ADP (O) [speech NOUN (B) quality NOUN (I)] , [FastSpeech PROPN
(B)] nearly ADV (O) matches NOUN (O) the DET (O) [autoregressive ADJ (B)
Transformer VERB (I) model NOUN (I)] .
Furthermore ADV , (O) [FastSpeech PROPN (B)] achieves VERB (O) 270x NUM (O)
speedup NOUN (O) on ADP (O) [mel PROPN - spectrogram PROPN (B) generation
NOUN (I)] and CCONJ (O) 38x NOUN (O) speedup NOUN (O) on ADP (O) final ADJ
(O) [speech NOUN (B) synthesis NOUN (I)] compared VERB (O) with ADP (O) the
DET (O) [autoregressive ADJ (B) Transformer VERB (I) TTS PROPN (I) model NOUN
(I)] , almost ADV (O) eliminates NOUN (O) the DET (O) problem NOUN (O) of ADP
(O) word NOUN (O) skipping NOUN (O) and CCONJ (O) repeating VERB , (O) and
CCONJ (O) can VERB (O) adjust NOUN (O) voice NOUN (O) speed NOUN (O) smoothly
ADV . (O)
We PRON (O) attach NOUN (O) some DET (O) [audio NOUN (B) files NOUN (I)]
generated VERB (O) by ADP (O) our DET (O) method NOUN (O) in ADP (O) the DET
(O) supplementary ADJ (O) materials NOUN . (O)
```

Figure: Extract of an annotated and IOB + POS tagged file.

Evaluation

Metrics and results

Common indexes have been used:

- Precision: % of selected items that are correct.
- Recall: % of correct items that are selected.

	Automatic annot		
		+	-
Manual annot	+	13+18	0+1
	-	4+2	-

Table: Confusion matrix of the evaluation results.

$$F_1 = 2 * \frac{(31/38) * (31/32)}{(31/38) + (31/32)} \simeq 0.88$$

meaning or resolve into the [target language]. To achieve such, there is a need to adopt coordinated cooperation of the separate [Human Language Technology] components. In particular, the most significant elements in a [speech translation system] encompass the [automatic speech recognition], [machine translation] as well as [text to speech]. With this understanding, the current paper explores

Figure: Expected output

meaning or resolve into the [target language]. To achieve such, there is a need to adopt coordinated cooperation of the separate Human Language Technology components. In particular, the most significant elements in a [speech translation system] encompass the [automatic speech recognition], [machine translation] as well as [text to speech]. With this understanding, the current paper explores

Figure: Tool's output

Figure: Sample of the evaluation outputs

Conclusion

Conclusion

- Good experience to understand and manipulate terminology.
- Few difficulties:
  - corpus creation,
  - exceptions which occur in this specific field (dash),
  - identification of the rules specific to our corpus,
  - manually checking.
- Improvements in our system — larger corpus, system based on more precise rules, etc.

Thank you!
Do you have any questions?

References I

Cram, Damien and Béatrice Daille (2016). "Terminology extraction with term variant detection". In: *Proceedings of ACL-2016 System Demonstrations*, pp. 13–18.