Cross-lingual topic identification in low resource scenarios

Sangeet Sagar, Santosh Kesiraju,
Ondřej Glembek, Lukáš Burget

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Cross-lingual topic ID ?

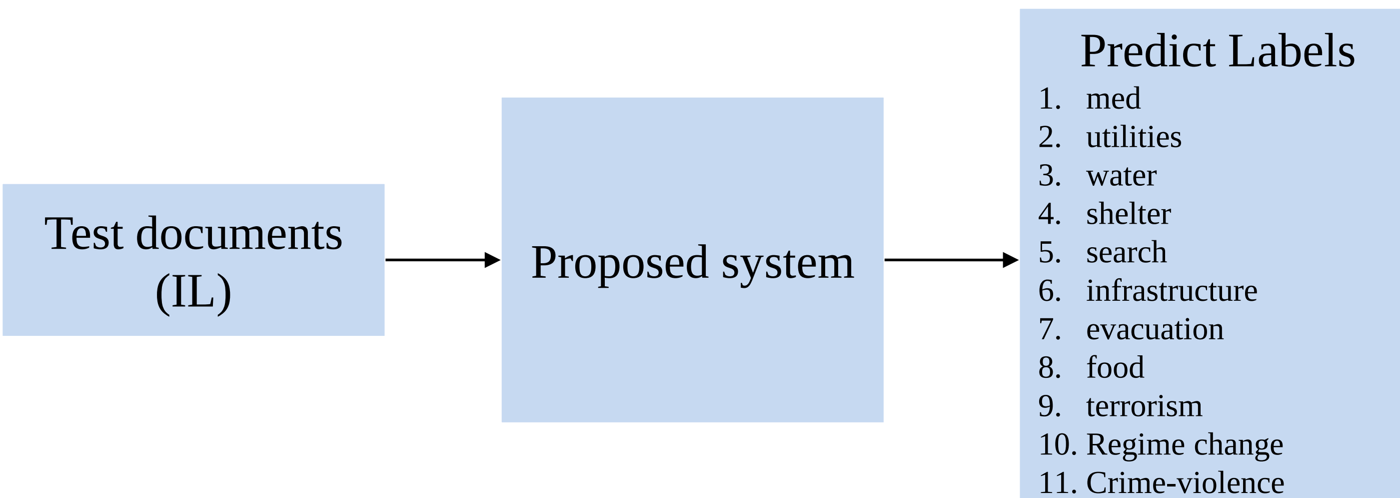
- IL – Incident language/ target language
- English – source language

Given

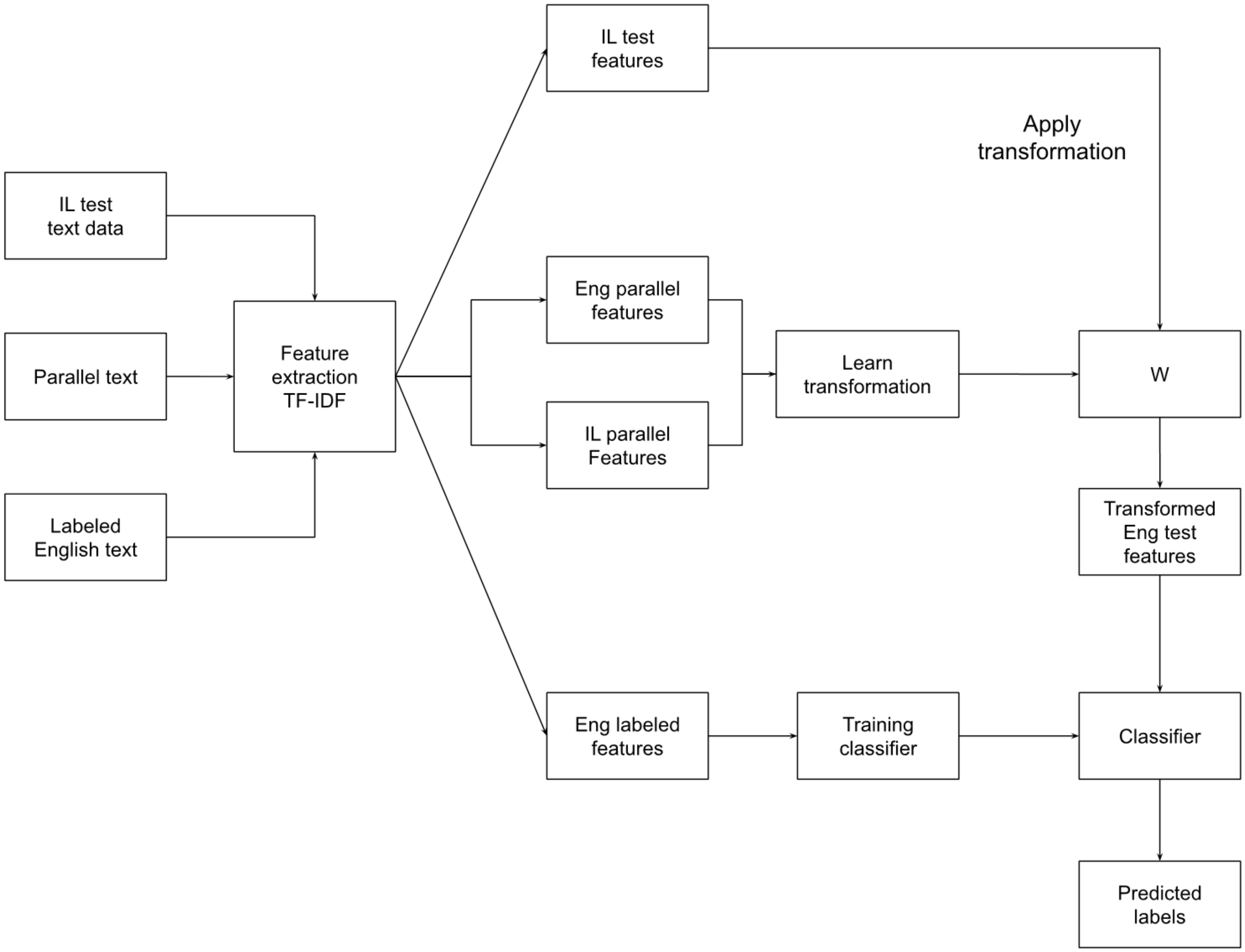
- parallel text (English - IL)
- labeled documents (English)

Task

- predict topic labels for test documents (IL)



Proposed System



- Parallel text

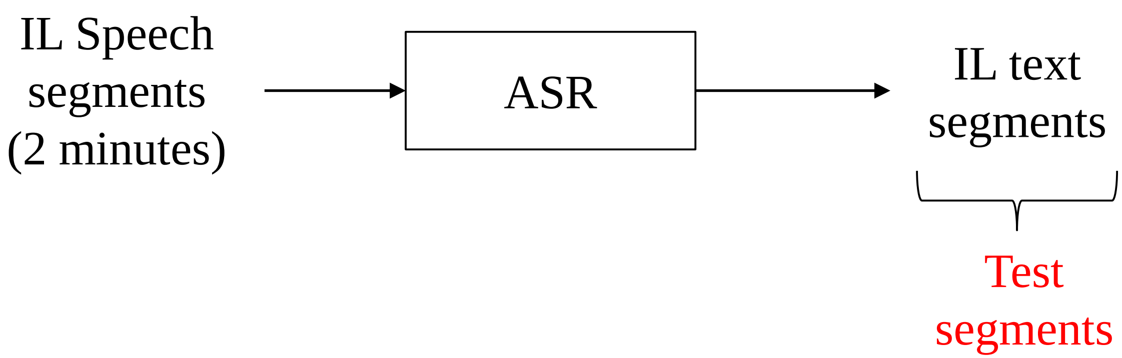
Table 1: Details of LORELEI parallel text data

Data	**Language**	**Number of parallel sentences**	**Writing system**
IL9	Kinyarwanda	29,3559	Latin
-	Zulu	27,4063	Latin
-	Hindi	11,563	Devanagari

- Training data (English) for topic ID
  - Dataset - LDC LORELEI
  - Comprising of 9,017 documents belonging to 11 classes

Dataset

Source of test documents



- Several speech segments makes one recording which represent one document.
- Objective is to predict topic(s) at segment level.
- ASR system description
  - GMM-HMM based – **ASR I**
  - DNN-HMM based – **ASR II**

Average Precision score

- A measure that combines recall and precision to interpret the performance of classifier
- Computed using

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

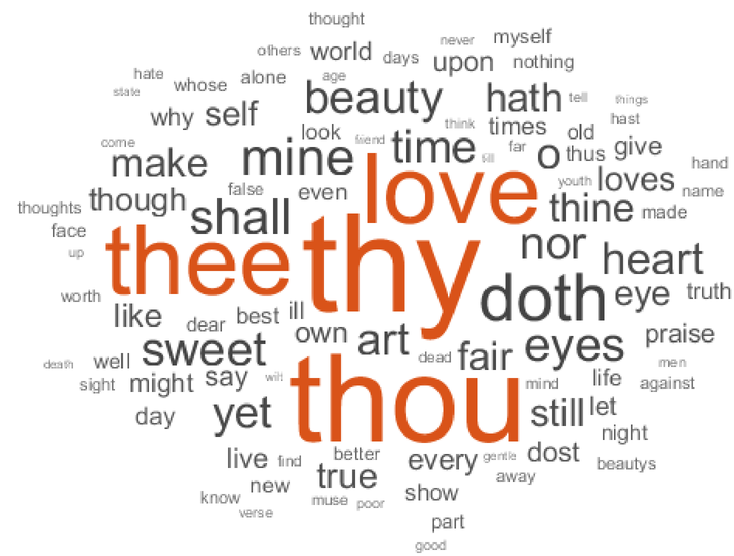
Where P_n and R_n are the precision and recall at the nth threshold.

- A higher AP is an evidence of a better classifier.
- *Weighted average precision* (WAP)

$$WAP = \frac{\sum_n AP_n * C_n}{\sum_n C_n}$$

Where AP_n is the average precision score of topic n and C_n is number of documents in that topic.

Bag-of-word (BoW) model



- Tokenization - character tri-grams
- Text documents → BoW → Counts → TF-IDF

Results: WAP using char-trigrams

Multi Label topic ID weighted average precision scores

Table 2: Multi Label topic ID weighted average precision on LORELEI language packs

Language	**full-set**	
	ASR I	ASR II
Kinyarwanda (IL9)	0.2299	0.1917
Zulu	0.2221	0.2165
Hindi	0.1075	0.1411

Learning transformation on topic related sub-set

- Not all sentences in parallel text are topic related.
- Select topic related text to examine if it helps to learn a better transformation
- Classify English parallel text – select ones that are 70% likely to belong to a topic
- Feature transformation using sub-set – subset of topic related sentences from parallel text

Comparison of weighted average precision (WAP) scores using transformation learned on full-set and sub-set

Table 3: Multi Label topic ID weighted average precision on LORELEI language packs

Language	**full-set**		**sub-set**	
	ASR I	ASR II	ASR I	ASR II
Kinyarwanda (IL9)	0.2299	0.1917	**0.2564**	0.2104
Zulu	0.2221	0.2165	**0.2438**	0.2302
Hindi	0.1075	**0.1411**	0.0971	0.0984

Strategy II - Merging segments

- Most segments of a test document have common labels†.
- Combine all such segments to form a large segment of text.
- Share the prediction scores of the large segments among its child segments.

Table 4: Weighted average precision scores upon combining all sentences of a document into a single sentence.

Language	**full-set**		**sub-set**	
	ASR I	ASR II	ASR I	ASR II
Kinyarwanda (IL9)	**0.4145**	0.3497	0.3630	0.2092
Zulu	0.2023	0.2727	0.2537	$\boxed{\textbf{0.2846}}$
Hindi	0.1410	$\boxed{\textbf{0.1905}}$	0.1316	0.1249

† C. Liu *et al*., "Low-Resource Contextual Topic Identification on Speech," *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018, pp. 656-663

Strategy III - Vocabulary selection

- Pick out topic-specific-tokens from English labeled LDC data and explicitly search for them in the parallel text.
- Such tokens are selected from each topics using:

$$S(w,t) = \frac{\sum_{d \in D_t} f_{wd}}{\sum_{\forall d} f_{wd}}$$

Where,

$S(w,t)$ is score of token w in topic t.

f_{wd} represents frequency of token w in document d.

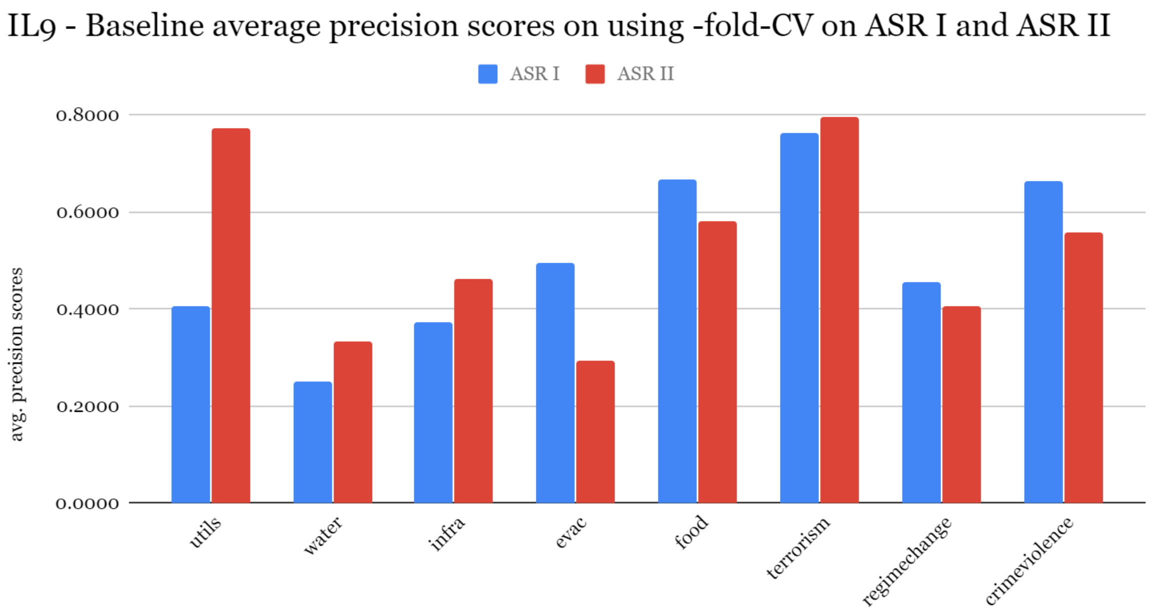
D_t means all documents belonging to topic t.

Table 5: Weighted average precision upon selection of topic specific tokens

Language	**full vocabulary**		**selected vocabulary**	
	ASR I	ASR II	ASR I	ASR II
Kinyarwanda (IL9)	0.4145	0.3497	**0.4524**	0.3767
Zulu	0.2023	**0.2727**	0.2033	0.2416
Hindi	0.1410	**0.1905**	0.1369	0.1629

Strategy III - Merging ASR outputs

- Test documents from ASR I has better performance for some labels while for other labels ASR II shows better results.
- Procedure
  - Examine WAP upon combination of ASR I and ASR II test documents.



Strategy III - Merging ASR outputs

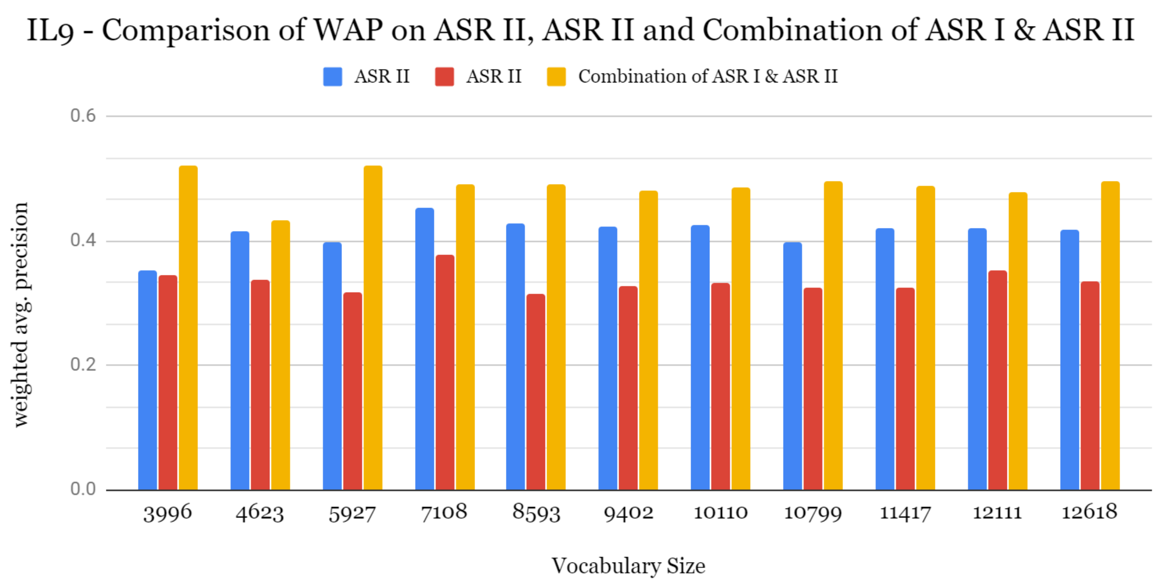


Table 6: Weighted average precision on combination of test documents

Language	ASR I	ASR II	**Combination of ASR I & ASR II**
Kinyarwanda (IL9)	0.4524	0.3767	$\boxed{\mathbf{0.5212}}$
Zulu	0.2033	**0.2416**	0.2206
Hindi	0.1369	**0.1629**	0.1254

Conclusion and future work

- Kinyarwanda (IL9) showed best results when test documents from ASR I & ASR II are merged.
- However, these strategies did not seem to show significant improve in results for Zulu and Hindi.
- Tremendous amount of work still needs to done for these languages.
- For Hindi, we should probably try syllable tri-grams