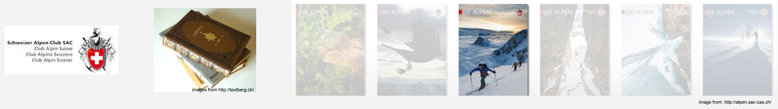
Leveraging Data-Driven Methods in Word-Level Language Identification for a Multilingual Alpine Heritage Corpus

Ada Wan
ada.wan@uzh.ch
adawan919@gmail.com

Task:

Word-level language identification (including Named Entity Recognition (NER)) of a multilingual corpus compiled with 150 years of publication from the *Swiss Alpine Club* (SAC)



images from http://textberg.ch/

image from: http://alpen.sac-cas.ch/

Motivation

Language Identification (LID) is important so long as the language processing applications and tools designed and used are language-specific. Many tokenizers, POS-taggers, lemmatizers, and NER systems suffer in performance when met with sporadic sequences of unknown elements.

While LID can be viewed under certain circumstances, e.g. on document-level, as a "solved task", LID on smaller spans of texts continues to pose challenges. That is especially the case when the texts contain instances of code-mixing/code-switching, i.e. when the writers mix in other language(s) or switch back and forth between at least two languages in communication. Code-mixed/code-switched segments can occur within a document, a sentence, or even within a word.

Real-World Data

The *Swiss Alpine Club* (SAC) has been publishing periodically (formerly yearly, now monthly) since 1864. Published texts, esp. in earlier issues, are inherently code-switched. That is, an issue can, for example, contain some articles in DE, some in FR, some in IT, or an author would mix in passages, sentences, or words in another language without translation.

Language format used in these periodicals (referred to as "yearbooks"):
1864 to 1956: mixed languages
1957 to 2011: parallel versions in DE & FR
2012 to present: parallel versions in DE, FR, & IT
Note that the phenomenon of code switching still persists, despite the presence of an "official" dominant language in the parallel versions.

Text+Berg digital has been converting the yearbooks into annotated XML files from these original file formats:
1864-2000: scanned & converted into text with commercial OCR software
2001-2009: texts extracted from PDFs
2010-present: XML files received directly from SAC

Approx. 40 million tokens of domain-specific, genre-diverse, multilingual, diachronic (about 150 years) data, containing:

- reports and essays on all aspects of alpinism, alpine nature and culture
- literary essays and anecdotal narratives on mountain expeditions
- book reviews
- poems
- practical travel tips such as hotel reviews and cabin directories
- scientific studies of living organisms, glacier and climate observations
- geo-historical descriptions on cols, mountains, parks, and on the flora and fauna of the Alps and other mountain regions
- technical accident and security reports
- financial reports from protocols of the annual club gatherings

14	uhr	25266	4	camp	7118	5	cima	3157	45	alps	645	41	sv	197
14	nacht	7296	4	mal	6198	5	passo	2373	45	mountains	597	41	lub	180
14	stunde	7217	4	hôtel	2769	5	vi	1906	45	lake	593	41	zei	102
14	sonne	7183	4	halte	1904	5	lago	1848	45	range	445	41	isa	97
14	morgen	6240	4	étions	1704	5	punta	1780	45	himalayan	399	41	lps	90
14	himmel	5676	4	chaud	1594	5	circa	1721	45	black	389	41	ais	89
14	nebel	5491	4	chance	1559	5	dalle	1668	45	karakoram	388	41	pü	84
14	lager	3598	4	tentative	1527	5	campo	1547	45	for	364	41	lur	75
14	verlassen	3208	4	trouvé	1515	5	rifugio	1491	45	mountaineering	360	41	dais	75
14	licht	2962	4	arrivée	1509	5	io	1424	45	moore	351	41	sün	74
14	wolken	2623	4	après-midi	1481	5	alpino	1070	45	no.	339	41	aint	73
14	morgens	2586	4	arriver	1412	5	gran	1003	45	sikkim	326	41	scu	70
14	rast	2577	4	perdu	1394	5	termine	905	45	garhwal	316	41	svizzer	63
14	stein	2512	4	compagnie	1258	5	ti	812	45	valley	308	41	izz	61
14	hotel	2454	4	attendre	1235	5	pur	789	45	ice	304	41	naiv	59
14	früh	2416	4	gravi	1219	5	pure	778	45	kenya	291	41	tuot	52
14	abends	2380	4	tempête	1210	5	no	736	45	snow	290	41	suot	52
14	kopf	2350	4	étape	1172	5	bel	686	45	douglas	289	41	eira	49
14	rucksack	2172	4	réussi	1149	5	né	664	45	ladakh	280	41	daint	47
14	halb	2091	4	souvenir	1130	5	porta	661	45	hudson	280	41	sco	46

Fig. 1: Examples of clusters labeled (with cluster number, word token, & term frequency in each of the 5 sets of 3 columns):
German (DE) French (FR) Italian (IT) English (EN) Romansch (RM)

Method

1. **unsupervised word vectors with optimized count-based methods:**
   weighted co-occurrence count of all word types and words in context transformed into a normalized association measure PPMI (positive pointwise mutual information)
2. **TSVD (truncated singular value decomposition) for dimensionality reduction**
3. **simple cluster-and-label approach:**
   K-means clustering, then assign one language label per word type
   (see Fig. 1 above)
   Possible cluster labels were DE, EN, FR, IT, RM, CH-DE (Swiss German), NE (named entity), and MIXED (i.e. none of the aforementioned classes).

Note: clustering experiments with non-parametric models (e.g. DPGMM (Dirichlet Process Gaussian Mixture Model)) resulted in higher number of clusters -- though more clusters did not necessarily entail purer clusters.
50 clusters were sufficient to achieve a satisfactory score when compared with the baseline and were easy enough to manually label.

Evaluation

5,073 words from 192 sentences considered to be likely to contain an intra-sentential code-switching instance according to the code-switching algorithm from Volk & Clematide (2014) (VCCS) were randomly selected for word-level LID evaluation.

Comparison with VCCS using off-the-shelf alternatives:

The system against which we evaluated the results of our LID system for Text+Berg SAC yearbook corpus (TBLID) uses *Lingua-Ident* by Michael Piotrowski, a statistical language identifier based on character n-gram frequencies for all sentences with more than 40 characters* (on the sentence-level) and *langid.py* (a Naive Bayes classifier based on byte n-grams) for the intra-sentential code-switched segments.
*Results for shorter sentences and for RM were found to be unreliable and these were hence assigned language tag of the previous sentence or that of the article.

DE, EN, FR, IT, RM, CH-DE, NE, MIXED

Sample sentence from yearbook 1925:

Gold standard:
Je n' en sais rien , mais l' énergie de son « Oel per oel e daint per daint » résonne encore à mon oreille . (22 words)

VCCS + *Lingua-Ident* + *langid.py* (incorrectly tagged as OC (Occitan)*)**:**
Je n' en sais rien , mais l' énergie de son « Oel per oel e daint per daint » résonne encore à mon oreille . (15/22)
*or IT if, for *langid.py*, the set of languages is restricted to the one assumed to be most relevant for the yearbooks: DE, EN, FR, IT, ES (Spanish), & LA (Latin)

TBLID:
Je n' en sais rien , mais l' énergie de son « Oel per oel e daint per daint » résonne encore à mon oreille . (17/22)

Named Entities (NEs):
Each word in a name for a person/location/organization is identified as NE.
Our general intuition behind determining the gold standard for NEs is: if an NE is to be found in a monolingual gazetteer, in what language would the gazetteer be?

TBLID (49/53, errors underlined), from DE yearbook 1975:
Indessen gibt dann der Guide bleu von 1962 einige Details über die Bedingungen für eine Besteigung bekannt , indem er von der alten Ausgabe die Erwähnung warmer Kleidung , guter Schuhe und Schutzbrillen übernimmt , aber alles in Verbindung mit « Grand Hotel Ätna , mehreren Restaurants bei der Casa Cantoniera , Schutzhütte Sapienza , Hütte Menza des CAI , Wintersport etc . »

TBLID (44/44), from yearbook 1900:
Bien que „ Über Eis und Schnee " n' indique , comme unique ascension du Sattelhorn que celle précitée du Prof. Schulz , ce sommet a cependant été atteint à deux reprises , il y a quelques années , de la cabane Oberaletsch (MM. Courvoisier et Girard , Courvoisier et Grisel) .

Results

Final accuracy scores (rounded to 2 digits), based on 5073 words:

	Lingua-Ident + langid.py	TBLID
strict	**89.73**	89.33
lenient	89.81	**89.91**

Conclusion

We have presented a simple data-driven approach that identifies the language of word types of a multilingual, diachronic, domain-specific, genre-diverse corpus of almost 40 million words with accuracies that are comparable/superior to the baseline that does not require any human supervision save for the minimal effort in labeling 50 clusters.

Selected references:

- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225.
- Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30. The Association for Computer Linguistics.
- Marco Lui. 2014. *Generalized Language Identification*. PhD thesis, The University of Melbourne.
- Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in codeswitched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October. Association for Computational Linguistics.
- Martin Volk and Simon Clematide. 2014. Detecting code-switching in a multilingual alpine heritage corpus. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 24–33, Doha, Qatar, October. Association for Computational Linguistics.

image (bottom) from http://blogs.ethz.ch/digital-collections/files/2011/06/Panorama_West1.jpg