

Text Based User Comments as a Signal for Automatic Language Identification of Online Videos

A. Seza Dođruöz
Xoogler
Turkey
a.s.dogruoz@gmail.com

Natalia Ponomareva
Google Inc.
USA
nponomareva@google.com

Sertan Girgin
Google Inc.
France

Reshu Jain
Google Inc.
USA

Christoph Oehler
Google Inc.
Switzerland

ABSTRACT

Identifying the audio language of online videos is crucial for industrial multi-media applications. Automatic speech recognition systems can potentially detect the language of the audio. However, such systems are not available for all languages. Moreover, background noise, music and multi-party conversations make audio language identification hard. Instead, we utilize text based user comments as a new signal to identify audio language of YouTube videos. First, we detect the language of the text based comments. Augmenting this information with video meta-data features, we predict the language of the videos with an accuracy of 97% on a set of publicly available videos. The subject matter discussed in this research is patent pending.

CCS CONCEPTS

• **Information systems** → *Multilingual and cross-lingual retrieval*;

KEYWORDS

Automatic language identification, machine learning, signal processing, natural language processing, multi-media content, YouTube

ACM Reference Format:

A. Seza Dođruöz, Natalia Ponomareva, Sertan Girgin, Reshu Jain, and Christoph Oehler. 2017. Text Based User Comments as a Signal for Automatic Language Identification of Online Videos. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3136755.3136788>

1 INTRODUCTION

There are more than 1 billion YouTube users around the world and many of them upload videos in diverse languages. 80% of these videos are uploaded outside the USA [8] and the YouTube website is available in 76 languages. Identifying the audio language of YouTube videos is essential to rank videos, provide subtitles for users who do not understand the original language of the video

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ICMI'17, November 13–17, 2017, Glasgow, UK

© 2017 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5543-8/17/11.

<https://doi.org/10.1145/3136755.3136788>

or have hearing impairment, and for providing relevant ads and recommendations.

Automatic *Language Identification* (LID) is part of Automatic Speech Recognition (ASR) systems which are deployed in multiple languages. However, LID based on audio signals is a difficult task for videos involving multiple speakers, interruptions, background music and noise. In addition, ASR systems are not widely available for all languages. YouTube users may assign a language themselves while uploading the video. However, this feature became available only recently and a large portion of YouTube videos do not have an uploader-assigned language yet. In addition, when an uploader assigns a language to a video, it does not guarantee the correctness of such an assignment.

Instead of relying on audio signals, we use text based comments posted by users to predict the audio language of the videos. This new signal is simple to implement and provides a quick and accurate coverage for videos in at least 21 language (97% accuracy, with precisions, recalls and F1s in the high 90s). In addition, this method can also be used for the majority of videos that were not assigned a language during the upload.

2 RELATED RESEARCH

When a top-level ASR system encounters requests in more than one language, LID can either be fed to the system externally by the user or it can be done automatically. Niesler et al. apply LID to South-African languages by combining acoustic and language modeling using HMMs trained over audio (speech) data [14] and achieve an accuracy of about 81%. Most recently, Gonzalez-Dominguez et al. developed a multilingual ASR architecture to recognize speech in several languages that works simultaneously with LID [7]. The system relies on pre-selected languages indicated by the user. Based on user input, LID recognizes the spoken language and an ASR system for each language is deployed to decode the input signal. On average, the LID system achieves an accuracy of 80% across languages on Google 5M LID corpus.

High accuracy in LID for spoken data is still very difficult with naturalistic multi-party human communication. Instead of speech based signals [10],[15], we propose text based comments as a new signal for audio LID of the videos. LID for text data has a wide array of applications ranging across Machine Translation for online resources [11] and building linguistic resources from the web [1]. To the best of our knowledge, we are the first to utilize text based user comments as a new signal for audio LID.

3 USING TEXT AS A SIGNAL FOR VIDEO LANGUAGE IDENTIFICATION

3.1 Data

Our data comprises publicly available YouTube videos with languages assigned by the uploaders, and publicly available user comments. If the users select incorrect languages (by mistake or intentionally), these labels may be noisy, but a robust classifier can tolerate such noise. First, we train a machine learning model on such noisy labels (Sections 4.1 and 4.2) and in section 4.3, we investigate how predictions of the classifier trained on noisy labels correspond to the ground truth labels (actual audio language).

During the clean up, we merged the low occurrence languages (e.g. Afrikaans, Albanian etc.) into "Other" class (together they represent less than 0.25% of the training data). Secondly, we did not distinguish between different dialects of the same language (e.g. French and Canadian French). Thirdly, we combined the most prevalent Indian languages (e.g. Hindi, Bengali) into an "Indian" class and Chinese languages into an aggregated "Chinese" class.

We also excluded the comments marked as abusive or moderated, filtered out all comments less than 5 words or 20 characters in length (e.g. "Ok"), auto-generated ("Shared on Google+"), spammy (URLs) and long comments (more than 10,000 characters) that are usually machine generated.

Our final training data contains approximately 10M videos and has 41 language classes. The top 6 languages (English, Spanish, Portuguese, Russian, German and French) represent approximately 80% of all the training data, with mean number of comments of 41.

3.2 Feature Engineering

From the videos and associated comments, we extract two types of features: language- and video- related ones.

Language Related Features. We classify each valid comment using an internal language-detecting classifier (ILDC) which returns a probability score for each language (100+ languages) [9]. ILDC is a pre-trained model (i.e. not trained on our data). We consider the following language-related features:

- *Base features:* Average LID scores for comments - a vector of sparse continuous features representing the weighted average scores across all videos' comments. The weight of the scores for each comment is proportional to the comment's length.
- Language with the maximum score - a nominal feature representing which language receives maximum score from *Base features*.
- Language score vectors for both the title of the video and the description of the video obtained from ILDC.

Video Related Features are features based on video metadata:

- The category assigned to a video by the uploader (e.g. NEWS, SPORTS etc).
- Number of comments posted under each video: More comments make us more confident for ILDC predictions.
- Number of views for the video: Videos with high viewing frequency usually have the same language for the audio, title and the description.

- Mean and standard deviation (SD) of the number of up-votes for the comments of the video. More up-votes signal authenticity (i.e. not machine generated) and trust in comment-based language features. Larger SD of up-votes might indicate less accuracy on comment-based LID (e.g. noisy comments that get no up-votes). Small SD and high mean are good signals for us to be confident about ILDC based LID.
- Mean and SD of the comments' length. We expect ILDC to be more reliable with longer comments, since it is more difficult to assign a language to short texts.

Before training models, we normalize our features and translate nominal features into binary. Our final training dataset had approximately 10M instances and 600+ features.

4 EXPERIMENTS

4.1 Baseline

To prove the feasibility of text-based LID, we created a simple baseline classifier that merely assigns the language receiving the maximum score based on *Base features* from ILDC. Such a baseline model does not require training because ILDC is pre-trained. Since there was no need for cross-validation, we used ILDC for all our training videos.

This baseline model achieves an accuracy of 91.43% which is much better than just predicting the majority class "English" for all videos (i.e. 40% accuracy). Table 2 contains baseline model precision, recall and F1 metrics for all the languages. In general, the first 20 languages exhibit very high precision and recall (high 80s and 90s), with the exception of Indonesian (recall of 76%), Japanese (recall of 64%) and Chinese (recall of 70%).

4.2 ML Cross Validated Models

We performed additional experiments to test the value of new features and whether a machine learning model can beat the performance of the baseline classifier. To compare the performance of different models, we used 5x2 CV with a subsequent t-test [5] for each pair of models (with Bonferroni correction).

We consider the following models: *Multiclass Perceptron* [2], *Winnow* [12], *Soft Margin SVM* [4] (Liblinear [6] with linear kernel and L2 loss and regularization), *Maximum Entropy Classifier* [13] and *Random Forest* [3] (10 and 100 trees). We use internal implementations (unless otherwise stated) of models with reasonable values of hyperparameters.

Table 1 provides the comparison of 5x2 CV results for different models. All pairwise t-tests show statistical significance, all classifiers improve over the baseline (Table 1). Among linear models, the best one is the Perceptron (4% accuracy improvement and improvement of precisions, recalls and F1s). For example, Perceptron improved the recall of Spanish by 8% and recall of Portuguese videos by 6%.

Random forests achieve a slightly better performance than perceptron. A random forest of 10 trees improves the accuracy by 0.44% (statistically significant) and marginally (less than a percent, significant) improves precision, recall and F1 measures on most prevalent languages. A forest of 100 trees provides statistically significant improvements over the forest of 10 trees, albeit the gain is

minimal (i.e. 0.22% for accuracy and less for precision, recall and F1 measures).

Perceptron and Forest provide the best generalization performance on our training data set. Since we obtain such a good performance with these simple models, we do not explore more powerful models (e.g. deep nets) for now.

All classifiers have very stable performances (SE are less than $1e - 4$). Therefore, we possibly do not need such a large training data set and can speed up the training time by subsampling the data without degrading the performance.

Using random forests, we performed feature importance analysis. Apart from *base features* and features that represent the language with the maximum score, the most important features were (in this order): standard deviation for the number of upvotes, title language features, categories of the video, average length of comments, average number of upvotes, number of counts and view count.

Next, we investigate the performance based on the number of training videos for the language. Table 2 summarizes CV performance on all the groups defined below.

Most prevalent languages (languages with at least 1% training videos): In combination, they represent 92.2% of the training data set. Languages from the most prevalent group exhibit high precision and recall (90% and higher). Japanese has the lowest recall of 88%, which is still a major improvement over the baseline recall of 64%.

Less prevalent languages (languages with inclusion between 1% and 0.1%): They cover another 6.4% of the training data. All languages apart from Indian, Serbian, Croatian and the accumulative group "Other" exhibit precisions and recalls around 80% and above. Note that the performance of "underperformer" Indian class is much better than that of a baseline classifier (64% vs 55% precision, 41% recall vs 8% recall). The same holds for "Other" class (50% vs 25% F1). The improvement for Indian languages may be due to the additional features (e.g. "language of the title" and "description"). The Indian class contains videos with comments in various languages spoken in India, in English and mixed languages (e.g Hindi-English), which creates noise. To improve the performance for this class, a separate scheme could be devised later.

Norwegian and Bulgarian (0.2% of the training data set) exhibit a performance on par with the prevalent group. These videos are possibly watched and commented on mostly by native speakers of these languages which may explain their strong performance.

The tail (all other languages): The tail of the distribution includes two languages that are very easy to detect from the comments: Lithuanian and Latvian. They exhibit precision and recall in high 80%. The rest of the languages have very low recall as typically observed in unbalanced data sets.

4.3 Golden Dataset

According to the CV experiments, we can accurately predict the user-assigned labels. However, since such labels may not represent the actual audio language of the video, we verified our approach by constructing a golden label set with the correct languages assigned by human raters. We chose the top 21 languages from our training data set excluding the aggregated Indian and Chinese classes. For each language, we randomly sampled 50 videos and sent them to 3 human raters. After the labeling, we reconstructed our test set by

assigning the language to the video (with majority agreement of the human raters). We reconstructed our full training set by excluding the videos that were in the golden set. Then, we trained the same models as in our CV experiments on this training set.

The overall accuracy on golden set is 97% for Forest 10 and 96.8% for a perceptron. Table 1 summarizes the resulting accuracy of different classifiers that were trained on the training data set with noisy labels and evaluated on the golden data set. Similar to the CV setting, the use of machine learning improves the performance over the baseline. This improvement is less pronounced since the golden data set had only top 21 languages. In addition, our CV experiments show that additional features and models provide more pronounced improvements for less prevalent and tail languages.

Table 1: Accuracy of models on golden dataset and CV.

Model	Golden set Accuracy	5x2 CV Accuracy
Baseline	0.9494	0.9143 \pm 0.0
Winnow	0.9680	0.9545 \pm 2.7e - 05
Perceptron	0.9680	0.9562 \pm 1.5e - 06
SVM	0.9638	0.9435 \pm 2.4e - 05
MaxEnt	0.9587	0.9440 \pm 3.2e - 06
Forest10	0.9700	0.9606 \pm 4.8e - 05
Forest100	0.9741	0.9628 \pm 6.5e - 06

Table 3 summarizes the perceptron results for different languages on the golden test data set. Based on human labeling of the videos for the golden data set, we obtained a rough estimate for the level of noise in the language labels. Noise represents situations when the uploader assigns a language to a video without spoken content (e.g. silent videos) and situations when the uploader assigns a language that is different than the language spoken in the video. English labels are very noisy (approximately 16% of the videos with English labels were either in a different language or had no speech content). Other prevalent languages also exhibit various levels of noise ranging between 0% and 14%.

All languages except English exhibit very high precision and recall (90% and higher). Users may utilize English to write comments for the videos in another languages. Therefore, videos with majority of English-looking comments get classified as English audio although some of them may have audios in other languages. As a result, the precision for English decreases. It is also possible that English videos are watched by speakers of other languages and commented on in different languages. This may also decrease the recall for English.

Notice that CV results indicate higher precision and recall for English (87% and 95%). Considering the level of noise in English labels (16%), the discrepancy is not surprising. Lower precision means predicting non-English audio videos as English. This is possible if our training data had wrongly assigned English labels. For example, some users label their videos as English even though they are not in English (to increase the international watch time of the videos) and our classifier learned this wrong labeling. Analyzing the confusion matrix, we notice that some videos are predicted as English although they are in Spanish (1), Italian (1), Indonesian (1), Romanian (1), Finnish (1), Japanese (3) and Korean (3).

Table 2: Baseline and perceptron 5x2 CV performance (precision, recall and F1). CV standard errors are negligible.

	Language	Baseline			Perceptron			%
		Pr	R	F1	Pr	R	F1	
Prevalent	English	0.872	0.954	0.911	0.947	0.959	0.953	39.62
	Spanish	0.967	0.901	0.933	0.969	0.981	0.975	10.79
	Portuguese	0.980	0.925	0.952	0.981	0.989	0.985	9.36
	Russian	0.952	0.948	0.950	0.969	0.976	0.973	7.49
	German	0.982	0.933	0.957	0.982	0.973	0.978	7.15
	French	0.973	0.891	0.931	0.972	0.958	0.965	4.91
	Italian	0.978	0.909	0.942	0.980	0.969	0.975	2.43
	Arabic	0.901	0.949	0.924	0.913	0.945	0.929	2.23
	Polish	0.962	0.943	0.953	0.969	0.977	0.973	1.83
	Turkish	0.959	0.920	0.939	0.956	0.970	0.963	1.79
	Korean	0.937	0.886	0.911	0.942	0.902	0.921	1.78
	Japanese	0.987	0.641	0.777	0.972	0.881	0.924	1.65
	Dutch	0.958	0.844	0.898	0.969	0.930	0.949	1.17
	Less prevalent	Thai	0.953	0.926	0.939	0.959	0.952	0.956
Vietnamese		0.864	0.938	0.900	0.884	0.945	0.913	0.97
Chinese		0.956	0.699	0.808	0.953	0.896	0.924	0.85
Indonesian		0.864	0.764	0.811	0.883	0.900	0.892	0.61
Czech		0.924	0.870	0.896	0.948	0.947	0.948	0.52
Romanian		0.932	0.832	0.879	0.943	0.928	0.935	0.48
Hungarian		0.967	0.943	0.955	0.974	0.960	0.967	0.47
Indian		0.550	0.082	0.143	0.639	0.407	0.498	0.45
Swedish		0.940	0.839	0.887	0.958	0.871	0.913	0.30
Finnish		0.966	0.905	0.934	0.972	0.942	0.956	0.24
Other		0.162	0.536	0.249	0.670	0.400	0.501	0.23
Greek		0.912	0.742	0.818	0.933	0.861	0.896	0.23
Ukrainian		0.807	0.226	0.353	0.939	0.732	0.823	0.22
Danish		0.957	0.847	0.899	0.972	0.916	0.943	0.22
Serbian	0.358	0.025	0.046	0.720	0.727	0.723	0.20	
Hebrew	0.908	0.925	0.916	0.937	0.915	0.926	0.18	
Norwegian	0.944	0.787	0.859	0.967	0.859	0.910	0.17	
Bulgarian	0.895	0.697	0.784	0.912	0.886	0.899	0.13	
Slovak	0.783	0.679	0.727	0.912	0.851	0.881	0.10	
Croatian	0.503	0.044	0.080	0.643	0.474	0.546	0.10	
Tail	Lithuanian	0.880	0.811	0.844	0.930	0.873	0.901	0.05
	Azerbaijani	0.476	0.718	0.572	0.578	0.568	0.573	0.04
	Malay	0.358	0.591	0.446	0.730	0.538	0.619	0.03
	Latvian	0.834	0.768	0.800	0.898	0.822	0.858	0.02
	Scots& Gael.	1.000	0.000	0.000	0.900	0.000	0.000	0.00
	Klingon	1.000	0.000	0.000	1.000	0.000	0.000	0.00
	Igbo	0.000	0.000	0.000	1.000	0.100	0.100	0.00
	Cherokee	1.000	0.000	0.000	0.900	0.000	0.000	0.00
	AVERAGE	0.832	0.679	0.696	0.907	0.773	0.792	

Even though Vietnamese, Indonesian and Turkish have high level of noise in the labels (14% and 10%), we predicted them with almost perfect precision and recall. These videos are probably watched and commented on mostly by native speakers of these languages. Therefore, it was easier to identify the languages of these videos despite the noisy labels.

Table 3: Perceptron and Forest 10 performance on golden set

Language	#	Noise	Perceptron		Forest 10	
			Precision	Recall	Precision	Recall
English	61	0.16	0.7500	0.8361	0.7778	0.8033
Spanish	46	0.02	0.9783	0.9783	0.9787	1.0000
Portuguese	48	0.02	0.9796	1.0000	0.9796	1.0000
Russian	47	0.06	0.9792	1.0000	0.9787	0.9787
German	45	0.04	1.0000	1.0000	1.0000	1.0000
French	47	0.02	0.9783	0.9574	0.9783	0.9574
Italian	48	0.02	0.9792	0.9792	0.9792	0.9792
Arabic	42	0.08	0.9545	1.0000	0.9333	1.0000
Polish	45	0.00	1.0000	1.0000	1.0000	0.9778
Turkish	42	0.10	1.0000	0.9762	1.0000	0.9762
Korean	45	0.06	0.9762	0.9111	0.9773	0.9556
Japanese	48	0.02	1.0000	0.9375	1.0000	0.9167
Dutch	44	0.04	1.0000	0.9318	1.0000	0.9545
Thai	47	0.02	1.0000	1.0000	1.0000	1.0000
Vietnamese	41	0.14	0.9762	1.0000	0.9762	1.0000
Indonesian	41	0.14	1.0000	0.9756	0.9762	1.0000
Czech	46	0.02	1.0000	0.9783	1.0000	1.0000
Romanian	46	0.02	0.9783	0.9783	0.9783	0.9783
Hungarian	44	0.10	0.9778	1.0000	0.9565	1.0000
Swedish	49	0.00	1.0000	0.9592	1.0000	0.9796
Finnish	46	0.04	0.9783	0.9783	0.9783	0.9783

5 CONCLUSION

For multi-media systems operating globally like YouTube, LID for the audio content is crucial to provide subtitles, ranking, recommendation and ads serving. Our experiments prove that our text based signal can be efficiently used instead or in addition to existing audio LID methods for multi-media services. Our model achieves 97% accuracy, it is easy to implement and drastically improves the coverage for LID in comparison to uploader-assigned language labels which do not provide adequate coverage and can be noisy. Furthermore, our model provides a reliable alternative for audio based LID in ASR systems which are not available for all languages and face various challenges in processing naturalistic multi-party conversations as observed in YouTube videos.

REFERENCES

- [1] Steven Abney and Steven Bird. 2010. The human language project: building a Universal Corpus of the world's languages. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 88–97.
- [2] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- [3] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>
- [4] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297. DOI: <https://doi.org/10.1023/A:1022627411411>
- [5] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation* 10 (1998), 1895–1923.
- [6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* 9 (June 2008), 1871–1874. <http://dl.acm.org/citation.cfm?id=1390681.1442794>
- [7] Javier Gonzalez-Dominguez, David Eustis, Ignacio Lopez-Moreno, Andrew Senior, Francoise Beaufays, and Pedro J Moreno. 2015. A Real-Time End-to-End

- Multilingual Speech Recognition Architecture. *Selected Topics in Signal Processing, IEEE Journal of* 9, 4 (2015), 749–759.
- [8] Google. 2016. Youtube statistics. (2016). Available at <https://www.youtube.com/yt/press/statistics.html>.
- [9] Google. 2017. Google Cloud Platform: Detecting Languages. <https://cloud.google.com/translate/docs/detecting-language>. (2017). Accessed: 2017-05-15.
- [10] David Imseng, Mathew Magimai.-Doss, and Hervé Bourlard. 2010. *Hierarchical Multilayer Perceptron based Language Identification*. Idiap-RR Idiap-RR-14-2010. Idiap.
- [11] Wang Ling, Guang Xiang, Chris Dyer, Alan W Black, and Isabel Trancoso. 2013. Microblogs as Parallel Corpora. In *ACL (1)*. 176–186.
- [12] Nick Littlestone. 1988. Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm. *Mach. Learn.* 2, 4 (April 1988), 285–318. DOI: <https://doi.org/10.1023/A:1022869011914>
- [13] Robert Malouf. 2002. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20 (COLING-02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–7. DOI: <https://doi.org/10.3115/1118853.1118871>
- [14] Thomas Niesler and Daniel Willett. 2006. Language identification and multilingual speech recognition using discriminatively trained acoustic models. In *Multilingual Speech and Language Processing*.
- [15] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li. 2013. Shifted-Delta MLP Features for Spoken Language Recognition. *IEEE Signal Processing Letters* 20 (2013), 15–18.