Word Level Language Identification in Online Multilingual Communication

Dong Nguyen[1] **A. Seza Doğruöz**[23]
(1) Human Media Interaction, University of Twente, Enschede, The Netherlands
(2) Tilburg School of Humanities, Tilburg University, Tilburg, The Netherlands
(3) Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
dong.p.ng@gmail.com, a.s.dogruoz@gmail.com

Abstract

Multilingual speakers switch between languages in online and spoken communication. Analyses of large scale multilingual data require automatic language identification at the word level. For our experiments with multilingual online discussions, we first tag the language of individual words using language models and dictionaries. Secondly, we incorporate context to improve the performance. We achieve an accuracy of 98%. Besides word level accuracy, we use two new metrics to evaluate this task.

1 Introduction

There are more multilingual speakers in the world than monolingual speakers (Auer and Wei, 2007). Multilingual speakers switch across languages in daily communication (Auer, 1999). With the increasing use of social media, multilingual speakers also communicate with each other in online environments (Paolillo, 2011). Data from such resources can be used to study code switching patterns and language preferences in online multilingual conversations. Although most studies on multilingual online communication rely on manual identification of languages in relatively small datasets (Danet and Herring, 2007; Androutsopoulos, 2007), there is a growing demand for automatic language identification in larger datasets. Such a system would also be useful for selecting the right parsers to process multilingual documents and to build language resources for minority languages (King and Abney, 2013).

In this paper, we identify Dutch (NL) en Turkish (TR) at the word level in a large online forum for Turkish-Dutch speakers living in the Netherlands. The users in the forum frequently switch languages within posts, for example:

> <TR> *Sariyi ver* </TR>
> <NL> *Wel mooi doelpunt* </NL>

So far, language identification has mostly been modeled as a document classification problem. Most approaches rely on character or byte n-grams, by comparing n-gram profiles (Cavnar and Trenkle, 1994), or using various machine learning classifiers. While McNamee (2005) argues that language identification is a solved problem, classification on a more fine-grained level (instead of document level) remains a challenge (Hughes et al., 2006). Furthermore, language identification is more difficult for short texts (Baldwin and Lui, 2010; Vatanen et al., 2010), such as queries and tweets (Bergsma et al., 2012; Carter et al., 2012; Ceylan and Kim, 2009). Tagging individual words (without context) has been done using dictionaries, affix statistics and classifiers using character n-grams (Hammarström, 2007; Gottron and Lipka, 2010). Although Yamaguchi and Tanaka-Ishii (2012) segmented text by language, their data was artificially created by randomly sampling and concatenating text segments (40-160 characters) from monolingual texts. Therefore, the language switches do not reflect realistic switches as they occur in natural texts. Most related to ours is the work by King and Abney (2013) who labeled languages of words in multilingual web pages, but evaluated the task only using word level accuracy.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 857–862,
Seattle, Washington, USA, 18-21 October 2013.

Our paper makes the following contributions: 1) We explore two new ways to evaluate the task for analyzing multilingual communication and show that only word accuracy gives a limited view 2) We are the first to apply this task on a conversational and larger dataset 3) We show that features using the context improve the performance 4) We present a new public dataset to support research on language identification.

In the rest of the paper, we first discuss the related work and describe our dataset. Secondly, we present our experiments. We finally conclude with a summary and suggestions for future work.

2 Corpus

Our data[1] comes from one of the largest online communities in The Netherlands for Turkish-Dutch speakers. All posts from May 2006 until October 2012 were crawled. Although Dutch and Turkish dominate the forum, English fixed phrases (e.g. *no comment, come on*) are also occasionally observed. Users switch between languages within and across posts. Examples 1 and 2 illustrate switches between Dutch and Turkish within the same post. Example 1 is a switch at sentence level, example 2 is a switch at word level.

Example 1:
<NL>Mijn dag kan niet stuk :) </NL>
<TR> Cok guzel bir haber aldim </TR>
Translation: <NL> This made my day:)
</NL><TR> I received good news
</TR>

Example 2:
<TR>kahvalti</TR><NL>met
vriendinnen by my thuis </NL>
Translation: <TR>breakfast </TR>
<NL> with my girlfriends at my home
</NL>

The data is highly informal with misspellings, lengthening of characters (e.g. *hotttt*), replacement of Turkish characters (*kahvalti* instead of *kahvaltı*) and spelling variations (*tankyu* instead of *thank you*). Dutch and Turkish sometimes share common spellings (e.g. *ben* is *am* in Dutch and *I* in Turkish), making this a challenging task.

Annotation
For this research, we classify words as either Turkish or Dutch. Since Dutch and English are typologically more similar to each other than Turkish, the English phrases (less than 1%) are classified as Dutch. Posts were randomly sampled and annotated by a native Turkish speaker who is also fluent in Dutch. A native Dutch speaker annotated a random set of 100 posts (Cohen's kappa = 0.98). The following tokens were ignored for language identification:

- Smileys (as part of the forum markup, as well as textual smileys such as ":)").
- Numeric tokens and punctuation.
- Forum tags (e.g. *[u]* to underline text).
- Links, images, embedded videos etc.
- Turkish and Dutch first names and place names[2].
- Usernames when indicated with special forum markup.
- Chat words, such as *hahaha*, *ooooh* and *lol* recognized using regular expressions.

Posts for which all tokens are ignored, are not included in the corpus.

Statistics
The dataset was randomly divided into a training, development and test set. The statistics are listed in Table 1. The statistics show that Dutch is the majority language, although the difference between Turkish and Dutch is not large. We also find that the documents (i.e. posts) are short, with on average 18 tokens per document. The data represents realistic texts found in online multilingual communication. Compared to previously used datasets (Yamaguchi and Tanaka-Ishii, 2012; King and Abney, 2013), the data is noisier and the documents are much shorter.

	#NL tokens	**#TR tokens**	**#Posts/(BL%)**
Train	14900 (54%)	12737 (46%)	1603 (15%)
Dev	8590 (51%)	8140 (49%)	728 (19%)
Test	5895 (53%)	5293 (47%)	735 (17%)

Table 1: Number of tokens and posts for Dutch (NL) and Turkish (TR), including % of bilingual (BL) posts

[1] Available at http://www.dongnguyen.nl/data-langid-emnlp2013.html

[2] Based on online name lists and Wikipedia pages

3 Experimental Setup

3.1 Training Corpora

We used the following corpora to extract dictionaries and language models.

- *GenCor*: Turkish web pages (Sak et al., 2008).
- *NLCOW2012*: Dutch web pages (Schäfer and Bildhauer, 2012).
- *Blog authorship corpus*: English blogs (Schler et al., 2006).

Each corpus was chunked into large segments which were then selected randomly until 5M tokens were obtained for each language. We tokenized the text and kept the punctuation.

3.2 Baselines

As baselines, we use *langid.py*[3] (Lui and Baldwin, 2012) and van Noord's *TextCat* implementation[4] of the algorithm by Cavnar and Trenkle (1994). TextCat is based on the comparison of n-gram profiles and langid.py on Naive Bayes with n-gram features. For both baselines, words were entered individually to each program. Words for which no language could be determined were assigned to Dutch. These models were developed to identify the languages of the documents instead of words and we did not retrain them. Therefore, these models are not expected to perform well on this task.

3.3 Models

We start with models that assign languages based on only the current word. Next, we explore models and features that can exploit the context (the other words in the post). Words with the highest probability for English were assigned to Dutch for evaluation.

Dictionary lookup (`DICT`)
We extract dictionaries with word frequencies from the training corpora. This approach looks up the words in the dictionaries and chooses the language for which the word has the highest probability. If the word does not occur in the dictionaries, Dutch is chosen as the language.

Language model (`LM`)
We build a character n-gram language model for each language (max. n-gram length is 5). We use Witten-Bell smoothing and include word boundaries for calculating the probabilities.

Dictionary + Language model (`DICT+LM`)
We first use the dictionary lookup approach (`DICT`). If the word does not occur in dictionaries, a decision is made using the language models (`LM`).

Logistic Regression (`LR`)
We use a logistic regression model that incorporates context with the following features:

- (Individual word) Label assigned by the `DICT+LM` model.
- (Context) The results of the `LM` model based on previous + current token, and current token + next token (e.g. the sequence "*ben thuis*" (*am home*) as a whole if *ben* is the current token). This gives the language model more context for estimation. We compare the use of the assigned labels (`LAB`) with the use of the log probability values (`PROB`) as feature values.

Conditional Random Fields (`CRF`)
We treat the task as a sequence labeling problem and experiment with linear-chain Conditional Random Fields (Lafferty et al., 2001) in three settings:

- (Individual word) A CRF with only the tags assigned by the `DICT+LM` to the individual tokens as a feature (`BASE`).
- (Context). CRFs using the `LAB` or `PROB` as additional features (same features as in the logistic regression model) to capture additional context.

3.4 Implementation

Language identification was not performed for texts within quotes. To handle the alphabetical lengthening (e.g. *lolllll*), words are normalized by trimming same character sequences of three characters or more. We use the Lingpipe[5] and Scikit-learn (Pedregosa et al., 2011) toolkits for our experiments.

[3] https://github.com/saffsd/langid.py

[4] http://www.let.rug.nl/∼vannoord/TextCat/

[5] http://alias-i.com/lingpipe/

	Word classification					**Fraction**				**Post classification**	
	TR		**NL**				**MAE**				
Run	P	R	P	R	Acc.	ρ	All	Mono.	BL	F_1	Acc.
Textcat	0.872	0.647	0.743	0.915	0.788	0.739	0.251	0.264	0.188	0.386	0.396
LangIDPy	0.954	0.387	0.641	**0.983**	0.701	0.615	0.364	0.371	0.333	0.413	0.475
DICT	0.955	0.733	0.802	0.969	0.858	0.827	0.196	0.200	0.175	0.511	0.531
LM	0.950	0.930	0.938	0.956	0.944	0.926	0.074	0.076	0.065	0.699	0.703
DICT + LM	0.951	0.934	0.942	0.957	0.946	0.943	0.067	0.067	**0.063**	0.711	0.717
LR + LAB	0.965	0.952	0.958	0.969	0.961	0.917	0.066	0.066	0.068	0.791	0.808
LR + PROB	0.956	0.976	0.978	0.959	0.967	0.945	0.048	0.044	0.064	0.826	0.849
CRF + BASE	**0.973**	0.974	0.977	0.976	0.975	0.940	0.043	0.027	0.119	**0.858**	**0.898**
CRF + LAB	0.964	0.977	0.979	0.967	0.972	0.933	0.046	0.033	0.111	0.855	0.891
CRF + PROB	0.970	**0.980**	**0.982**	0.973	**0.976**	**0.946**	**0.039**	**0.025**	0.103	0.853	0.895

Table 2: Results of language identification experiments.

3.5 Evaluation

The assigned labels can be used for computational analysis of multilingual data in different ways. For example, these labels can be used to analyze language preferences in multilingual communication or the direction of the switches (from Turkish to Dutch or the other way around). Therefore, we evaluate the methods from different perspectives.

The evaluation at word and post levels is done with the following metrics:

- *Word classification* precision (P), recall (R) and accuracy. Although this is the most straightforward approach to evaluate the task, it ignores the document boundaries.
- *Fraction of language in a post*: Pearson's correlation (ρ) and Mean Absolute Error (MAE) of proportion of Turkish in a post. This evaluates the measured proportion of languages in a post when the actual tags for individual words are not needed. For example, such information is useful for analyzing the language preferences of users in the online forum. Besides reporting the MAE over all posts, we also separate the performance over monolingual and bilingual posts (BL).
- *Post classification*: Durham (2003) analyzed the switch between languages in terms of the amount of monolingual and bilingual posts. Our posts are classified as NL, TR or bilingual (BL) if all words are tagged in the particular language or both. We report F_1 and accuracy.

4 Results

The results are presented in Table 2. Significance tests were done by comparing the results of the word and post classification measures using McNemar's test, and comparing the MAEs using paired t-tests. All runs were significantly different from each other based on these tests ($p < 0.05$), except the MAEs of the `DICT+LM` and `LR+LAB` runs and the MAEs and post classification metrics between the CRFs runs.

The difficulty of the task is illustrated by examining the coverage of the tokens by the dictionaries. 24.6% of the tokens (dev + test set) appear in both dictionaries, 31.1% only in the Turkish dictionary, 30.5% only in the Dutch dictionary and 13.9% in none of the dictionaries.

The baselines do not perform well. This confirms that language identification at the word level needs different approaches than identification at the document level. Using language models result in a better performance than dictionaries. They can handle unseen words and are more robust against the noisy spellings. The combination of language models and dictionaries is more effective than the individual models. The results improve when context was added using a logistic regression model, especially with the probability values as feature values.

CRFs improve the results but the improvement on the correlation and MAE is less. More specifically, CRFs improve the performance on monolingual posts, especially when a single word is tagged in the wrong language. However, when the influence of the context is too high, CRFs reduce the performance in bilingual posts.

This is also illustrated with the results of the post classification. The LR+PROB run has a high recall (0.905), but a low precision (0.559) for bilingual posts, while the CRF+PROB approach has a low recall (0.611) and a high precision (0.828).

The fraction of Dutch and Turkish in posts varies widely, providing additional challenges to the use of CRFs for this task. Classifying posts first as monolingual/bilingual and tagging individual words afterwards for bilingual posts might improve the performance.

The evaluation metrics highlight different aspects of the task whereas word level accuracy gives a limited view. We suggest using multiple metrics to evaluate this task for future research.

Dictionaries versus Language Models

The results reported in Table 2 were obtained by sampling 5M tokens of each language. To study the effect of the number of tokens on the performance of the DICT and LM runs, we vary the amount of data. The performance of both methods increases consistently with more data (Figure 1). We also find that language models achieve good performance with only a limited amount of data, and consistently outperform the approach using dictionaries. This is probably due to the highly informal and noisy nature of our data.

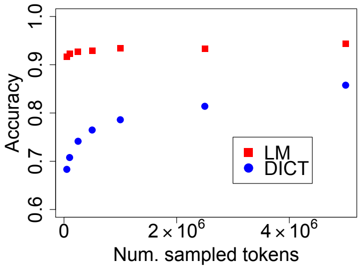

Figure 1: Effect of sampling size

Post classification

We experimented with classifying posts into TR, NL and bilingual using the results of the word level language identification (Table 2: post classification). Posts were classified as a particular language if all words were tagged as belonging to that language, and bilingual otherwise. Runs using CRFs achieved the best performance.

We now experiment with allowing a margin (e.g. a margin of 0.10 classifies posts as TR if at least 90% of the words are classified as TR). Allowing a small margin already increases the results of simpler approaches (such as the LR-PROB run, Table 3) by making it more robust against errors. However, allowing a margin reduces the performance of the CRF runs.

Margin	0.0	0.05	0.10	0.15	0.20
Accuracy	0.849	0.873	0.876	0.878	0.865

Table 3: Effect of margin on post classification (LR-PROB run)

Error analysis

The manual analysis of the results revealed three main challenges: 1) Our data is highly informal with many spelling variations (e.g. *moimoimoi*, *goooooooooooolllll*) and noise (e.g. *asdfghjfgshahaha*) 2) Words sharing spelling in Dutch and Turkish are difficult to identify especially when there is no context available (e.g. a post with only one word). These words are annotated based on their context. For example, the word *super* in *"Seyma, super"* is annotated as Turkish since *Seyma* is also a Turkish word. 3) Named entity recognition is necessary to improve the performance of the system and decrease the noise in evaluation. Based on precompiled lists, our system ignores named entities. However, some names still remain undetected (e.g. usernames).

5 Conclusion

We presented experiments on identifying the language of individual words in multilingual conversational data. Our results reveal that language models are more robust than dictionaries and adding context improves the performance. We evaluate our methods from different perspectives based on how language identification at word level can be used to analyze multilingual data. The highly informal spelling in online environments and the occurrences of named entities pose challenges.

Future work could focus on cases with more than two languages, and languages that are typologically less distinct from each other or dialects (Trieschnigg et al., 2012).

6 Acknowledgements

The first author was supported by the Netherlands Organization for Scientific Research (NWO) grant 640.005.002 (FACT) and the second author through a postdoctoral research grant in E-Humanities (Digital Humanities) by Tilburg University (NL). The authors would like to thank Mariët Theune and Dolf Trieschnigg for feedback.

References

J. Androutsopoulos, 2007. *The multilingual internet. Language, Culture and communication online*, chapter Language choice and code-switching in German-based diasporic web-forums., pages 340–361. Oxford: Oxford University Press.

P. Auer and L. Wei. 2007. Introduction: Multilingualism as a problem? Monolingualism as a problem? In *Handbook of Multilingualism and Multilingual Communication*, volume 5 of *Handbooks of Applied Linguistics*, pages 1–14. Mouton de Gruyter.

P. Auer. 1999. From codeswitching via language mixing to fused lects toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4):309–332.

T. Baldwin and M. Lui. 2010. Language identification: the long and the short of the matter. In *Proceedings of NAACL 2010*.

S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*.

S. Carter, W. Weerkamp, and M. Tsagkias. 2012. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, pages 1–21.

W.B. Cavnar and J. M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*.

H. Ceylan and Y. Kim. 2009. Language identification of search engine queries. In *Proceedings of ACL 2009*.

B. Danet and S. C. Herring. 2007. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press Oxford.

M. Durham. 2003. Language choice on a Swiss mailing list. *Journal of Computer-Mediated Communication*, 9(1).

T. Gottron and N. Lipka. 2010. A comparison of language identification approaches on short, query-style texts. In *Proceedings of ECIR 2010*.

H. Hammarström. 2007. A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007*.

B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. Mackinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of LREC 2006*.

B. King and S. Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL 2013*.

J. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*.

M. Lui and T. Baldwin. 2012. langid.py: an off-the-shelf language identification tool. In *Proceedings of ACL 2012*.

P. McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

J.C. Paolillo. 2011. “Conversational” codeswitching on Usenet and Internet Relay Chat. *Language@Internet*, 8(3).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

H. Sak, T. Güngör, and M. Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *GoTAL 2008*, volume 5221 of *LNCS*, pages 417–427. Springer.

R. Schäfer and F. Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC 2012*.

J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

D. Trieschnigg, D. Hiemstra, M. Theune, F. Jong, and T. Meder. 2012. An exploration of language identification techniques for the Dutch folktale database. In *Adaptation of Language Resources and Tools for Processing Cultural Heritage workshop (LREC 2012)*.

T. Vatanen, J. J. Väyrynen, and S. Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of LREC 2010*.

H. Yamaguchi and K. Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings of ACL 2012*.