Automatic Language Identification

M.A. Nejla Qafmolla

Tirana University**,** Faculty of Foreign Languages**,** English Department **–** Tirana, Albania

Abstract

Automatic Language Identification (LID) is the process of automatically identifying the language of spoken utterance or written material. LID has received much attention due to its application to major areas of research and long-aspired dreams in computational sciences, namely Machine Translation (MT), Speech Recognition (SR) and Data Mining (DM). A considerable increase in the amount of and access to data provided not only by experts but also by users all over the Internet has resulted into both the development of different approaches in the area of LID – so as to generate more efficient systems – as well as major challenges that are still in the eye of the storm of this field. Despite the fact that the current approaches have accomplished considerable success, future research concerning some issues remains on the table. The aim of this paper shall not be to describe the historic background of this field of studies, but rather to provide an overview of the current state of LID systems, as well as to classify the approaches developed to accomplish them. LID systems have advanced and are continuously evolving. Some of the issues that need special attention and improvement are semantics, the identification of various dialects and varieties of a language, identification of spelling errors, data retrieval, multilingual documents, MT and speech-to-speech translation. Methods applied to date have been good from a technical point of view, but not from a semantic one.

Keywords: Human Language Technologies, Automatic Language Identification (LID), Data Mining (DM), Spoken Language Identification (SLID), Written Language Identification (WLID)

Introduction

Automatic Language Identification (LID) is the process of automatically identifying the language of spoken utterance or written material. (Amine et al.: 95) LID has received much attention due to its application to major areas of research and long-aspired dreams in computational sciences, namely Machine Translation (MT), Speech Recognition (SR), and Data Mining (DM), to mention a few. A considerable increase in the amount of and access to data provided not only by experts but also by users all over the Internet has resulted into both the development of different approaches in the area of LID – so as to generate more efficient systems – as well as major challenges that are still in the eye of the storm of this field. Nowadays, LID systems are being used in connection with different fields; although the same basic approaches introduced and developed in the 1990s are still in use. Despite the fact that the current approaches have accomplished considerable success, future research concerning some issues – especially a greater incorporation of semantic content in the different LID systems – remains on the table. The field of the LID activity dates back to the 1970s, and a considerable number of methods have been developed in its furtherance. Due to the requirements that rule the following project, the goal shall not be to describe the historic background of this field of studies, but rather to provide an overview of the current state of LID systems, as well as to classify the approaches developed to accomplish them.

Spoken and Written LID

Due to the fact that both spoken and written information serves as the input of LID systems, some initial differences must be made. It must be pointed out, however, that the methods used to accomplish LID do not necessarily differ in essence – with some exceptions. In other words, the approaches used are substantially the same. The difference lies at the initial stage of the architecture of the overall LID system varies.

Spoken LID

Spoken LID (SLID) is carried out in basically two steps: training and recognition. (Zissman: 118) Initially, speech samples of different languages are provided to the system. Such samples are the training material, meaning the corpora that serve as background information. Such information is analyzed and characteristics, usually in the form of vectors, are extracted. With this information, language models are produced, to which future speech input will then be compared to. The model that most likely appears to be similar to the input will be the one selected by the system. (Zissman: 118-121) There are some techniques that are specifically focused on SLID. The similarities between the language model and the speech input were initially measured in terms of similar spectra. However, further approaches resorted to the segmentation of the input speech signal: comparison was carried out on the bases of such smaller segments and recognition was possible by the use of Artificial Neural Networks (ANN) (Zissman: 121). Another approach resorted to the use of phonotactic constraints, which are the set of language limitations as to which phonemes can occur in a given context and which cannot. (Zissman: 121-122). This method is usually used in combination with a Phone Recognizer (PR) per language. PRs are mostly n-gram-based or otherwise combined with Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs), which are also applicable to Written LID (WLID) and shall be explained later on.

Written LID

Like SLID, WLID also entails a training stage, when corpora is gathered and analyzed, and an identification stage, at which point the models generated before are compared to the new written input (Řehůřek & Kolkus: 357-358). Some approaches are particular to SLID, but since the popularity and efficiency of statistical methods were sufficiently stated during the mid-1990s, they were used for both types of input data: after extracting relevant features, data is available in the form of strings and, therefore, analyzed statistically with basically the same approaches.

Approaches to LID

The approaches used to carry out the LID activity can be categorized as follows: those which apply language knowledge and those which use statistical methods. As mentioned before, some of them concern specifically SLID. Despite their differences, some common attempts can be mentioned: to divorce as much as possible the LID activity from language knowledge; to deal with errors, such as misspelled words, as effectively as possible; to deal with multilingual texts; to use as little storage and time as possible without losing efficiency.

Spectra Comparison

This approach is applied to spoken input. The basic procedure entails the use of speech samples of about 10 msec. from either one language or more and the extraction of relevant features. With these features, phonemic or orthographical transcriptions are produced and all this data is used as background corpora. The same procedure is applied to the input data. All this training data is then used for the identification task, i.e. new speech input is compared to the stored data and the language which is most similar to it is the selected one. (Zissman: 120-121)

Segmentation and ANNs

This technique is similar to the previous one, insofar as it is also used with SLID and it also entails a training phase based on speech samples. However, segmentation is applied so as to capture a given language unique features in terms of prosody, pitch, duration and amplitudes, spectral features in vowels and consonants, amongst others, in smaller portions of both the samples and the input. The training data is then compared and the language with the highest degree of likelihood is selected. (Muthusamy: 4) Each segmented input is then matched to a given stored set of data via ANN. (Muthusamy: 53-67)

Phonotactic Rules and PRs

Each language imposes certain constraints as to which phonemes can appear in a given context. These constraints have been used for the SLID activity. The PR initially tokenizes the input data and applies techniques which are also used for WLID, namely n-grams, GMMs and HMMs. (see below: HMMs, GMMs, and SVMs)

Short-word-based LID

A basic language-dependent approach towards LID is the use of so-called common or short words, typically determiners, conjunctions and prepositions. (Grefenstette: 3). The basic procedure this approach entails is fairly simple and it involves an initial tokenization of the training and input data, followed by the assignment of probability values to each token, representing such small words. Finally, the resulting statistics are compared so as to determine the language of a given text. (Grefenstette: 3) The idea of using the otherwise referred to as function or stop words (Řehůřek & Kolkus: 358) is based on Zipf's Law: such rule basically states that there is always a set of words in a given language consisting of words which are more frequent than others. The fundamental ground behind this method is based on the notion that human beings need only little information to determine in which language a given text is in, and they need not even be proficient at such language to do so. (Řehůřek & Kolkus: 358, Dunning: 1)

This approach does not need the aid of more complex language rules such as syntax and semantics. (Řehůřek & Kolkus: 358). Although this technique is still the subject of work, as it shall be explained later on, its simplicity makes it difficult for a LID system of such nature to deal with multilingual texts (Dunning: 3), or even texts with errors, for a misspelled word would be left out of statistical results (Cavnar & Trenkle: 1). Furthermore, a system using this method might encounter problems when dealing with short texts, in which occurrence of stop words might be relatively low so as to generate sensitive statistics (Dunning: 4). With the intention of dealing with these issues more successfully and efficiently, further approaches were developed.

N-gram-based LID

An n-gram is an n-character slice of a longer string containing blanks to the beginning and ending so as to categorize it as a beginning- or ending of-word-n-gram (Cavnar & Trenkle: 2). This system was primarily developed in literature and experiments in the 1990s and remains a useful and experimented approach in LID. The system basically works as follows:

A set of texts are used as the sources to build a model;

Such texts are tokenized, leaving out digits and punctuation marks;

Tokens are scanned and as many n-grams as possible are generated;

A ranking is obtained with the most frequent n-grams.

Example 1:

bi-grams: _T, TE, EX, XT, T_
tri-grams: _TE, TEX, EXT, XT_, T_ _
quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

Examples of different types of n-grams from the word 'text'. (Cavnar & Trenkle: 2).

Example 2:

Joh 2

ohn 2

hn 2

n k 2

ki 2

kis 2

iss 2

sse 2

sed 2

ed 2

d M 1

Ma 1

Mar 1

ary 1

ry. 1

y. 1

.J 1

Example of n-grams generated from the string 'John kissed Mary.' [INT 1]

The result of these basic four steps in the creation of a profile or language model, on the basis of which a given text in a given language, is later on compared to and, therefore, the language in which it is written determined. Since the language models are derived by assuming only that the n-grams are sequences of bytes, no language specific pre-processing is required (Dunning: 6). Due to its popularity, Dunning (1995) introduced the use of HMMs as a way to calculate the probability of a sequence of strings. In addition with the Bayesian deciding rule, when the system is faced with two possibilities, it shall select the most likely cause. (Dunning: 8-10) This renders the system nondeterministic and more dynamic when faced with alternatives.

This method has certain benefits, namely:

It is an alternative to generating statistics based on whole words, since the latter approach would be rendered unsatisfactory when faced with texts coming from noisy sources and, therefore, containing mistakes or invariances such as misspelled words, loan words, among others. (Cavnar & Trenkle: 12);

Whole words statistics would also require longer texts in order to produce sensitive results, a difficulty that is overcome to a certain extent by n-gram-based systems (Cavnar & Trenkle: 12);

It is an alternative to the systems that require building a lexicon or using a set of morphological processing rules, such as stemming, which would entail through knowledge on both the corpora texts and those subject to LID (Cavnar & Trenkle: 12);

Not only is it possible to identify a text's language, but also to classify it on the bases of its topic, by measuring the most frequent n-grams in a given field in terms of subject similarity (Cavnar & Trenkle: 9);

The addition of the Markov model provided the chance of mathematically manipulating the algorithm that determines the sequence of n-grams in an easier way. (Dunning: 6-7)

HMMs, GMMs, and SVMs

These three statistical techniques have been used in combination with basically all the approaches mentioned so far, due to the fact that they can be used in both SLID, as well as in WLID. Basic advantages are quicker and more efficient LID activity, with smaller storage consumption. As already stated, they are also easy to manipulate mathematically.

HMM is a stochastic model which consists in the parsing of both the training data and the input text. Once nodes are generated, transition probabilities between each of the nodes are determined. From one node to another, the system determines the transition probabilities based solely on the present state. (Dunning: 6-8) [INT 2]

GMM is a technique to accomplish clustering of subpopulations by repeatedly running the algorithm along the input data (written or spoken). The method works basically as follows: cluster centers are determined in the input data; multiple iterations are applied to the initially estimated centers so as to maximize the initial result and obtain more clearly defined clusters; the clusters ultimately determine the data as belonging to a given hypothesized language. (Zissman: 124-125)

Support Vector Machines (SVMs), more frequently used in SLID, are also applied to achieve the maximum degree of clustering. The function of the SPM is to determine patterns in the input and match them to those found before in the training data. (Amine et al.: 98)

Cross-Entropy Systems

Systems using entropy are usually combined with the n-gram method and serve as an aid to HMM and GMM algorithms. The incorporation of this concept attempts to help the system when faced with alternatives and to use a probabilistic technique to predict the sequence of strings. The concept of cross-entropy entails the degree of uncertainty involved in choosing a symbol: the more considerable the entropy, the more uncertain it is to select a given symbol (Teahan: 2). In addition to the process, as described above, the use of entropy implies that the probabilities related to a given character that has followed another given character in the past are used to as to predict which character will follow in future situations with the same characteristics. (Teahan: 4)

Dictionary Method

A further development on the arena of the word-based and language-dependent approach was presented by R. Řehůřek and M. Kolkus. The idea is not only to use frequent stop words from a given language, but rather to implement an algorithm to detect how relevant those words are in a non-binary, graded manner. (Řehůřek & Kolkus: 361). Therefore, this approach resorts to language knowledge and makes use of the precision/recall equation used by information retrieval systems so as to identify the language of a text. The basic algorithm entails the following concepts:

$$rel(word, language) : W \times L \mapsto \mathbb{R}$$

Source: (Řehůřek & Kolkus: 361)

The algorithm presents *W*, representing all words present in the training or background data, and *L*, being the set of considered languages. The concept *rel* represents the relevance value of the words.

Applications

There are two main categories of applications where the LID process can be performed: on the one hand speech-based applications, and on the other text-based applications.

Spoken LID

SLID is used in many applications and even though it has been widely used, it has encountered many problems, especially when dealing with dialectal variances, accent identification and differences from speaker to speaker (gender, age, social background, etc.).

Speech-to-Speech Translation

LID is applicable in speech-to-speech translation at the very first stage. It captures the spoken utterance, in this case, the source language, identifies it and then the translation system converts it into the target language.

Dialect and Accent Identification

The task of dialect and accent identification is that of identifying the spoken dialect or accent by using examples of the input language. Investigation of a dialect or of a non-native accent can contribute to forensic speech science as well. A speaker's first language can be identified and consequently his/her nationality by analyzing the characteristics present in the foreign spoken utterance. (Amino & Osanai: 236)

Telephone Speech

Companies make use of telephone-based applications to reduce the costs involving the hiring of human staff. The aim of these systems is to identify the language of a caller and if applicable route the call to the appropriate receiver, who is fluent in that language. (Santhi & Raja: 2) Banking institutions and airline companies usually make use of this system.

Written LID

Due to the increase in the amount of written data, the need for WLID systems is required, in order to identify and categorize it. Additionally, such data is available in many different languages and sometimes containing mistakes, such as typos, misspelled words, etc. Like SLID, WLID must deal with these difficulties and challenges.

Data Mining

In this case, the aim of LID systems is to identify the languages in large amounts of data and to support identification multilingual texts, texts coming from sources, such as Optical Character Recognition (OCR) systems, or even handwritten texts (see below Script ID). Google, for instance uses LID to identify a language of a certain Web page and if the language is different from that of the IP of the user, Google offers to translate it automatically. Furthermore, data can be categorized and subcategorized by topic or genre. (Teahan: 9)

Machine Translation

LID is used at the beginning of the MT process for identifying the language in which the text is in and even to identify the multilingual chunks. [INT: 3]

Authorship Ascription

Different authors have different styles and statistical properties that can be assigned to them. By analyzing as much written work from a given author as possible, such statistical properties are established and authorship can be determined. This is useful to detect plagiarism and even the identity of an author. (Teahan: 8)

Spell-checking and Correction

LID is also useful and an essential component in spell-checking and correction systems in text editors. It aims at identifying a language regardless of the mistakes contained in the text and offers the user to correct them. These mistakes may also occur on account of text coming from noisy sources, such as OCR. (Teahan: 11; Bergsma, et al.: 65-68)

Word Segmentation

Word segmentation is concerned with determining where the word begins and ends. One of the main challenges of LID systems in this context is when dealing with languages, such as Asian languages, whose words do not have clear-cut boundaries. (Teahan: 16) Word segmentation is important for data retrieval systems.

Script ID

A major advantage of handwritten LID systems is to provide more sources for data retrieval and to detect features that are present in a given writer. (Ramanathan: 933). Since handwriting differs from one writer to another and since we do not always write a certain character exactly in the same way, creating a general and reliable recognition system is very challenging. Commonly used features in character recognition are: zoning features, structural feature, directional features, crossing points and contours. (Ramanathan: 933)

Online LID

LID systems are also offered online (see examples below State of the Art). Users can type in or otherwise copy a text into the web page and the LID system identifies the language it is written in. Another LID technique is used by "Google Translate", so as to detect the source language (from 80 languages supported) in case the user does not know it, and later on the MT system proceeds with the translation process. [INT 4]

State of the Art

Since its beginnings, LID research and development has produced a variety of applications. Nowadays, the increasing number of users of mobile devices and computers has produced a growing interest in developing applications directed to their needs and interconnectivity established amongst them. As has already been mentioned, online LID systems can be easily spotted on the Internet. With a varying degree of sophistication and accuracy, the Internet offers the public a user-friendly and even websites and companies that can be found online:

Rosette Language Identifier	*http://www.basistech.com/text-analytics/rosette/language-identifier/)*
Lextek Language Identifier	*http://www.lextek.com/langid/li/*
Collection of LIDs	*http://transdict.com/guessers.php*
Automatic Source Language Identification by LEC	*http://www.lec.com/help.asp?app=Translate&family=Translate&page=Features/LanguageIdentification.htm*
LID by Translated Labs	*http://labs.translated.net/language-identifier/*

LID systems are, though in a lesser amount, also available for mobile devices. An example is the application Language Detector, which supports over 50 languages. Popular enough is the mobile application of Google Translate. Both the online and mobile versions can detect an incorrect input language, as mentioned before. Additionally, the mobile application can perform this task quite accurately even with handwritten input. Moreover, mobile applications that resort to LID technology can be found in predictable keyboards. An example is the customizable keyboard for Blackberry, which detects the language the user is typing in real time. [INT 5] Another example is also available for Android users with *Adaptxt*: the user downloads this application and the languages to be used from a range supported by the application. Then the user can type in those languages freely. If, for example, the user selects one language within a set of downloaded languages as the input one, but chooses to type in another one of such set, the application will be able to predict it and offer the corresponding alphabet and dictionary. [INT 6]

A major concern in this area is text processing. Apple has provided a recent development to this respect with a strong emphasis in multilingual documents. By applying an automatic language identifier, the user can type in different languages within the same document without having to constantly change the settings. This makes not only the word processing task much more dynamic, but also the spellchecking activity less constrained, which can be accomplished either automatically or at the user's request. [INT 7; INT 8]

With companies' focusing more and more on their audiences' interest, it is now possible for experienced users to access language identification code language so as to generate their own software and improve the existing one. Google's Compact Language Detection tool and Python's language detection code are not available for users [INT 9; INT 10] and results can already be witnessed. [INT 11]

As already mentioned DM is another important and challenging sphere. Multilingual documents are at the core of LID research and NER has already received attention from developers and researchers. The techniques used are basically statistical, but DM calls for a bigger incorporation of semantic content. In furtherance of increasing such content and focusing on cultural differences, some LID systems resort now to the analysis of publicly available data, such as Twitter and Wikipedia. [INT 12] (Bergsma et al.: 67)

Future

LID systems have advanced and are continuously evolving. Some of the issues that need special attention and improvement are semantics, the identification of various dialects and varieties of a language, identification of spelling errors, data retrieval, multilingual documents, MT and speech-to-speech translation.

Semantics is one of the main issues that need particular attention in the context of LID systems. In order to identify a language, it is often necessary to analyze the context and the content of the information provided. Methods applied to date have been good from a technical point of view, but not from a semantic one. Code switching poses a challenge not only from the point of view of its multilingual nature, but rather, and most importantly, from the point of view of its semantic complexity: in a text, a social network post or an utterance, people might switch their communicational code for social reasons, such as register, social standard and context, style, etc. [INT 13]

With the increasing use of social networks (Twitter, LinkedIn, Facebook, etc.), multilingual speakers switch between languages in online environments, consequently there is a growing demand for LID in larger datasets, rather than small ones, which can also be useful for creating language resources for minority languages. (Nguyen & Doğruöz: 857-858).

Furthermore, the highly informal spelling, grammatical errors, unedited text by ordinary people and the occurrence of NE pose challenges to LID systems. (Nguyen & Doğruöz: 861). Thus, what researchers are trying to do is to improve the system's ability to cope with informal writing, multilingual texts, code switching, very short texts and unbalanced data. (Bergsma et al.: 65-66)

DM constitutes another major challenge to LID systems concerning a number of technical issues, such as clustering, data summarization, classification, finding dependency networks, analyzing changes and detecting anomalies (Mohamed Jafar & Sivakumar: 1), as well as non-technical ones. Due to the growing amount of data produced by users in social networks, such as Twitter, Facebook, Wikipedia, LinkedIn, and the like, it is imperative for LID systems to deal with semantic content, grammar mistakes, informal spelling, dialectal variances, with minority languages and languages that are quite similar to each other or share the same family (Hosford: 34-35; Řehůřek & Kolkus: 360), with loan words. (Zeng et al.: 198-200) It is worth noticing that the material produced by users is a rich source from which researchers and developers can be profited when information on certain languages is scarce. (Nguyen & Doğruöz: 857-858)

Script ID, OCR documents and classification are important fields of research. The identification facilitates automatic transcription of multilingual documents and search for documents on the Web containing a particular script. (Ramanathan: 1892). Today, LID systems face difficulties in identifying the information provided by these sources, due to different writing styles, character size, shapes, fonts, spacing between the lines and words on the one hand, and errors and typos, on the other, thus leading to unreliable results. The existing systems have demonstrated good performance, but of course there is still room for further work. For the identification process to be considered successful and efficient, it must show reliable results in spite of textual errors and the quality of the OCR or handwritten document, and it must use as little storage and processing time as possible. (Cavnar & Trenkle: 1-2, 11-12). Additionally, one of the main areas of improvement is to develop a method for accurately identifying text lines and words in a document, which can also be extended to page segmentation. (Ramanathan: 1895) It is also an important issue for current LID systems to accomplish a better rate of efficiency when determining the primary language in a source text. [INT 3]

Named Entity Recognition (NER) is another application in which there is still need for improvement. Its aim is to help the computer recognize NEs and classify them through contextual rules and syntax information into classes, such as person, organization, location, abbreviation, measure, number, term, date and time, etc. Current NER systems can handle multi-tokens entities, but are not able to identify entity boundaries. (Yohan et al.: 173, 179).

Furthermore, other areas in the context of LID systems which need improvement are MT and speech-to-speech translation: the aim is to successfully identify languages, dialects, sub-dialects or varieties, specific terminology, proper names, etc., and then proceeding with the translation process; and as far as speech-to-speech translation is concerned, handling with real-time conversations in different environments, including meetings, laptop or mobile conversations, web conferences, webinars, trainings, tourist information, in hotels, airports, etc.

References

[1] *Amine, A / Elberrichi, Z. / Simonet M.* 2010 "Automatic Language Identification: An Alternative Unsupervised Approach using a New Hybrid Algorithm." In: International Journal of Computer Science and Applications, Technomathematics Research Foundation. 7(1). January 2010. pp. 94-107, 2010. Available online: http://www.tmrfindia.org/ijcsa/v7i16.pdf Last Access: 3 March 2014.

[2] *Amino, K. / Osanai, T.* 2011 "Realization of the Prosodic Structure of Spoken Telephone Numbers by Native and Non-Native Speakers of Japanese." In: Proceedings of the 17th International Congress of Phonetic Science. Hong Kong. China. 17-21 August, 2011. 236-239 Available Online: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Amino/Amino.pdf Last Access: 22 June 2015.

[3] *Bergsma, Sh./McNamee, P. / Bagdouri, M. / Fink, C. / Wilson, Th.* 2012 "Language Identification for Creating Language-Specific Twitter Collections." In: Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012). Montreal, Canada. 7 June 2012. Available Online: http://aclweb.org/anthology//W/W12/W12-2108.pdf Last Access: 3 March 2014.

[4] *Cavnar, W.B. / Trenkle, J.M.* 1994 "N-Gram-Based Text Categorization." In: Proceedings of SDAIR-94 Annual Symposium on Document Analysis and Information Retreival. Michigan, USA, 1994, pp. 161-175. Available

Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.2994&rep=rep1&type=pdf Last access: 3 March 2014.

[5] *Dunning. T.* 1994 "Statistical Identification of Language". Technical Report CRL MCCS-94--273, Computing Research Laboratory. New Mexico State University. USA. March 1994. Available Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.1958&rep=rep1&type=pdf Last access: 3 March 2014.

[6] *Grefenstette, G.* 1995. "Comparing Two Language Identification Schemes." In: 3rd International Conference on Statistical Analysis of Textual Data. Rome, Italy. 11-13 December 1995. Available Online:http://webcache.googleusercontent.com/search?q=cache:71qw69YFXQIJ:xrce.fr/content/download/18502/133266/file/Gref---Comparing-two-language-identification-schemes.pdf+&cd=1&hl=de&ct=clnk&gl=de. Last Access: 3 March 2014.

[7] *Hosford, A.W.* 2011 "Automatic Language Identification (LID) Through Machine Learning." Master Thesis. School of Informatics, University of Sussex, UK. May 2011. Available Online: https://www.sussex.ac.uk/webteam/gateway/file.php?name=hosford-proj.pdf&site=20. Last Access: 6 March 2014.

[8] *Mohamed Jafar, O.A. / Sivakumar, R.* 2000. "Ant-based Clustering Algorithms: A Brief Survey" In: International Journal of Computer Theory and Engineering, Vol. 2, No. 5, 1793-8201. Available Online: http://www.ijcte.org/papers/242-G730.pdf Last Access: 3 March 2014.

[9] *Muthusamy, Y.K.* 1993. "A Segmental Approach to Automatic Language Identification." Ph.D. Thesis. Oregon Graduate Institute of Science & Technology. USA. October 1993. Available Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.4402&rep=rep1&type=pdf Last Access: 3 March 2014.

[10] *Nguyen, D. / Doğruöz, A.S.* 2013. "Word Level Language Identification on Online Multilingual Communication." In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, 18-21 October 2013. Available Online: http://www.dongnguyen.nl/publications/nguyen-emnlp2013.pdf Last Access: 3 March 2014.

[11] *Ramanathan, P.* 2013 "Automatic Identification of Handwritten Scripts." In: Middle-East Journal of Scientific Research 16 (12). ISSN 1990-9233. IDOSI Publications. 2013. pp. 1892-1895. DOI: 10.5829/idosi.mejrs.2014.19.7.1480. Available Online: http://www.idosi.org/mejsr/mejsr16(12)13/56.pdf Last Access: 3 March 2014.

[12] *Řehůřek, R. / Kolkus, M.* 2009 "Language Identification on the Web: Extending the Dictionary Method." In: Proceedings of the Computational Linguistic and Intelligent Text Processing. 10th International Conference, CICLing 2009, Mexico City. 24-28 February 2009. Available Online: http://folk.ntnu.no/sandsmar/langdetect.pdf Last Access: 3 March 2014.

[13] *Santhi, S. / Raja, S.* 2013. "An Automatic Language Identification using Audio Features." In: Proceedings of the International Conference on Information Systems and Computing (ICISC). India. 1 January 2013. Pp. 358-364. Available Online: http://www.ijetae.com/files/Conference%20ICISC-2013/IJETAE_ICISC_0113_60.pdf Last Access: 6 March 2014.

[14] *Teahan, W. J.* 2000. "Text Classification and Segmentation using Minimum Cross-Entropy." In: Proceedings of the RIAO-00, 6th International Conference "Recherche d'Information Assistée par Ordinateur". Paris, France. 12-14 April 200. Available Online: http://vuz.zaznai.ru/tw_files2/urls_5/28/d-27599/7z-docs/1.pdfLast Access: 5 March 2014.

[15] *Yohan, P.M. / Sasidhar, B / Basha, Sk. A. H. / Govardhan, A.* 2014. "Automatic Named Entity Identification and Classification using Heuristic Based Approach for Telugu." In: IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No. 1. January 2014. pp. 173-180. Available Online: http://www.ijcsi.org/papers/IJCSI-11-1-1-173-180.pdf Last Access: 3 March 2014.

[16] *Zeng, X. / Yang, J. / Zuo, L., Xu, Y.* 2013. "A Telephone Speech Corpus of China's Minority Languages for Automatic Language Identification." In: The 2013 AASRI Winter International Conference on Engineering and Technology (AASRI-WIET 2013). Saipan, USA. 28-29 December 2013. Available Online: http://www.atlantis-press.com/php/pub.php?publication=aasri-wiet-13&frame=http%3A//www.atlantis-press.com/php/paper-details.php%3Fid%3D10916 Last Access: 3 March 2014.

[17] *Zissman, M.A.* 1995. "Automatic Language Identification of Telephone Speech" In: The Lincoln Laboratory Journal. Volume 8, Number 2. 1995. pp. 115-144. Available Online: http://www.ll.mit.edu/publications/journal/pdf/vol08_no2/8.2.1.languageidentification.pdf Last Access: 3 March 2014.

Web Sources

[1] Language Identification (LID) Examples: http://www.cavar.me/damir/LID/ Last Access: 3 March 2014.
[2] Markov Model: Permanent link: http://en.wikipedia.org/w/index.php?title=Markov_model&oldid=593667257 Last Access: 3 February 2014.
[3] Language Identification in Multilingual Texts http://www.google.com/patents/US20120095748 Last Access: 5 March 2014.
[4] Lifehacker: http://lifehacker.com/388526/google-translate-automatically-detects-and-translates-languages Last Access 6 March 201
[5] Inside BlackBerry Help Blog: http://helpblog.blackberry.com/2013/02/customizing-blackberry-10-language-settings/ Last Access: 3 March 2014.
[6] Adaptxt: http://www.adaptxt.com/about-us/ Last Access: 3 March 2014.
[7] Apple's Daily Report: http://appledailyreport.com/apple-wants-your-mac-to-able-to-automatically-identify-languages/ Last Access: 3 March 2014.
[8] Automatic Language Identification for Dynamic Text Processing http://www.google.com/patents/US20090307584 Last Access: 5 March 2014.
[9] Compact Language Detection 2: https://code.google.com/p/cld2/ Last Access: 22 June 2015.
[10] Python: https://pypi.python.org/pypi/chromium_compact_language_detector/0.1.1 Last Access: 2 March 2014.
[11] Cognitive Science and Coding: http://cogscicoder.blogspot.de/2009/03/automatic-language-identification-using.html Last Access: 3 March 2014.
[12] Automatic Language Identification: http://research.microsoft.com/en-us/projects/lid/ Last Access: 3 March 2014.
[13] Code-switching: https://en.wikipedia.org/w/index.php?title=Code-switching&oldid=596564987 Last Access: 22 February 2014.

List of Abbreviations

DM	Data Mining
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LID	Automatic Language Identification
MT	Machine Translation
NE	Named Entity
NER	Named Entity Recognition
OCR	Optical Character Recognition
PR	Phone Recognizer
SLID	Spoken Language Identification
SR	Speech Recognition
WLID	Written Language Identification