Text segmentation for Language Identification in Greek Forums

Pavlina Fragkou
Technological Educational Institution of Athens (TEI-A),
Dept. of Informatics Systems,
Ag. Spyridonos, 12210, Egaleo, Athens, Greece.
pfragkou@teiath.gr

Abstract

In this paper, we examine the benefit of applying text segmentation methods to perform language identification in forums. The focus here is on forums containing a mixture of information written in Greek, English as well as Greeklish. Greeklish can be defined as the use of Latin alphabet for rendering Greek words with Latin characters. For the evaluation, a corpus was manually created by collecting web pages from Greek university forums and most specifically, pages containing information that combines Greek with English technical terminology and Greeklish. The evaluation using two well known text segmentation algorithms leads to the conclusion that despite the difficulty of the problem examined, text segmentation seems to be a promising solution.

1 Introduction

Language identification can be defined as the process of determining which natural language given content is in. Traditionally, identification of written language - as practiced for instance in library science - has relied on manually identifying frequent words and letters known to be characteristic of particular languages. More recently, computational approaches have been applied to the problem, by viewing language identification as a special case of text categorization, a Natural Language Processing approach that relies on a statistical method.

Greeklish, which comes from the combination of the words Greek and English, stands for the Greek language written using the Latin alphabet. The term Greeklish mainly refers to informal, ad-hoc practices of writing Greek text in environments where the use of the Greek alphabet is technically impossible or cumbersome, especially in electronic media. Greeklish was commonly used on the Internet when Greek people communicate by forum, e-mail, instant messaging and occasionally on SMS, mainly because older operating systems didn't have the ability to write in Greek, or in a Unicode form like UTF-8. Nowadays, most Greek language content appears in native Greek alphabet.

This paper is organized as follows: Section 2 provides information regarding related work, Section 3 provides a description of the method followed and the algorithms used, Section 4 provides evaluation metrics and obtained results, while Section 5 provides concluding remarks and future work.

2 Related Work

Language identification cannot be considered as a novel scientific area. Language identification of text has become increasingly important as large quantities of text are processed or filtered automatically for tasks such as information retrieval or machine translation. The problem has been researched long both in the text and in the speech domain.

Several works appear in the literature each of which dealing with a different type of problem. In Fereira da Silva and Pereira Lopes (2006a; 2006b), the authors examine language variation in two distinct problems: (a) identification of whether a text is written in Portuguese or in a Brazilian dialect; (b) small touristic advertisements on the web, addressing foreigners but using local language to name most local entities. Their approach uses the Quadratic Discrimination Score to decide which cluster (language) must be assigned to the document they want to classify. Space properties of the clusters are based on a document similarity measure which is calculated using character n-grams. The authors

Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, pages 23–29,
Hissar, Bulgaria, 13 September 2013.

conclude that discriminate elements depend on each specific context.

In Huges et al. (2006), the authors review a number of methods for enabling language identification in written language resources by focusing on cases such as: (a) the detection of the character encoding of a given document; (b) language identification for minority languages or unspecified language(s). They noticed that there is no one to one relation between a language and an encoding.

One of the most important papers on statistical language identification is presented by Dunning (1994). Dunning uses Markov Models to calculate the probability that a document originated from a given language model. In order to perform statistical language identification, a set of character level language models is prepared from training data during the first step. The second step involves the calculation of the probability that a document derives from one of the existing language models i.e., the probability that a String S occurs being from an alphabet X.

Another fundamental approach was proposed by Cavnar and Trenkle (1994). The authors calculated the N-gram profile of a document to be identified and compared it to language specific N-gram profiles. The language profile which has the smallest distance to their sample text N-gram profile indicates the language used.

A closely related work to ours is the one presented in Carter et al. (2011). In this work the authors introduce two semi-supervised priors to enhance performance at microblog post level: (i) blogger-based prior, using previous posts by the same blogger, and (ii) link-based prior, using the pages linked to from the post. The authors used the TextCat algorithm[1] and tested their models on five languages (Dutch, English, French, German, and Spanish), and a set of 1,000 tweets per language. Results showed that their priors improve accuracy but that there is still room for improvement.

Additionally, in the work presented in Winkelmolen and Mascardi (2011), the authors applied the well known Naive Bayes Classifier to perform language identification. The authors experimented on very short texts as well as on a corpus that they created from movie subtitles belonging to 22 different languages. To evaluate the impact of the use of different corpora, they compared the trigrams provided by TextCat with those obtained by their method. They concluded that a more accurate identification was obtained from their trigrams.

To the author's best knowledge, the only work that uses the notion of segmentation for the language identification task is presented in Zue and Hazen (1993), where a segment-based Automatic Language Identification (ALI) system has been developed. The system was designed around a formal probabilistic framework. The system incorporates different components which model the phonotactic, prosodic, and acoustic properties of the different languages used in the system. Practically the system investigates when an utterance should be segmented and how these segments can be characterized by a set of broad phonetic classes. The system was trained and tested using the OGI Multi-Language Telephone Speech Corpus. An overall system performance of 47.7% was achieved in identifying the language of test utterances.

The Greeklish phenomenon has been investigated in Chalamandaris et al. (2004), where the aim was to develop a module able to discriminate any Greeklish text from any other language. In order to surpass the problem of inconsistency in writing Greeklish, the authors made use of an alternative representation of every Greeklish word, namely a phonetic one. The performance of this module was tested with large multilingual corpora, where the initial Greek text was transliterated automatically according to four different sets of rules. The dataset consisted of: (a) public mailing lists; (b) private emails; (c) web pages in Greeklish written by more than 60 different persons in mixed Greeklish and English; (d) a large multilingual corpus whose content was varying from private and public emails, to web pages, newspapers, manuals, general documents, reports, and educational material for Greek high-school.

3 Method

In this paper we present an approach for language identification by using the technique of text segmentation. The text segmentation problem can be stated as follows: "*given a text which consists of several parts (each part corresponding to a different subject) it is required to find the boundaries between the parts*". In other words, the goal is to divide a text into homogeneous segments so that each segment corresponds to a particular subject while contiguous segments correspond to different subjects. In this manner, documents relevant to a query can be retrieved

[1] http://odur.let.rug.nl/~vannoord/ TextCat/

from a large database of unformatted (or loosely formatted) text. The problem appears often in information retrieval and text processing. One problem belonging to this category is language identification. To the author's best knowledge, it is the first time that text segmentation techniques are used to solve a language identification problem concerning text and not acoustic transcripts.

3.1 Text Segmentation Algorithms

The majority of text segmentation algorithms usually have as a starting point the calculation of the within segment similarity. This calculation is based on the assumption that parts of a text having similar vocabulary are likely to belong to a coherent topic segment. A significant difference between text segmentation methods is that some evaluate the similarity between all parts of a text, while others between adjacent parts. To penalize deviations from the expected segment length, several methods use the notion of "length model".

For our experiments we have chosen two well known topic change segmentation algorithms, the C99b implemented by Choi (2000; 2001) and the one proposed by Utiyama and Isahara (2001). Other algorithms presented in the literature proved to perform better in the Choi's benchmark corpus for the topic change segmentation task, such as the one implemented by Kehagias et al. (2004a; 2004b). However, the two selected algorithms benefit from the fact that they do not require training and their implementation is publicly available.

More specifically, Choi's C99b algorithm (2000; 2001) uses lexical cohesion as a mechanism to identify topic boundaries. This method uses the vector space model to projected words; sentences are then compared using the cosine similarity measure. Similarity values are used to build a similarity matrix. More recently, Choi improved C99b by using the Latent Semantic Analysis (LSA) achievements to reduce the size of the word vector space (Choi, 2001). Once the similarity matrix is calculated, an image ranking procedure is applied to obtain a rank matrix, which is a proportion of neighbors with lower values. The hypothesis is that LSA similarity values are more accurate than cosine ones.

Utiyama and Isahara (2001) propose a method that finds the optimal segmentation of a given text by defining a statistical model which calculates the probability of words belonging to a segment. Utiyama and Isahara's algorithm (2001) searches for segmentations with compact language models. The assumption here is that a segment is characterized by the distribution of words contained in it. Thus, different segments belonging to different topics have different word distributions. To find the maximum-probability segmentation, they calculate the minimum-cost segmentation by obtaining the minimum-cost path in a graph.

3.2 Corpus

As it was mentioned earlier, our work focuses on language identification on Greek forums. To the author's best knowledge, a publicly available corpus that examines the same problem does not appear in the literature. For this reason we created a corpus by collecting web pages taken from Greek university forums. The emphasis here was in collecting pages talking about a specific topic using Greek, Greeklish as well as English terminology. Thus, we collected 109 pages from the websites of the following institutions:

- University of Piraeus (28 pages)
- Technological Educational Institute of Athens (22 pages)
- National Technical University (NTUA) (3 pages)
- Aristotle University of Thessaloniki (69 pages)

Overall, our corpus consists of 17036 sentences, with the longest one containing 2582 characters. All the aforementioned web pages present strong variation in length as well as in the thematic category. In each of the aforementioned pages, an initial preprocessing was performed. Most specifically, sentences which were common or similar in each post, such as the post's theme (i.e, its subject), the date and time, the user login and other user's characteristics were removed. At a subsequent step, an annotation was performed where boundaries were placed at positions where the language used by the user changed.

Moreover, for English short function words such as prepositions, adverbs, adjectives as well as common verbs (e.g., the verbs "to be", "to have") in their variant forms were removed from the corpus. Additionally, stop word removal from a manually created list for Greek was performed. The stop list used for Greek is very similar to the one used for English. Stemming was also performed for English (i.e., substitution of a word by its root form) based on Porter's algorithm (Porter, 1980). Even though Greek is a heavily inflected language which means that a word may appear in many different forms, no further preprocessing (i.e., stemming and lemmatization) was performed for Greek.

Examination of the corpus led to interesting observations. A common observation is that users end their comments by the addition of a proverb as well as with facial expressions indicating their mood. However, in a number of cases, users writing their comment in Greek often finish their comment with an English proverb. On the contrary, users writing their comment in Greeklish often finish their comment with a Greek proverb. This makes the annotation (i.e., the choice of the boundary position) even harder because a boundary must be positioned before the proverb instead of being positioned at the end of user's post. Table 1 provides some examples of the different types combinations of comments and their corresponding proverbs written either using the same or using different languages for each pair comment-proverb of a post.

Another observation is the co-relation between the user's student identity and the language used. More specifically, we noticed that on the one hand, students belonging to technical departments choose to write their comments in Greek (but use a lot of technical terminology in English). On the other hand, the majority of law students write their comments in Greeklish. Users often start their comment in Greeklish and continue their post in Greek. Additionally, user's first word in the post corresponds to the login of the user to which they reply to. A frequent phenomenon is that users writing in Greek, also write English words using the Greek alphabet (for example, the word "thanks" is found as "θενκς"). Finally, emotional expressions are written in English (such as lol, evil, oops etc).

The purpose of the paper is the examination of whether a text segmentation algorithm is capable of identifying equivalent parts of text, where each part is written in different language. Since the topic in each web page of the corpus remains the same, the segmentation task here is to identify segment boundaries where each segment constitutes a text part written in Greek, or Greeklish, or English. Since text segmentation methods focus on sentence similarity or word distribution, the aim here is to identify where language changes according to the words appearing in a web page. In other corpora where language is common in all text parts, each segment corresponds to a different topic. In those contexts, change in word usage signals topic change and not language usage change.

4 Experiments

In this section we present the experiments we conducted to evaluate our method. We evaluate the application of a segmentation algorithm using the following three indices: Precision, Recall and Beeferman’s Pk metric (Beeferman et al., 1997; Beeferman et al., 1999). Those metrics are commonly used in the text segmentation problem. Precision and Recall metrics are properly defined for the segmentation task. More specifically, Precision is defined as “*the number of the estimated segment boundaries which are actual segment boundaries*” divided by “*the number of the estimated segment boundaries*”. Recall is defined as “*the number of the estimated segment boundaries which are actual segment boundaries*” divided by “*the number of the true segment boundaries*”. The F measure which combines the results of Precision and Recall is not used here, due to the fact that both Precision and Recall penalize equally segment boundaries that are “close” to the actual i.e., true boundaries with those that are less close to the true boundary. For that reason, Beeferman proposed an new metric named Pk which measures segmentation *in*accuracy; intuitively, Beeferman's Pk measures the proportion of “*sentences which are wrongly predicted to belong to different segments (while actually they belong to the same segment)*” or “*sentences which are wrongly predicted to belong to the same segment (while actually they belong in different segments)*” (for a precise definition of Beeferman Pk metric see (Beeferman et al., 1997; Beeferman et al., 1999)). A variation of Beeferman's Pk metric, named WindowDiff index has been proposed by Pevzner and Hearst (2002). The WindowDiff metric remedies several problems of Beeferman's Pk and is also used in our evaluation. More specifically, the WindowDiff metric penalizes false positives and near misses equally. Since Beeferman’s Pk and WindowDiff metrics measures segmentation *in*accuracy, low values of those metrics exhibit high performance of the algorithm examined.

Table 2 contains the obtained results after applying the two text segmentation algorithms in our corpus (where preprocessing has been performed as it was described in Section 3.2) using the four evaluation metrics described above.

Metric	Choi's algorithm	Utiyama & Isahara's algorithm
Precision	34.67%	23.88%
Recall	10.05%	62.35%
Pk	33.14%	46 %
WindowDiff	33.76%	62.9%

Table 2: Evaluation results

From the obtained results we can conclude that the segmentation accuracy differs from the one obtained in text segmentation corpora such as in Choi’s benchmark (Choi, 2001). Choi’s benchmark is used for text segmentation where the aim is to identify topic change. Reported results regarding Choi’s benchmark can be found in Kehagias et al. (2004a; 2004b). It is worth mentioning that the aforementioned text segmentation algorithms are usually examined in problems where the number of segments, as well the number of sentences per segment do not exhibit strong variations.

In order to understand the obtained results, we calculated the minimum, maximum, and average number of segments as well the number of sentences per segment and their standard deviation. Table 3 contains the aforementioned statistics.

	Number of segments per document	Number of minimum sentences per segment	Number of maximum sentences per segment
Mininum	1	1	2
Maximun	428	11	402
Average	38,69	1,14	28,43
Standard deviation	49,54	0,989	28,18

Table 3: Statistics regarding the corpus

From the information listed in Table 3 we can see that our corpus presents strong heterogeneity as far as the number of segments per document and the number of sentences per segment are concerned. In other words, text segmentation for this corpus constitutes a difficult task, justifying the relative low performance obtained by the text segmentation algorithms.

The performance of the text segmentation algorithms presents strong interest. This is due to the fact that in traditional text segmentation corpora Choi's algorithm achieves lower performance compared to the one obtained by Utiyama and Isahara's algorithm. However, in the current problem the exact opposite phenomenon occurs. A possible explanation may be that Utiyama and Isahara's algorithm performs global optimization of a local cost function contrary to the local optimization of global information performed by Choi's algorithm. It may be possible that local optimization of global information may be more suitable for the nature of our corpus.

5 Conclusions - Future Work

In this paper we presented an attempt to perform language identification on a corpus which combines information written in Greek, English, and Greeklish using text segmentation algorithms. The novelty of our approach lies in the nature of our corpus as well as the use of this type of algorithms for the language identification task. Despite the difficulty of problem, we believe that the use of text segmentation algorithms constitutes a promising solution which however deserves further examination.

We outlook several directions of future work. The first direction considers the investigation of alternative segmentation algorithms.

The second considers comparison of our approach with other language identification tools. Arguably, the best known tool is van Noord’s Text Cat, an implementation based on character n-gram sequences. Other well known implementations include BasisTech’s Rosette Language Identifier[2] and a number of web based language identification services such as those created by Xerox[3] and Ceglowski[4]. Language::Ident is another interesting language identification tool[5] implemented by Michael Piotrowski. The program already comes with trained language models and so far supports 26 languages. Supported identification methods are N-grams, common words, and affixes.

A third direction of future work considers a more sophisticated preprocessing of Greek using a POS tagger and a lemmatizer such as the one developed by Orphanos (Orphanos and Christodoulakis, 1999; Orphanos and Tsalidis, 1999). Finally we consider the examination of other Greek corpora.

2 http://www.basistech.com/language-identifier/

3 http://open.xerox.com/Services/LanguageIdentifier.

4 http://search.cpan.org/~mceglows/Language-Guess-0.01/

5 http://search.cpan.org/~mpiotr/Lingua-Ident-1.7/Ident.pm

References

D. Beeferman, A. Berger, and J. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34: 177-210.

D. Beeferman, A. Berger, and J. Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, 35-46.

S. Carter, E. Tsagkias, and W. Weerkamp. 2011. Semi-Supervised Priors for Microblog Language Identification. 2011. In *Dutch-Belgian Information Retrieval workshop (DIR 2011).*

W. B. Cavnar, and J.M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*.

A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos, and G. Carayannis. 2004. Bypassing Greeklish!. In *Proceedings of LREC 2004: 4th International Conference on Language Resources And Evaluation. Lisbon, Portugal*.

F.Y.Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 26-33.

F.Y.Y. Choi, P. Wiemer-Hastings, and J. Moore. 2001. Latent semantic analysis for text segmentation. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, 109-117.

T. Dunning. 1994. *Statistical Identification of Language*. New Mexico State University. Technical Report MCCS 94-273.

J. Ferreira da Silva, and G. Pereira Lopes. 2006. Identification of Document Language is Not yet a Completely Solved Problem. In *Proceeding of the CIMCA '06 Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*.

J. Ferreira da Silva, and G. Pereira Lopes. 2006. Identification of Document Language in Hard Contexts. In *SIGIR workshop on New Directions in Multilingual Information Access*, Seattle, USA.

B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. Mackinlay. 2006. Reconsidering language identification for written language resources. In *Proceeding of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 485-488.

A. Kehagias, A. Nicolaou A., P. Fragkou, and V. Petridis. 2004. Text Segmentation by Product Partition Models and Dynamic Programming. *Mathematical and Computer Modeling*, 39: 209-217.

A. Kehagias, P. Fragkou, and V. Petridis. 2004. A Dynamic Programming Algorithm for Linear Text Segmentation. *Journal of Int. Information Systems*, 23: 179-197.

G. Orphanos, and D. Christodoulakis, D. 1999. Part-of-speech disambiguation and unknown word guessing with decision trees. In *Proceedings of EACL'99*.

G. Orphanos, and C. Tsalidis 1999. Combining handcrafted and corpus-acquired lexical knowledge into a morphosyntactic tagger. In *Proceedings of the 2nd Research Colloquium for Computational Linguistics in United Kingdom (CLUK)*.

Porter, M.F. 1980. An algorithm for suffix stripping *Program*, 14(3) 130-137.

L. Pevzner, and M. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

M. Utiyama, and H. Isahara. 2001. A statistical model for domain - independent text segmentation. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 491-498.

F. Winkelmolen, and V. Mascardi. 2011. Statistical Language Identification of Short Texts. *In Proceedings of ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, vol. 1-Artificial Intelligence, 498-503.

W. Zue and T.J. Hazen. 1993. Automatic Language Identification Using a Segment-Based Approach. In *Proceedings Eurospeech 1993*, 1303-1306.

Example	Message	Proverb	Case	Web page source
1	" καταρχας ειναι παρα πολυ σημαντικο που επιτελους ειδαμε και μια λυση πρακτικου!!!Αλλα Δημητρα μηπως σου ειναι ευκολο να "ανεβασεις" και το πρακτικο? Θα ηταν πολυ χρησιμο για εμας που το χρωσταμε...... Χαμόγελο Ευχαριστω εκ των προτερων.	*Go confidently in the direction of your dreams.... Live the life you have imagined*	Message in Greek, proverb in English	http://www.dapnomikis-thess.gr/forum/index.php?topic=54.0
2	Lacrimosa το συγκεκριμενο μαθημα ειναι λιγο δυσκολο. προσωπικα σαν μαθημα το βρηκα αρεκετα ενδιαφερον, αλλα αυτο ειναι προσωπικη εκτιμηση.	«Δε συμφωνώ ούτε με μια λέξη από όλα όσα λες, αλλά θα υπερασπίζω, και με το τίμημα της ζωής μου ακόμα, το δικαίωμά σου ελεύθερα να λες αυτά που πρεσβεύεις» Βολταίρος"	Both message and proverb in Greek	http://www.dapnomikis-thess.gr/forum/index.php?topic=54.0
3	se mia apegnwsmeni prospatheia na diavasw to sugkekrimeno ma8ima k meta apo polu kopo mporw na dilwsw oti : auto to ma8ima einai APAISIO!!!	"Be the change you want to see in the world!"	Message in Greeklish, proverb in English	http://www.dapnomikis-thess.gr/forum/index.php?topic=31.0
4	Dhmhtra nomizw pws to xe h tzwrtzakakh to a tmhma!ylh den poly yparxei pantws klassiko sos einai h athinaikh dhmokratia k h sparth me th gortyna na akolouthei ligo pio pisw.....	"Einai h palia froura pou epistrefei me fora...TO KANAME TOTE,MPOROUME KAI TWRA!!"	Both message and proverb in Greeklish	http://www.dapnomikis-thess.gr/forum/index.php?topic=31.0
5	einai kati simeiwseis gia to mathima dne kserw kata poso tha boithisoun alla elpizw...	ΗΡΘΕ Η ΩΡΑ ΤΗΣ ΑΝΑΤΡΟΠΗΣ...1η ΞΑΝΑ Η ΔΑΠ ΤΗΣ ΝΟΜΙΚΗΣ...	Message in Greeklish, proverb in Greek	http://www.dapnomikis-thess.gr/forum/index.php?topic=13.0

Table 1: List of examples of users comments and their corresponding proverbs