Language Classification in Multilingual Documents

Gorkem Ozbek, Itamar Rosenn, Eric Yeh

Summary

We investigate the use of machine learning techniques for language classification in a multilingual setting. We consider three contexts in which this task may be performed: identifying the language of monolingual documents, identifying the language of individual tokens, and finally identifying and correctly classifying spans of monolingual text in multilingual documents. Our broad goal is to achieve the third task with optimal efficiency and accuracy. We pursue this goal by building and examining various morphologically-based classification methods that attempt to classify, within a document, the language identity of individual words whose language is not known, thereby approximating a potentially multilingual setting.

1 Introduction

In today's increasingly integrated and multilingual world, when developing natural language technology, one cannot always assume that the text one will encounter will be solely in one language. Maintaining this assumption and uniformly applying single-language-specific text-processing techniques may result in erroneous handling of terms in other languages. For example, in information retrieval tasks, single-language approaches can result in lower precision and recall scores: an English language preprocessor may not effectively capture subtleties relevant to languages such as Turkish, where linguistic structure differs drastically even at the word level.

Traditionally, the task of language identification has been applied in settings where the entire document is assumed to be in a single language. However, with the advent of the World Wide Web, instances of mixed-language documents have become more prevalent. For example, the online edition of the German magazine *Der Spiegel* uses a sidebar of text written in English [see http://www.spiegel.de/]. In addition, phrases are often appropriated from one language into the context of another, such as the English phrase "sexiest man alive", which appears in a *Der Spiegel* article about George Clooney. The reality of multilingual text introduces the task of identifying when the language of a span of text in a document differs from the primary language of that document; to our knowledge, this multilingual identification task has not yet been sufficiently explored. At one extreme, this problem can be reduced to identifying likely language of origin for a single observed token in a possibly multilingual setting; this approach motivates our present work.

2 Feature Engineering

Previous Work

A large body of previous language identification work has focused on statistical classifiers primarily operating over character-level, non-linguistically motivated features, such as n-gram character models [1, 2, 3]. These methods generally perform well only after a certain number of characters has been seen by the classifier [3]. However, as mentioned previously, these efforts have focused on instances where the text to be classified is considered to be of a single language. Furthermore, in a possibly multilingual setting, these methods would also be unreliable because a single-language portion of the text may be too short to contain sufficient characters for the character-based classification methods.

Our Approach

The character-based approach achieves near-perfect classification accuracy within about 100-200 characters, without exploiting any features that are idiosyncratic to linguistic characteristics of different languages. In the hopes of achieving high-accuracy classification within a short textual span, as is needed for multi-lingual classification, we would like to develop a classification system based on linguistic factors that differ among languages, using light-weight features that can accrue significant statistics within short word spans of text. Using morphological features is an obvious choice given these desired criteria.

Linguistic theory teaches us that word tokens of a given language are comprised of smaller elements called morphemes [see figure 6], which are the smallest components of a language that carry semantic value. Thus, a language model may be estimated using morphemes as lexical units, instead of the words in which they appear.

We restrict our inquiry to four languages: English, Finnish, German, and Turkish. Our approach is to construct a feature set for each of these languages by obtaining a morpheme-based language model estimated from a unilingual corpus of the language. The prima facie difficulty with this goal is that construction of a broad, linguistically informed morphological lexicon for a given language requires a considerable amount of work by trained experts. Thus, for our purposes and for the general task of engineering a successful morpheme-based language classifier, obtaining such a lexicon is prohibitively costly, particularly in the context of many possible languages or highly evolving languages.

An alternative approach has recently been explored in the literature: designing generative, minimally supervised algorithms that attempt to automatically discover morphemes in a corpus. We rely on one such algorithm to construct individual feature sets – morpheme "language models" – for each of our languages in an efficient and unsupervised manner. The algorithm has been developed by Mathias Creutz, who demonstrates its high accuracy for various languages [5].

Creutz's Algorithm

The algorithm uses segmentation and is formulated within a probabilistic framework. Two features of the algorithm enable it to handle various morphologically disparate languages, making it especially appropriate for our purposes:
First, the algorithm treats words as arbitrary sequences of alternating stems and affixes, making it more flexible with respect to languages having different levels of inflection. Within our own framework, Turkish and Finnish are far more inflective than German and English. Second, the algorithm considers sequential dependencies between functional categories of morphemes (an approach known as *morphotactics*), which increases its accuracy in an arbitrary multilingual setting and provides us with a richer feature set than a simple collection of morphemes. The algorithm assumes that morphemes fall into two categories: stems and affixes, and the latter category is divided into prefixes and suffixes.

Creutz's algorithm uses an HMM to model morpheme sequences, without assuming prior knowledge of the segments (morphemes) themselves, nor of their individual functional categories. The algorithm performs a baseline segmentation, then estimates probabilities of observing a particular morph given its category, and the probability of a transition from one morph category to another, using EM.

Our Features

The main advantages to Creutz's approach for our purposes are that it allows us to extract an estimated morphological and morphotactic feature set for each language, without any supervision or prior linguistic knowledge. We develop an analyzer that uses Creutz's algorithm to identify morphological units and their appropriate functional categories within each language-specific training document.

3 Methodology

Corpora

Each of our models, discussed below, uses documents made available for Morpho Challenge 2007 [4]. The site provides sets of unilingual documents in English, Finnish, German, and Turkish, which we use for train and test corpora in each of our language identification procedures.

Baseline

In order to establish a baseline for the language identification task, we implement a Naïve Bayes classifier using n-gram (i.e. unigram, bigram and trigram) character models. In accordance with current state of the art language classification systems, we choose Naïve Bayes for our Baseline and morphological models. We also use Laplacian noise modeling throughout. For the baseline, our n-gram character models are built from word lists constructed from English, Finnish, German, and Turkish training corpora. Each word list contains unique words (i.e. word types) encountered in the corpus for one of the four languages, along with the token frequencies for the words. The classifier is then trained with these character models for each of the four languages.

Test documents in each language, similar to the training documents, are also obtained from Morpho Challenge. Accuracy vs. number of corpus characters read is measured to establish the success of the baseline approach for varying amounts of data. The tokenized version of the corpus, in the form of a wordlist, is also used to examine baseline performance with respect to classifying individual tokens.

Morph Classification

As a first attempt to improve upon the character n-gram approach of the baseline and obtain a classification system appropriate for multilingual settings, we obtain a morpheme feature set for each language by applying the Creutz algorithm to each individual-language training document. We limit our feature set only to a morpheme count for each language, which serves as a simple morphological "language model" for the language. The morpheme count list contains a list of the unique morphemes found in the document, along the frequency of each morpheme. We then implement a Naïve Bayes classifier using our morpheme counts as features. In the testing phase, each language-specific test document is first analyzed using the Creutz procedure in order to identify best guesses of the correct morpheme segmentation of each word in the test document. We then apply our Naïve-Bayes classifier, along with the morpheme feature list for each language, to the segmented test document. The classifier identifies the language of each word in the document according to its maximum likelihood classification. Note that the classifier does not rely on the assumption that the entire document or even sequences of words of the document are all in one language. Therefore, although the test documents are all unilingual, the classifier itself performs exactly as it would in a multilingual setting (restricted to our candidate languages), even in the extreme case where the identity of each word was entirely independent of the identities of other words in close proximity to it.

Morph + Morphotactics Classification

In addition to the simple morpheme feature classifier, we develop a classifier that takes advantage of the morphotactic information that the Creutz analyzer provides. For this classifier, we also obtain a feature set of 3-gram morpheme category sequences for each language. This set is obtained by examining the language-specific training documents after they have been Creutz-analyzed, and creating a count of all 3-gram morphological category sequences within the document. Thus, for each of our four candidate languages, the Creutz analyzer provides us with a morphological "language model" based on morpheme frequency count and morphotactic sequence frequency. We then implement a Naïve Bayes classifier to use the original morpheme-count as a feature set together with our new 3-gram morphotactic feature set, for each of our four languages. As before, in the testing phase, each language-specific test document is first analyzed using the Creutz procedure in order to discover morphemes and their functional categories (prefix, suffix, stem). Then, the Naïve-Bayes classifier, along with the morpheme and morphotactic language features, is applied to the tagged document. The classifier identifies each word according to its maximum likelihood language identity, using both the morphological and morphotactic features of the word.

Filtering

In an attempt to reduce computation time, as well as to improve the accuracy of our morphological "language models", we also applied filtering to the training documents. For each language-specific training document, we obtained a new document that filtered out words occurring less than 5, 100, and 1000 times. We performed new rounds of testing classification for each of these filter levels to see if efficiency and performance are increased.

4 Results and Discussion

When attempting to identify the language of a single document, the performance of the baseline character n-gram model increases as the number of characters observed increases (see Figure 1), resembling performance curves seen in previous studies. Baseline identification of single tokens is far less successful, as suggested by the low accuracies observed in figure 1 when 50 or fewer characters have been seen.

The results of the Naïve Bayes classification task using our morphological feature sets, in comparison to the results of baseline n-gram Naïve Bayes classification, are shown in Figures 2-5. The first notable observation is that in unfiltered settings, the morpheme-feature classifier (Morph) and the morpheme-feature plus morphotactic-feature classifier (Morph + Mts) achieve at least a slight improvement in performance over the baseline. Our classifiers achieve the most success in classifying Turkish words: Morph does slightly better than the baseline, which is roughly at 80%, and Morph + Mts achieves very close to 100% accuracy. Filtering does not greatly alter these results, suggesting that Creutz's method gives a reliable morphological-morphotactic model of Turkish.

With respect to Finnish and German, unfiltered classification using Morph and Morph + Mts achieves slightly higher accuracy than the baseline, but not near the levels of success seen with Turkish. Furthermore, adding morphotactic features does not increase performance beyond classification using only morphemes. Filtering seems to reduce accuracy somewhat with both of these languages, suggesting that the full training documents supplied to Creutz's algorithm yield the most information in terms of a morphological language model.

The results for English are the most discouraging. As Figure 2 shows, neither Morph nor Morph + Mts classification achieves accuracy scores as high as baseline, with the worst performance given by Morph + Mts, with only 40% accuracy. Once filtering is introduced, Morph + Mts accuracy gradually climbs up toward the accuracy of the other two classification systems. This suggests that the 3-gram morphotactic model is particularly unsuitable for English morphotactics, since at a high level of filtering the richness of Morphotactic feature set disappears, and the system works similar to simple Morph classification.

5 Future Work

Our success on the Turkish test document suggests that the true potential for better performance stems from constructing an appropriate *morphotactic* model for a language. If the near-perfect accuracy of Morph + Mts in classifying Turkish could be replicated for other languages, then morphological and morphotactic feature sets would be more accurate than character n-grams or other current methods, such as "most common substring", at identifying the language of a document. Furthermore, because such success is achieved at the word level, these methods would be appropriate for classifying the language of word sequences of any length within a multilingual document, an achievement that n-gram character classification cannot replicate. However, we are as yet unable to develop accurate classifiers for our other candidate languages. This may be because the 3-gram morphotactic model is appropriate for the morphology of Turkish, but not for other languages; a different n-gram model must be used.

Therefore, the first and crucial avenue for future work is to examine and test this hypothesis. Without a reliable and accurate n-gram model of morphotactics, our proposed morphological approach does not have much to offer in comparison to the "character based" state of the art. However, if future work does identify successful n-gram morphotactic models for other languages, other methods can be applied to attempt to increase performance. For example, rather than Naïve Bayes classification, one can use SVM for the task, which turned out to be too computationally costly for the scope of our work.

The lightweight, informationally dense attributes of morphological features suggest that morphological approaches to text classification may be extremely promising. However, with respect to the language classification task, the reality of this potential is as yet uncertain.

References

.

[1] N-Gram-Based Text Categorization, William B. Cavnar, John M. Trenkle. http://citeseer.ist.psu.edu/68861.html.

[2] Statistical Identifaction of Language, Ted Dunning, Computing Research Lab, New Mexico State University.

[3] Applying Monte Carlo Techniques to Language Identification, Arjen Poutsma, SmartHaven, Amsterdam.

[4] Unsupervised Morpheme Analysis, Morpho Challenge 2007, Kurimo, Creutz, Varjokallino. http://www.cis.hut.fi/morphochallenge2007.

[5] Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency, Matthias Creutz, Proceedings from ACL-03, 280-287, Sapporo, Japan. July 2003.

Figures

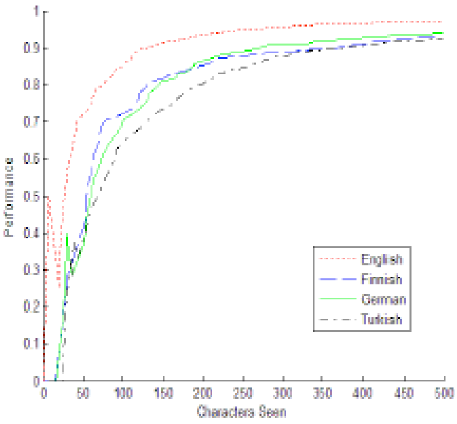

Figure 1

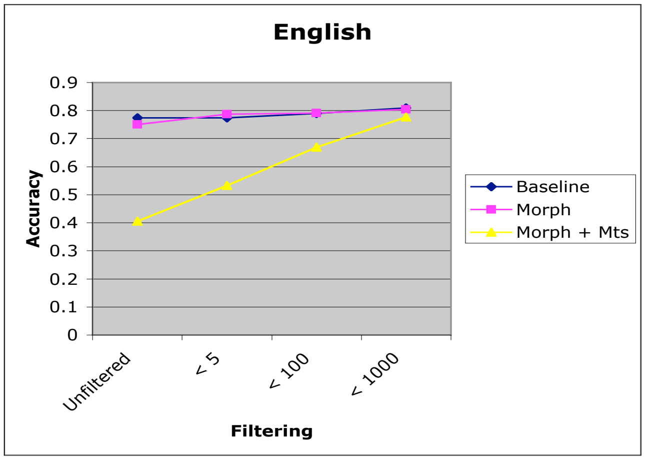

Figure 2

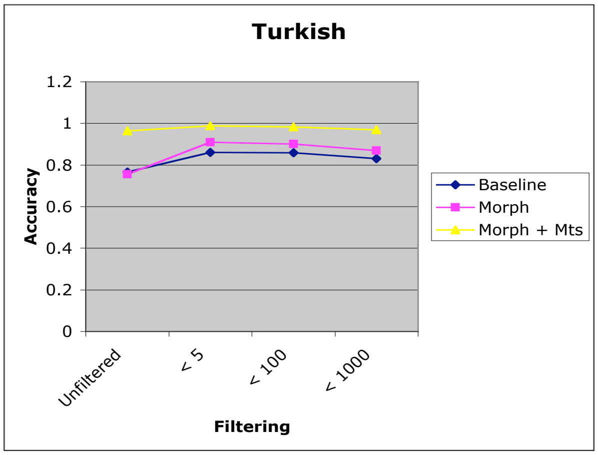

Figure 3

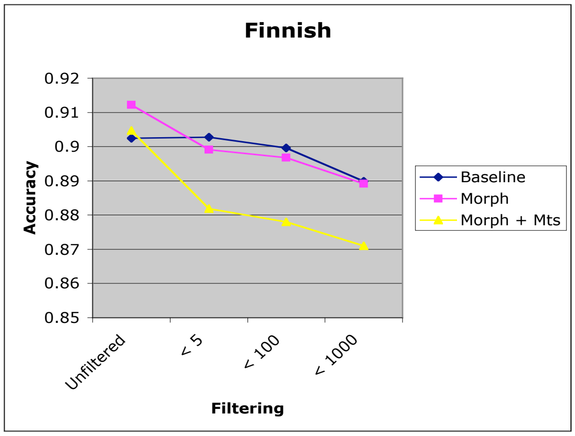

Figure 4

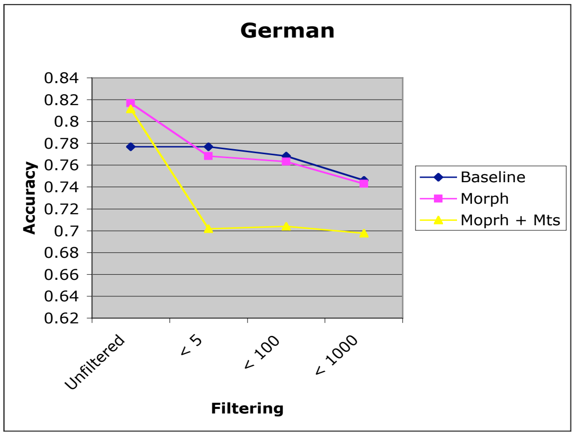

Figure 5

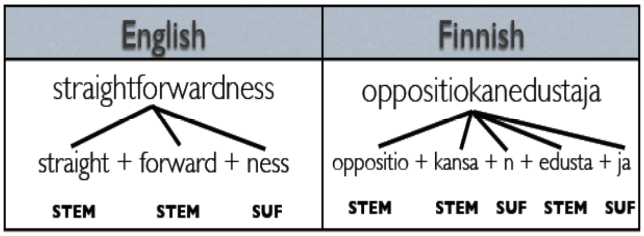

Figure 6