Authorship Verification with Entity Coherence and Other Rich Linguistic Features

Notebook for PAN at CLEF 2013

Vanessa Wei Feng and Graeme Hirst

[1] University of Toronto
weifeng@cs.toronto.edu
[2] University of Toronto
gh@cs.toronto.edu

Abstract We adopt Koppel et al.'s unmasking approach [5] as the major framework of our authorship verification system. We enrich Koppel et al.'s original word frequency features with a novel set of coherence features, derived from our earlier work [2], together with a full set of stylometric features. For texts written in languages other than English, some stylometric features are unavailable due to the lack of appropriate NLP tools, and their coherence features are derived from their translations produced by Google Translate service. Evaluated on the training corpus, we achieve an overall accuracy of 65.7%: 100.0% for both English and Spanish texts, while only 40% for Greek texts; evaluated on the test corpus, we achieve an overall accuracy of 68.2%, and roughly the same performance across three languages.

1 Introduction

Authorship verification, a sub-task of authorship identification, deals with the demand of identifying whether a two documents are written by the same author or not. Typically, a set of documents which are known to be written by the author of interest is given, and an authorship verification system needs to determine whether a given unknown document is written by this author.

We follow the unmasking approach described by Koppel et al. [5], which is designed specifically for the task of authorship verification, as the major framework of our authorship verification system. However, rather than using word-frequency features as Koppel et al. did, we unmask a pair of documents by using a set of linguistic features, including our own coherence features and well-established stylometric features. Moreover, as sophisticated coreference resolution tools, available only for English, are required for extracting our novel coherence features, we first translate non-English texts into English, and from the translations, we extract the coherence features.

2 Methodology

2.1 Unmasking

Unmasking [5] is a technique developed specifically for the task of authorship verification. Its underlying idea is that, if two documents were written by the same author, then any features a classifier finds that (spuriously) discriminate their authorship must be weak and few in number. On the other hand, if the texts were written by different authors, then many more features will support their (correct) discrimination.

Our modified unmasking approach is as follows: From all known documents in the training corpus, we extract: (1) $\mathscr{S}_{same}$, the set of pairs of documents written by same authors; (2) $\mathscr{S}_{diff}$, the set of pairs of documents written by different authors. For each document pair $\langle d_i, d_j \rangle$, written by author A_i and A_j respectively, the two documents, d_i and d_j, are segmented into equal-sized and non-overlapping small chunks. If the number of chunks of either document is less than 5, an up-sampling is first performed to pad the size to at least 5. If the number of segmented chunks of each document is unequal, a balanced sample is obtained by randomly discarding surplus chunks in the larger set. The following procedure is repeated N times.

1. A *weak classifier* with a set of unmasking features is trained to label each chunk as being from document d_i or d_j. The sampling is repeated five times, and the averaged leave-one-out cross-validation accuracy is reported to represent the discrimination performance.
2. Remove the top 3 most discriminating features of the weak classifier, and repeat Step 1 using the remaining features.

Pair $\langle d_i, d_j \rangle$ is therefore unmasked by the degradation of the cross-validation accuracy after each iteration of feature removal. Such degradation of accuracies is encoded by a numeric vector using the original representation in [5]. Finally, a binary classifier, called a *meta-classifier*, is trained to differentiate *same* degradation curves ($A_i = A_j$) vs. *different* degradation curves ($A_i \neq A_j$).

2.2 Enhanced features for unmasking

Our important extension to Koppel et al.'s unmasking approach is the enhancement of the features used in building the weak classifiers for unmasking the degradation curves (Step 1 in Section 2.1). Although Koppel et al.'s word frequency feature set achieved competitive performance for verifying novel-length texts, we found that word frequency features are too unreliable for much shorter texts. Therefore, we use a more comprehensive feature set, including various rich linguistic features, which can be partitioned into two categories:

Coherence features As we have shown in earlier work [2], coherence features, based on the local entity transition patterns derived from Barzilay and Lapata's entity grids [1], can be useful discourse-level authorship features. The entity grid model is based on the assumption that a text naturally makes repeated reference to the elements of a set of entities that are central to its topic. It represents local coherence as a sequence of transitions, from one sentence to the next, in the grammatical role of these references. For example, an entity may be mentioned in the subject of one sentence and then in the object of the next — or not at all in the next. These coherence features are encoded as a vector consisting of the relative proportions of a set of predefined entity transition patterns. As in our earlier work, [2], we use Reconcile-1.0[3] to extract entities in texts and resolve coreferences.

Since we are not aware of any available coreference resolution systems for languages other than English, it is nontrivial to extract coherence features for Spanish and Greek texts. However, we believe that, while surface-form authorship features, such as word usage, are generally obfuscated in the process of being translated to English, coherence features, as a kind of discourse-level feature, are relatively well preserved. Therefore, we first use the Google Translate service[4] to obtain their English translations, then perform the entity extraction on these translations.

Stylometric features In addition, we use a set of well-established stylometric features, the majority of which are from our earlier work [4], including (1) *Basic features*: the average sentence length (in words), the average word length (in characters), lexical density, word length distribution; (2) *Lexical features*: frequencies of function words, hapax legomena, and hapax dislegomena; (3) *Character features*: frequencies of various characters; (4) *Syntactic features*: part-of-speech entropy, frequencies of part-of-speech bigrams, and frequencies of syntactic production rules. English texts are parsed by the Stanford CoreNLP toolkit[5], and Spanish texts are parsed by the FreeLing toolkit[6]. We use the AUEB Greek part-of-speech tagger[7] to obtain the part-of-speech tags for Greek texts (full syntactic parsing is not available for Greek texts).

The total number of features is 538 for English, 568 for Greek, and 399 for Spanish texts.

2.3 Parameter configurations

There are a few parameters that can be adjusted in our approach. We tested several parameter combinations, and decided to use the following configurations which achieved the best performance on the training data.

Chunk sizes: English and Spanish texts are chunked into 200 words, while Greek texts are chunked into 100 words. In both cases, any leftover in a document is discarded.

Unmasking iterations: The unmasking procedures in Section 2.1 are repeated N times in order to obtain the degradation curve. For English texts, $N = 20$, while for both Greek and Spanish texts, $N = 10$.

Classifiers: For the *weak classifier* used in unmasking degradation curves, we use the linear-kernel LibSVM classifier, implemented by Weka 3.7.7 [3], and the top 3 most discriminating features to be removed in each unmasking iteration are chosen as the

[3] http://www.cs.utah.edu/nlp/reconcile/

[4] http://translate.google.com/

[5] http://nlp.stanford.edu/software/corenlp.shtml

[6] http://nlp.lsi.upc.edu/freeling/

[7] http://nlp.cs.aueb.gr/software.html

3 features with the highest absolute-value weight in the linear kernel. For the *meta-classifier* to differentiate *same* degradation curves vs. *different* degradation curves, we use the Bagging classifier offered by the Weka package. All these classifiers use the default parameters setting.

3 Experiments and Result

We use only the training corpus released by PAN 2013 Authorship Identification task (10 English cases, 20 Greek cases, and 5 Spanish cases) and no other complementary materials. A separate model is built for each language.

In training, for a particular language, we use all *known* documents written in this language in the training corpus to unmask *same* and *different* degradation curves. Since there are typically many more *different* degradation curves than *same* ones, we sampled at most 500 *same* curves and 1000 *different* curves.

For evaluation, we unmask each *unknown* document in the corpus, against the given *known* documents of the same case, to get a degradation curve corresponding to this unmasking. We then use the trained *meta-classifier* to classify the resulting degradation curve as *same* or *different*, and thus determine the final answer.

We produce a yes/no answer for all cases, and obtained an overall accuracy of 65.7% for all 35 cases in the corpus. Specifically, for each language, we obtained 100.0% for both English and Spanish, while only 40% for Greek texts. The final evaluation result of our system is an overall accuracy of 68.2%, with roughly the same performance across three languages.

Because the unmasking features in our authorship verification system are designed specifically for English texts, by using well-established stylometric features and our novel coherence features, we expect that by better understanding a particular language, especially those with fewer available NLP tools, more effective unmasking features can be designed to achieve competitive performance on non-English languages as well.

References

1. Barzilay, R., Lapata, M.: Modeling local coherence: an entity-based approach. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 141–148. ACL 2005, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
2. Feng, V.W., Hirst, G.: Patterns of local discourse coherence as a feature for authorship attribution. Literary and Linguistic Computing (2013)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1) (2009)
4. Hirst, G., Feng, V.W.: Changes in style in authors with Alzheimer's disease. English Studies 93(3), 357–370 (2012)
5. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. The Journal of Machine Learning Research 8, 1261–1276 (2007)