Towards the automatic detection of the source language of a literary translation

Gerard Lynch[1,2] & Carl Vogel[1,2,3]
(1) School of Computer Science and Statistics[1], Trinity College, Dublin 2, Ireland
(2) Centre For Next Generation Localisation[2],
(3) Centre for Computing and Language Studies[3]
gplynch@scss.tcd.ie, vogel@scss.tcd.ie

ABSTRACT
Experiments on the detection of the source language of literary translations are described. Two feature types are exploited, n-gram based features and document-level statistics. Cross-validation results on a corpus of twenty 19th-century texts including translations from Russian, French, German and texts written in English are promising: single feature classifiers yield significant gains on the baseline, although classifiers containing a combination of feature types outperform these, bringing L1 detection accuracy to ∼80% using ten-fold training set cross validation. Average test set results are slightly lower but still comparable to the cross-validation results. Relative frequencies of a number of salient features are studied, including several English contractions (*I'll*, *that's*, etc.) and uncontracted forms; we articulate hypotheses, anchored in source languages, towards explaining differences.

KEYWORDS: Computational stylometry, translation studies, source language detection, text classification.

1 Introduction

This study focuses on experimentation towards the detection of source language influence in literary translations into English from the late nineteeth and early twentieth centuries. We assembled a corpus of novels from this period, consisting of fifteen translations, five each from Russian, German and French, and five works written originally in English.[1] We carry out cross-validation experiments to determine robust features which identify the L1 of the texts.

We use document-level metrics such as sentence length and readability scores together with n-gram features such as the frequency of sequences of POS tags and closed-class words, features which are not directly related to the topics and themes contained within the texts. The present experiments attempt to correctly attribute the L1 of texts; this entails correctly classifying a text as translated or not. In order to minimize the effect of authorial or translatorial style in this study, we have not selected more than one work by the same author or translator.

Four criteria for corpus selection were as follows. Firstly, text should be available in an machine-readable format and in the public domain. Secondly, from the previous point, this dictates that text will most likely stem from prior to the early twentieth century, due to US copyright law. Thirdly, each text should have a unique author and in the case of translations, translator, i.e. no repeated authors or translators. Finally, text should be of sufficient length, at least two hundred kilobytes in size, i.e. preferably a novel or novella. In many cases, particular translators had translated numerous works by a single author and indeed also occasionally by several authors. Thus, it was necessary to choose texts so that each author and translator remained unique.[2] Table 1 lists the texts, all sourced from Project Gutenberg.[3]

Section 2 describes prior research. Section 3 explains our own experimental methodology. Section 4 details the results of experiments carried out on detection of the L1 of a corpus of texts translated from Russian, German and French together with texts in original English.

2 Previous research

Recent work in computational and corpus linguistics has focused on the analysis of comparable corpora[4] of translated and original text (see Kilgarriff (2001) on comparability assessment).

Olohan (2001) identifies patterns in *optional* usage in comparable English corpora, citing examples such as the use of complementizer *that*[5] as discriminatory between translations and original texts, with translations containing a higher incidence of the complementizer construction, using t-tests to identify features which differ with statistical significance. This method depends on selective expert hypotheses about which features discriminate texts of L2 English.

Guthrie, Guthrie, Allison, and Wilks (2007) evaluated their general method of ranked feature differences on the problem of assessing whether translations of L1 Chinese newspaper texts

[1]We will henceforth refer to the source language of the text as the L1.

[2]This was more complicated for Russian, for example, with the translator Constance Garnett having translated works by Dosteyevsky and Turgenev, amongst others, resulting in the bypassing of a title of such repute as *Anna Karenina* for the less well-known novella *The Cossacks* by Tolstoy, due to the fact that Garnett was already represented as the sole available translator of Turgenev.

[3]www.gutenberg.org, last verified August 2012

[4]These are corpora of the same style and genre, containing a proportional amount of translated and original text.

[5]*He said **that** he was ill* vs. *he said he was ill* vs. *the illness **that** killed him was swift*: the first contains a complementizer-that and the last, a relativizer-that.

Title	Author	Source	Pub.	Translator	T.pub.
Great Expectations	Charles Dickens	English	1861	n/a	n/a
The Picture of Dorian Gray	Oscar Wilde	English	1891	n/a	n/a
Jude the Obscure	Thomas Hardy	English	1895	n/a	n/a
Treasure Island	R.L Stevenson	English	1883	n/a	n/a
Middlemarch	George Eliot(M. Evans)	English	1874	n/a	n/a
The Idiot	Fyodor Dostoyevsky	Russian	1869	Eva Martin	1915
The Man Who Was Afraid	Maxim Gorky	Russian	1899	Hermann Bernstein	1901
Fathers and Children	Ivan Turgenev	Russian	1862	Constance Garnett	1917
The Cossacks	Leo Tolstoy	Russian	1863	Louise and Alymer Maude	n/a
A Man of our Time	Mikhail Lermontov	Russian	1841	J.H Wisdom/M. Murray	1917
The Count of Monte Cristo	Alexandre Dumas	French	1844	Anon	1846
Madame Bovary	Gustave Flaubert	French	1857	Eleanor Marx-Aveling	1898
Fr Goriot	Honoré de Balzac	French	1853	Ellen Marriage	1901
The Hunchback of Notre Dame	Victor Hugo	French	1831	Isabel F. Hapgood	1888
Around the World in Eighty Days	Jules Verne	French	1873	George M. Towle	1873
Effi Briest	Theodor Fontane	German	1896	William A. Cooper	1914
The Merchant of Berlin	Luise Mühlbach	German	1896	Amory Coffin	1910
Venus in Furs	Leopold V. Sacher-Masoch	German	1870	Fernanda Savage	1921
The Rider on the White Horse	Theodor Storm	German	1888	Margarete Münsterberg	1917
Debit and Credit	Gustave Freytag	German	1855	Georgiana Harcourt	1857

Table 1: Corpus of texts

in L2 English could be identified in a set of L1 English news texts (35K words of Chinese translated to English and 50K words of English L1). Features focused on what we consider document-level features (ie. percentages of words in major grammatical categories, ratios of frequencies between grammatical categories, most frequent POS trigrams and bigrams, etc). Feature vectors are constructed to represent each text and its relative complement, with separate vectors for the percentages and ratios and the ranked frequency features. A derived vector records a score based on the Spearman rank correlation coefficient between the text and its complement for each of the sorts of frequency list. Two texts are compared by calculating the average differences between feature vectors and adjusting with the derived scores from the ranked frequency list differences. In each configuration of the evaluation, one translation was presented without annotation along with 50 L1 English texts, texts separated as 1000 word samples. The translated text appeared in the top three ranked positions, representing greatest anomaly, in 93% of experiments, and in the top ten positions in 100%. Our own work is comparable in the features analyzed, but uses a classification approach that labels the source language of each text. rather than giving each text a rank in its evidence of being a translation.

Baroni and Bernardini (2006) explore whether machine learning methods may discover translated texts more robustly than people. They investigate a corpus of translated and original articles from the Italian current affairs publication *Limes* using machine learning methods similar to this study, and report high degrees (≥85%) of classification accuracy between the two categories, identifying features such as clitic pronouns and adverbial forms as distinguishing features between the translated and original sections of the corpus. Only one of ten humans in an evaluation exercise outperformed the ML system on all measures.

In previous work on detecting the L1 of translations using computational methods similar to those used in our study, van Halteren (2008) examined source language markers in the Europarl corpus, obtaining high accuracy in L1 detection($\geq 90\%$) across translations and original texts in multiple European languages, using features such as n-grams of words and POS tags alone. Frequent n-grams included *framework conditions* in the English corpus translated from German, and the n-gram *certain number*, which occurred to a higher extent in the translations from French and Spanish than the German, Italian and Dutch texts. However more recent work by Ilisei, Inkpen, Corpas Pastor, and Mitkov (2010) on stylistics of translations in Spanish technical and medical translations motivated the use of features other than simple n-grams in our work. These comprise of a number of statistics calculated on a document level, features which are listed in Table 2 We also broaden the scope of our study to literary translations, which we believe will pose a greater challenge to the task of L1 detection than the Europarl corpus which is more homogenous in style and comprising only parliamentary transcriptions.

3 Methods

We use Weka (Hall et al. (2009)) as a machine-learning toolkit, coupled with the TagHelperTools package (Dönmez et al. (2005)) which provides support for processing natural language data in Weka. We calculated values for the document-level features (Table 2) using our own script which relies on the TreeTagger POS tagger (Schmid (1994)) for the tagging of text. Within Weka, we use the Ranker algorithm coupled with the χ^2 metric to rank the features by classification power. These rankings are then listed in Tables 4 and 5 For the experiments, we used the Weka SMO classifier, which is an implementation of a Support Vector Machine (SVM) classifier, the Simple Logistic classifier and the Naive Bayes classifier.

Feature	**Description**	**Feature**	**Ratio Description**
Avgsent	Average sentence length	*Typetoken*	word types : total words
Avgwordlength	Average word length	*Numratio*	numerals : total words
CLI	Readability metric	*Fverbratio*	finite verbs : total words
ARI	Readability metric	*Prepratio*	prepositions : total words
		Conjratio	conjunctions : total words
		Infoload	open-class words : total words
		dmarkratio	discourse markers : total words
		Nounratio	nouns : total words
		Grammlex	open-class words : closed-class words
		simplecomplex	simple sentences : complex sentences
		Pnounratio	pronouns : total words
		lexrichness	lemmas : total words
		simplecomplex	simple sentences : complex sentences
		simpletotal	simple sentences : total sentences
		complextotal	complex sentences : total sentences

Table 2: Document-level features

3.1 Features and corpus treatment

We use 19 document-level features in this analysis listed in Table 2. Two readability indices, the Automated Readability Index, (Smith and Senter (1967)) and the Coleman-Liau Index, (Coleman and Liau (1975)) were used. We also use n-gram features such as word-unigrams and part-of-speech bigrams. We remove any proper nouns in the word n-gram feature list, as any character or place-names could unambiguously distinguish a text. We do this after the word unigram features are calculated. The frequency of untranslated terms and titles from the source language, place-names or names of characters could prove highly useful in predicting the source language of a text, however these we would expect to vary depending on the topics and themes within the text.[6] We therefore focus on highly frequent n-grams, such as prepositions, determiners and frequent verb forms, which we expect to be more robust predictors of the source language of a text.

To balance the corpus for each source language, we selected a random contiguous section of 200 kb of text from each work in the study and divided this up into 20 chunks of 10 kb each. This results in 100 textual segments per source language. Corpus balancing is important when using metrics such as type-token ratio which vary with relation to text length. We trained on 360 of the text chunks retained a separate set of 40 chunks from the corpus divided evenly across the four languages and works[7] for test purposes.

3.2 Classification tasks

The features described are used to label texts written in English according to their source language. This is more refined than labelling a text as translated or not since we want to know not just whether it is a translation, but further, if it is a translation, the identity of its L1.

4 Experiments

4.1 Single and combined feature sets

Using the SVM classifier we obtain 66% accuracy using ten-fold cross validation for the four categories using our 19 document level statistics only. The Naive Bayes classifier performs worse, giving 54% accuracy. The Simple Logistic classifier performs the best here, with 68% accuracy. Given that the baseline for this task is 25%, 68% can be deemed a promising result, although the results are lower for the hold-out set, at 62% for the Simple Logistic classifier. The merged feature sets produce better results in this task, the best performing combination being Run 13, which consists of the top 50 features as ranked by the chi-squared metric in Weka taken from: (i) the top one hundred POS bigrams; (ii) all 19 document-level features; (iii) the top fifteen word unigrams. This yielded an overall classification accuracy average after ten-fold cross validation of 86.3% using the Simple Logistic classifier, with a test set classification accuracy of 80% using the SVM classifier.

4.2 Discussion of distinguishing features

Table 9 shows that the German translations have a much higher frequency of the word *toward* as opposed to the other texts. A likely explanation for this is dialectal: two translators of the German texts were American,[8] while the other translations from German were published in the US, by translators whose nationality is not defined.

Table 7 displays the relative frequencies of both *that's* and *it's* and the expanded versions of the same. Olohan (2001) has shown that these forms tend to be less prevalent in translated English

[6] A novel translated from French may be set in a Francophone locale and contain tokens like *Madame*, *Rue*, etc.
[7] This consists of two segments from each work.
[8] Amory Coffin and William Cooper

Run	Training	Test	Classifier	Feature Set	Accuracy
1	Full	10-f cv	Baseline	n/a	25%
2	Full	Test	NB	19 doc-level	55%
3	Full	Test	SVM	19 doc-level	60%
4	Full	Test	SimpLog	19 doc-level	62%
5	Full	10-f cv	NB	19 doc-level	54%
6	Full	10-f cv	SVM	19 doc-level	66%
7	Full	10-f cv	SimpLog	19 doc-level	**68%**
8	Full	Test	NB	Top50(100 POS-bi+19doc+15wuni)	72%
9	Full	Test	SVM	Top50(100 POS-bi+19doc+15wuni)	80%
10	Full	Test	SimpLog	Top50(100 POS-bi+19doc+15wuni)	67%
11	Full	10-f cv	NB	Top50(100 POS-bi+19doc+15wuni)	81%
12	Full	10-f cv	SVM	Top50(100 POS-bi+19doc+15wuni)	80%
13	Full	10-f cv	SimpLog	Top50(100 POS-bi+19doc+15wuni)	**86.3%**
14	Full	Test	NB	30(100 POS-bi+19doc+15wuni)	60%
15	Full	Test	SVM	30(100 POS-bi+19doc+15wuni)	70%
16	Full	Test	SimpLog	30(100 POS-bi+19doc+15wuni)	72.5%
17	Full	10-f cv	NB	30(100 POS-bi+19doc+15wuni)	70%
18	Full	10-f cv	SVM	30(100 POS-bi+19doc+15wuni)	75%
19	Full	10-f cv	SimpLog	30(100 POS-bi+19doc+15wuni)	75%

Table 3: Summary of classification accuracy: Full corpus

in general, however in this case they may be less/more prevalent in translations from different languages. Russian has a much larger proportion of *that's* and *it's*, although it's proportion of *it is* is also relatively high. One possible explanation for this is that in French and German, *that is* and *it is* are two words,[9] whereas in the Russian language, one word zto serves both purposes.

Table 8 displays the frequencies for the contractions *I'm* and *I'll* in the four corpora. Again Russian contains the highest frequency for the two contractions among the languages. This may again be a source language artifact: In German there is no equivalent contraction, *Ich bin* for I am, and in French *je suis*, both two word phrases. In Russian *I am* is corresponds to ya,[10] with

[9] Ger. *es ist* or *das ist* and Fre. *il est* or *qui est*.

[10] Pronounced *ya* with a short a sound.

Chi	Rank	Token	Chi	Rank	Token
191.1184	1	toward	60.2458	11	though
101.8571	2	prepratio	56.4456	12	that's
79.6687	3	nounratio	54.1083	13	RB-CC
78.6035	4	lexrich	52.0254	14	i'll
78.1577	5	thousand	50.1781	15	PRP-CC
69.6095	6	it's	49.9458	16	conjratio
66.4622	7	towards	49.868	17	nodded
62.1622	8	numratio	49.224	18	i'm
62.1324	9	fverbratio	48.7354	19	law
61.1304	10	ari	48.6329	20	FW-FW

Table 4: Features 1-20 for Table 3, run 13

Chi	Rank	Token	Chi	Rank	Token
48.3455	21	VBP-VB	33.2283	36	typetoken
47.5911	22	suddenly	33.1439	37	simpletotal
47.1891	23	scream	32.2981	38	complextotal
46.9136	24	CD-CD	30.9333	39	simplecomplex
46.7665	25	don't	27.0928	40	what's
46.6164	26	resumed	26.4912	41	somewhere
43.3339	27	got	26.2167	42	you're
42.7951	28	drink	26.16	43	thought
37.8411	29	sense	25.7212	44	ain't
37.8411	30	infoload	25.6271	45	gazed
37.8411	31	presently	25.6141	46	beneath
37.8409	32	he's	25.3143	47	there's
37.6963	33	whispered	25.2518	48	say
36.2862	34	avgsent	24.1848	49	won't
35.8047	35	anyone	24.125	50	now

Table 5: Features 21-50 for Table 3, run 13

L1	No. of tokens
German	185413
French	180813
English	148565
Russian	183448

Table 6: Number of tokens in each L1 sub-corpus

I will also being one word, budu.[11] This is a possible reason for the abundance of contracted forms in the translations with Russian as L1.

Table 9 displays the frequencies for the next four words in the list. It is difficult to ascertain whether these are true source language artifacts, although the frequency of *drink* in the translations from Russian may reflect a rather unsavoury national stereotype. It is interesting also that the characters in the German translations tend to agree with an affirmative head movement more often than French or Russian. The high frequency of *thousand* in the French corpus is likely as a result of references to large denominations of the French *franc*.

[11]Pronounced *boodoo*.

Text	**it is**	**it's**	**that is**	**that's**
English	0.002358	0.000361	0.000754	0.000538
German	0.002931	0.000194	0.001106	0.000116
French	**0.003236**	0.000092	**0.001370**	0.000167
Russian	0.003216	**0.001058**	0.001112	**0.001052**

Table 7: Relative frequency of that's/it's

Language	I am	I will	I'm	I'll
English	0.003112	0.000452	0.000318	0.000555
French	0.002500	**0.001416**	0.000061	0.000088
German	0.003463	0.001219	0.000092	0.000205
Russian	**0.003598**	0.000883	**0.000627**	**0.000725**

Table 8: Relative frequency of I'll/I'm

Text	drink	nodded	resumed	thousand	toward	toward
English	0.000194	0.000075	0.000048	0.000075	0.000000	**0.000441**
French	0.000083	0.000011	**0.000227**	**0.000785**	0.00002	0.00038
German	0.000129	**0.000248**	0.000027	0.000167	**0.0006**	0.000010
Russian	**0.000627**	0.000033	0.000016	0.000076	0.00015	0.00029

Table 9: Common word frequencies

Conclusion

Our hybrid approach towards detecting the source language of a literary translation resulted in high classification accuracies using ten-fold cross validation on our translation corpus and also comparably high accuracies on our test set from the same corpus. We have identified a number of trends in our corpus, such as the frequency of certain English contractions (*I'm*, *it's* etc) which may be attributable to source language influence.

As noted at the outset, our work is comparable to research published by Guthrie et al. (2007). If one were to derive a classification of each item from the point at which their method achieved 100% inclusion of the translated item among the top ten items in terms of anomalies pointed out using the vectors of document level features, then precision is at 9%, but recall is at 100%, and accuracy is at 80%. However, note that this depends on two categories: L1 English or L2 English (translated from L1 Chinese). Our experiments provide a further label for which language provided the texts L1 source.

Comparing our results to the work by Baroni and Bernardini (2006), there are similarities, although the tasks were different, we focused on source language detection and they focused on detecting whether a text was a translation or original. Classification results for our task were lower than theirs, they obtained ca. 87.5% accuracy using an ensemble of classifiers and two categories, we obtained ca. 80% accuracy with four categories. Comparing discriminating features, we found optional contractions in English to be discriminatory amongst source languages, while they found optional items in Italian such as clitic pronouns to be markers of *translationese*.

Ongoing work focuses on corpora containing a variety of genres, as well as more source languages, and cross-validation experiments on unseen texts. We also wish to examine longer n-gram sequences such as bigrams and trigrams of words and parts-of-speech, with the possibility of supporting non-contiguous sequences or skip-grams, as used by van Halteren (2008).

Acknowledgments

We are grateful for support from the Science Foundation Ireland (Grant 07/CE/I1142) to the Centre for Next Generation Localisation (`www.cngl.ie`) and for anonymous review feedback.

References

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259.

Coleman, M., & Liau, T. (1975). A computer readability formula designed for machine scoring.. *Journal of Applied Psychology*, *60*(2), 283.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!*, pp. 125–134. International Society of the Learning Sciences.

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised Anomaly Detection. In *IJCAI*, pp. 1624–1628.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of Translationese: A Machine Learning Approach. *Computational Linguistics and Intelligent Text Processing*, 503–511.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133.

Olohan, M. (2001). Spelling out the optionals in translation: a corpus study. *UCREL Technical Papers*, *13*, 423–432.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, Vol. 12, pp. 44–49. Manchester, UK.

Smith, E., & Senter, R. (1967). Automated readability index.. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1.

van Halteren, H. (2008). Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 937–944. Coling 2008 Organizing Committee.